PROTEIN STRUCTURE ELUCIDATION BY COMBINING COMPUTATIONAL AND EXPERIMENTAL METHODS

by

Stephanie Judith Han Hirst DeLuca

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

May 2015

Nashville, Tennessee

Approved:
Professor Albert Beth, Chair
Professor Walter Chazin
Professor Borden Lacy
Professor Terry Lybrand
Professor Charles Sanders

ACKNOWLEDGEMENTS

I would first like to thank Dr. Jens Meiler, my PI, for convincing the Vanderbilt Chemical and Physical Biology Program to accept me, even after a late application submission. I am extremely grateful for the opportunities I have had in his lab and at Vanderbilt in general, and my graduate research experience has been what it has due, in large part, to his mentorship. Additionally, my fellow lab members have been pivotal to my success. Not only are the members of the lab productive, helpful, friendly, and collaborative, they also watch out for and take care of each other.

I have had the privilege of collaborating with members of Prof. Drs. Anette Beck-Sickinger and Daniel Huster from Leipzig University, as well as their students, Dr. Daniel Rathmann and Gerrit Vortmeier. Because of their superb collegiality, we have maintained top-notch, productive partnerships for several years.

I am also indebted to the faculty who sat on my qualifying exam and thesis committees. This includes Drs. Hassane Mchaourab, Walter Chazin, Chuck Sanders, Borden Lacy, Terry Lybrand, and Al Beth. Under the guidance of this group of professors, I have learned how to confidently and passionately present my research and feel that I have truly grown, both as a scientist and as a person. In particular, Dr. Chazin, ever the student advocate, he has always been “tough but fair” and has expressed his confidence in me. Dr. Beth, my thesis committee chair, has been essential for helping me power through the last months of my Ph.D. career. In addition, Dr. Bruce Damon, the CPB Program’s DGS, has been a useful resource.

The amazing Lindsay Meyers has always been a great resource for understanding the confusing bureaucracy of graduate school. Organized, intelligent, and friendly, I was constantly turning to her for her assistance. Also, when I entered graduate school, I did not think I would need the BRET Office counselor, but I was wrong. Thankfully, Darrell Smith and Mistie Germek have been wonderful in their own way.

I would not have chosen this path if it were not for my love of scientific inquiry. I consider my first scientific mentor to be my high school chemistry teacher, Mrs. Vicki Farina, who managed to draw me into the world of science without my even noticing. My professors at UAB were also essential to me pursuing this path. I would like to specifically thank Dr. Gary Gray for spending numerous hours talking about chemistry, science fiction and fantasy, and religion and spirituality, as well as being a superb chemistry professor. I am also indebted to Dr. Christie Brouillette for agreeing to be my senior thesis research adviser and teaching me to be tough and resilient. Finally, Professor Dr. Peter R. Schreiner at Justus-Liebig Universität was immensely kind in allowing me to serve my Fulbright year in his lab, even though I came in with no experience in computational chemistry. Because of the time I spent in his lab, I decided to stop eschewing computational methods of scientific research.

Where would I have ended up without the loving support of my friends? Graduate school is certainly tough at times, but because of the greatest network of friends one could ask for, I have traversed this path and made it to the end. Special thanks to Drs. Liz (Dong) Nguyen and Nathan Alexander for making 5154G the best office in the CSB. Dr. Steven Combs has also been great for all kinds of conversations. I am so glad to have had female company in the male-dominated lab thanks to Brittany Allison and Amanda

Duran. Thuy Nguyen, Veronica Combs, Meryl Harsadi, and Kate Mittendorf were always there to help keep my life balanced as well. I would like to thank my fellow aikidoka at Nashville Aikikai. Nashville Aikikai provided me with a great place to take a break from science and forced me to focus on something else entirely. In addition, I have made some great friends and fellow advocates in the National Federation of the Blind of Tennessee.

I thank God every day for my lovely family. They are such a blessing to me, and I was so fortunate to have been adopted by the Hirsts back in 1986. My parents, Laurie and David, have always been incredibly supportive of me and my interests. They have never failed to tell me that they were proud of me, even when I wasn't proud of myself. My brothers, Brian and Issac, are the best that any sister could ask for. My younger sister, Laura Kate, and my five nephews continue to inspire me to live by example every day. I have been fortunate to have also been "adopted" into the DeLuca family, whom I met because of my husband, Sam. Alice and Ed are like a second set of parents, and Marie is an awesome sister-in-law. They have helped me through this process and are also a wonderful resource of knowledge and experience. I also consider Callisto, my cat, to be part of the family. Without her to be my alarm clock in the morning and greet me at the door in the evenings, I am sure graduate school would have been much less enjoyable.

No words can really describe how much I love my husband, Sam DeLuca. I am so glad we met back in 2008, and even though we tried to avoid joining the same lab, I wouldn't have had it any other way. I do not know if I would have made it through graduate school without him by my side, helping me through every step of the way. Even though we have traded the same words of advice back and forth at different times over the years, it is still good to hear. I look forward to our future ventures together.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The overall focus of this dissertation was to develop, test, and apply computational methods integrated with experimental data for peptide and protein structure determination.

Chapter I outlines the need for novel structural biological methods that can lead to the characterization of peptides and membrane proteins. The prolactin releasing peptide, or PrRP, and the PrRP receptor, as well as ghrelin and the ghrelin receptor, are briefly introduced as examples of biomedically relevant systems for which no experimental structures are available. The first chapter also provides an overview of computational structural biology, including comparative modeling, *de novo* folding, and overcoming the obstacles of effective model scoring and conformational sampling. Portions of the chapter concerning comparative modeling and energy evaluation were taken from a *Nature Protocols* publication entitled, "Small-molecule ligand docking into comparative models with Rosetta," which was written by Steven Combs*, Samuel DeLuca*, Stephanie DeLuca*, Gordon Lemmon*, David Nannemann*, Elizabeth Nguyen*, Jordan Willis*, Jonathan Sheehan, and Jens Meiler. Authors with an (*) after the last name contributed equally to the publication and are considered equally contributing authors. The author of this dissertation contributed significantly to the development, documentation, revision, and dissemination of the reported protocol.

Chapter II is based on the publication, "The activity of the prolactin releasing peptide correlates with its helicity," by Stephanie DeLuca, David Rathmann, Annette Beck-Sickinger, and Jens Meiler. The author of this dissertation and Daniel Rathmann contributed equally to the work reported therein. The author of this dissertation

conducted all modeling and analysis needed to interpret the experimental information. The text was written in a collaborative effort, such that the first two authors listed shared first authorship.

Chapter III concerns the characterization of ghrelin structure and dynamics in a lipid vesicle environment. Ghrelin, like PrRP, is a peptide hormone that is involved in obesity and metabolic disease. Therefore, its three-dimensional structure is of special interest in the field of drug discovery. The manuscript entitled, "Integrating solid-state NMR and computational modeling to investigate the structure and dynamics of membrane-associated ghrelin" by Gerrit Vortmeier, Stephanie DeLuca, Constance Chollet, Holger A. Scheidt, Annette Beck-Sickinger, Jens Meiler, and Daniel Huster, describes the joint-effort work of both the Meiler and Huster laboratories, at Vanderbilt University and Leipzig University, respectively. The author of this dissertation performed all modeling of ghrelin using chemical shifts from solid-state NMR spectroscopy, which was conducted by the Huster laboratory. In addition, she contributed significantly to data interpretation, writing, and editing of the manuscript text and is therefore considered to be co-first author with Gerrit Vortmeier.

RosettaEPR is introduced and described in detail in Chapter IV. It is a computational tool for protein structure determination that employs the Rosetta *de novo* folding algorithm with an EPR distance knowledge-based potential. The content of Chapter IV is based on the publication, "RosettaEPR: An integrated tool for protein structure determination from sparse EPR data" by Stephanie Hirst, Nathan Alexander, Hassane Mchaourab, and Jens Meiler. The author of this dissertation was solely

responsible for the implementation, testing, reporting, and submission of the manuscript for publication.

In Chapter V, RosettaTMH, a novel membrane protein *de novo* folding algorithm in the Rosetta software suite is introduced. RosettaTMH assembles membrane protein topologies via the translation or rotation of entire transmembrane helices at a time. In later stages of folding, peptide fragments are inserted into the *de novo* folded protein backbone, much in the same way that soluble proteins are generated with the traditional Rosetta method. The author of this dissertation was solely responsible for the implementation, benchmarking, and description of RosettaTMH and the work associated therewith. The content of Chapter V is based on a manuscript submitted to *PLoS ONE* entitled, "RosettaTMH: Membrane protein structure elucidation by combining EPR distance restraints with assembly of transmembrane helices" by Stephanie DeLuca, Samuel DeLuca, and Jens Meiler.

Chapter VI concludes the main text of this dissertation. The author of this dissertation, including the concluding chapter, summarizes the work described in detail in Chapters II through V, as well as the implications of the results of that work. The author proposes future experiments and outlines the overall goals, motivation, and contributions of the herein reported research.

Appendix A is based on the publication entitled, "Ligand-mimicking receptor variant discloses binding and activation mode of prolactin releasing peptide" by Daniel Rathmann, Diane Lindner, Stephanie DeLuca, Kristian Kaufmann, Jens Meiler and Annette Beck-Sickinger. The structural basis of activation of the prolactin releasing peptide receptor by the binding of its endogenous peptide hormone, PrRP, was presented.

The binding model was generated by taking an iterative approach to computational modeling, hypothesis generation, and experimental validation. The author of this dissertation performed the majority of the modeling using the experimental data provided by collaborators in the Beck-Sickinger laboratory at Leipzig University. She also contributed significantly to the manuscript text, thus earning her a co-first authorship on the publication, along with Daniel Rathmann and Diane Lindner.

Appendices B, C, D, E, and F comprise the protocol captures for the scientific work reported in Chapters II, III, IV, V, and Appendix A, respectively. The author of this dissertation developed, used, and documented all protocols in this dissertation's appendix.

CHAPTER I

INTRODUCTION

Parts of this chapter were published in (Combs*, DeLuca, S.L.*, DeLuca, S.H.*, Lemmon*, Nannemann*, Nguyen*, Willis*, Sheehan, and Meiler, 2013). *These authors contributed equally.

Structural biology as a valuable approach to biomedical research

With the completion of the Human Genome Project in 2001, the field of structural biology has played an increasingly prevalent role in advancing our understanding of the molecular basis of disease. In 2004, there were fewer than 30,000 depositions in the Protein Data Bank (PDB) (1). In contrast, today, there are more than 100,000 publicly available three-dimensional (3D) structures, and these numbers continue to grow at exponential rates. There were over 5,600 structures deposited in the PDB within the first seven months of 2014 alone (www.pdb.org).

Why is the scientific community so interested in knowing what proteins look like? One of the main driving forces is the role that protein structure determination has played in drug discovery (2, 3). The structure-function relationship of proteins lies at the core of our understanding of biological processes and the basis of disease. If we can "see" a protein's 3D structure, we might be able to develop better drugs and therapeutics that target it. Indeed, numerous experimentally determined structures, such as that for HIV-1 protease (4-6), neuroamidase (for influenza) (7, 8), and epidermal growth factor receptor (9), have been used for drug lead design and optimization. Furthermore, by structurally

characterizing proteins and their interactions with other molecules (e.g., small molecules, peptides, other proteins, etc.), we can enhance our exploration of larger scale mechanisms, such as intra-cell signaling and cell-to-cell communication.

Membrane proteins play a major role in human disease

Proteins can be divided into two main groups. Soluble proteins exist in their native state in solution, such as in the cytosol of a cell. On the other hand, integral membrane proteins, hereafter referred to as membrane proteins (MPs), reside in lipid bilayers found in the permeable membranes of cells and organelles. MPs are involved in a plethora of physiological functions, including maintaining the proper electrochemical balance across cell membranes (10), transporting molecules into and out of the cell (11), and facilitating extra- and intra-cellar communication (12, 13). Their malfunction has been implicated in a myriad of diseases, including schizophrenia (14), depression (15-17), diabetes (18-20), cystic fibrosis (21), and cancer (22). It is therefore not surprising that MPs make up about one-third of all proteins encoded in the human genome and are the biological targets of over half of drugs and therapeutics.

G-protein coupled receptors are a major class of membrane proteins

G-protein coupled receptors, or GPCRs, comprise one of the biggest MP families. There are over 800 GPCRs in the human proteome, which are found all over the body, including the brain, heart, and ovaries. Approximately 30% of drugs are designed to interact with this important group of proteins (23-26). All GPCRs have seven transmembrane helices (TMHs) connected by alternating extra- and intra-cellular loops

(ECLs and ICLs, respectively). Upon interaction with a ligand, the receptors are said to be "activated," at which point they are bound by G-proteins at the C-terminal helical "tail." This then leads to a signal transduction cascade inside the cell. Activation is associated with a conformational change in the receptor, which allows for the transmission of extracellular signals to the intracellular space (27).

GPCRs can be divided into five classes based on their sequence homology. These classes include class A (rhodopsin-like), B (secretin-like), C (metabotropic glutamate-like), D (fungal mating pheromone), E (cAMP), and F (frizzled/smoothened) (23, 27). Class A GPCRs constitute the largest class and bind small-molecule ligands, peptides, and even photons. The two example GPCRs discussed in this chapter, the prolactin releasing peptide receptor and the growth hormone secretagogue receptor, belong to class A.

The prolactin releasing peptide receptor plays an important role in biological processes

The prolactin releasing peptide receptor (PrRPR), also known as GPR10 or Hgr3, was first discovered to be the receptor of the prolactin releasing peptide (PrRP) via reverse pharmacology techniques. This involved screening several tissue extracts against the then "orphan" GPCR and testing for cell signaling. While PrRPR mRNA was found in the spinal cord, adrenal gland, and femur, it was most abundant in the anterior lobe of the brain's pituitary, which is located below the hypothalamus (28). Even though PrRPR appears to be related to prolactin release, it has other biological functions as well. For example, PrRPR-knockout mice exhibited increased body weight and fat mass compared to wildtype (wt) mice after 11 and 15 weeks (29). Along with the increased body weight

and obesity, the knockout mice had decreased glucose tolerance and increased leptin, cholesterol, LDL, and HDL. Interestingly, food intake was actually decreased in female knockout mice. In addition, the Otsuka-Long-Evans-Tokushima Fatty (OLETF) rat strain, which serves as a rodent model of type II diabetes, encodes a mutant form of GPR10, in which the amino terminus is truncated. Binding of isotopically labeled PrRP was not detected in the reticular thalamus of OLETF rats compared to controls, indicating that the receptor mutation may play a role in the diabetes-like phenotype (20).

Prolactin releasing peptide--the endogenous agonist of the PrRPR

PrRP was originally isolated from bovine hypothalamus (28), but its mRNA has since been found in other tissues. In rats, PrRP mRNA was detected in the pituitary, the medulla oblongata, and the hypothalamus (30, 31). In humans, it was found in the medulla oblongata and the pancreas (31). Interestingly, PrRP is also able to activate RF- and FF-amide receptors, such as the human NPFF2 receptor (32). As the endogenous agonist of the PrRPR, it is not unexpected that circulating PrRP levels are also associated with energy and body weight homeostasis and metabolic diseases, such as obesity and diabetes. For example, reduced PrRP mRNA levels are found in fasted male rats and obese Zucker rats (33). Further, injection of PrRP into the central nervous system of PrRPR-knockout mice resulted in decreased food intake, and repeated administration of the peptide appeared to cause increased energy expenditure, as measured by core body temperature and oxygen consumption (29).

PrRP is a member of the RF-amide peptide family, the members of which are so-named for their arginine-phenylalanine C-terminal residues. There are two isoforms of

PrRP, PrRP20 and PrRP31, which have 20 and 31 residues, respectively. Both isoforms bind to the PrRPR with a potency of less than 10 nM. The seven C-terminal residues of PrRP, PrRP14-20 (PrRP25-31), were also able to stimulate the receptor but exhibited reduced binding (30). Later, structure-activity relationship (SAR) studies indicated that the 13 C-terminal residues are able to agonize the PrRPR. The authors found that amidation of the C-terminus is required for activity and that mutation of R^{19} / R^{30} and F^{20} / F^{31} was not well-tolerated (34).

Nuclear magnetic resonance (NMR) structural studies of PrRP revealed a stable helical conformation with a flexible N-terminus (35), but the Cartesian coordinates of the resulting structural ensemble was not made publicly available. Another study also published an image of the peptide structure but no experimental details or coordinates (36). Therefore, in order to study the structure of PrRP and its interaction with the PrRPR, the published chemical shifts (CSs) and nuclear Overhauser effect distances (NOEs) were used to generate an ensemble of models that agreed with the experimental data (Figure 1). Taken together with circular dichroism (CD) spectroscopic studies and receptor activity data collected using several PrRP mutants, it was found that the ability of PrRP to activate the PrRPR depends on its helical propensity (37) (Chapter II). The models were computationally docked into a comparative model of the PrRPR to provide, in addition to a large set of pharmacological data, a structural biological perspective of the binding mode of PrRP to PrRPR. Substitution of $D^{6.59}$ (Ballesteros-Weinstein numbering (38)) on the PrRPR to arginine resulted in a constitutively active receptor. While the mutant receptor exhibited little to no activity when wt PrRP20 was added to expressing cells, activity was recovered with D^{19}PrRP20, indicating that residue 19 on

PrRP20 and residue 6.59 on the PrRPR form an electrostatic interaction. Further, double-cycle mutagenesis experiments pointed to a second interaction partner on the receptor, $E^{5.26}$ (19) (Figure 1, Appendix B).

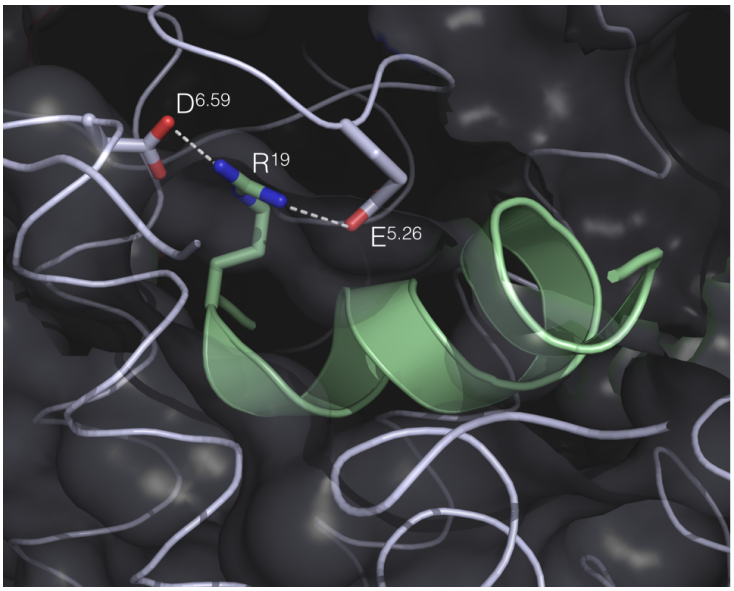

Figure 1: Dual binding mode of PrRP to the PrRP receptor
The 13 C-terminal residues (8-20) of PrRP (green), *de novo* folded using RosettaNMR (37) (Chapter II) docked into the PrRP receptor binding site (lavender) (39) (Appendix B). $D^{6.59}$ and $E^{5.26}$ on the PrRP receptor and R^{19} on PrRP are shown as sticks and labeled accordingly. Receptor helices are rendered as ribbons.

The ghrelin receptor is implicated in variety of diseases and physiological functions

The growth hormone secretagogue receptor 1a (GHSR1a), also known as the ghrelin receptor, is another GPCR. It is primarily located in the hypothalamus (40, 41), but lower expression levels have also been observed in the thyroid and adrenal glands, the myocardium, and the spleen (42, 43). This receptor was first found to be activated by the synthetic growth hormone releasing peptide, which stimulates growth hormone secretion from the pituitary gland (44). Since then, it has been found to be involved in appetite regulation, energy expenditure, and reward-driven behaviors (45, 46), such as

alcohol intake in mice (47). Synthetic agonists of the receptor may be useful in improving learning and exploratory behavior (48), and ghrelin receptor antagonists reversed ghrelin-induced increase in food intake (49). GHSR1a exhibits an inherently high basal level of activity (50, 51). Interestingly, upon ligand binding, the ghrelin receptor can induce cell signaling via multiple pathways, which indicates that it may be capable of functionally biased signaling (52). Given the broad range of functions in which this receptor is implicated, it is important for the biomedical research community to understand its mechanism(s) of activation in the context of its structure and dynamics.

Ghrelin is a unique peptide hormone that activates the ghrelin receptor

In 1999, the peptide hormone, ghrelin, was discovered to be an endogenous ligand of GHSR1a. Several studies have demonstrated ghrelin's importance in regulating appetite. For example, increased levels of ghrelin in blood plasma were measured in human subjects shortly before mealtime, which then decreased afterwards (53). However, in general (i.e., not before meals), obese individuals exhibited decreased amounts of ghrelin in their plasma relative to lean individuals (54). In addition to its orexigenic effects, administration of ghrelin has been shown to lead to improved memory retention (55, 56).

Ghrelin, like PrRP, is a short peptide, having a primary sequence of 28 amino acids. It is the only known peptide hormone that has a lipid modification. Even though desacylated ghrelin is more prominent in the bloodstream, Ser^3 of the peptide must be acylated in order to activate GHSR1a (41, 57). While the original discovery of ghrelin pointed to an octanoyl group on Ser^3 (41), fatty acids of different lengths can also be

added via ghrelin O-acyltransferase (58-60). In addition to the acylation at Ser^3, a peptide core consisting of residues 1 to 4 is able to activate GHSR1a *in vitro* (57).

The structure of ghrelin remains under debate. NMR and CD spectroscopy of both ghrelin and desacyl ghrelin indicate that both peptides are highly unstructured in aqueous solution (61). More recent studies unsurprisingly indicate that the peptide becomes increasingly helical with the addition of trifluorethanol (TFE) (62). This helical model is supported by molecular dynamics (MD) simulations performed in a membrane/water environment (63), and low-resolution ^1H NMR studies in cells support a helical secondary structure (64). New structural and dynamical information from solid-state NMR (ssNMR) also point to a highly mobile structure, in which a semi-helical peptide is bound to the membrane of lipid vesicles (Chapter III).

Membrane protein and peptide structure elucidation by computational methods

Membrane proteins pose a special challenge for traditional structural biological methods

Despite their clear biological significance, including the two specific examples given above, MPs make up less than 2% of all proteins of known structure. This statistic points to a number of technical and methodological difficulties encountered in protein over-expression, purification, and structural elucidation. For example, complications often include low expression levels, protein aggregation, instability, and insolubility. Unlike soluble proteins, MPs must be reconstituted into membrane mimetics that do not perturb their native conformations. This often involves extensive screening of appropriate conditions (65). Further, once enough protein can be obtained for structural characterization, optimal conditions for obtaining diffracting crystals (for X-ray

crystallography) or assignable spectra (for NMR) must be established. Additional protein engineering, such as site-directed mutagenesis (66), T4-lysozyme (67, 68) fusion, or antibody binding (66), may be required in order for a MP to crystallize, which may or may not significantly disturb the protein's native conformation.

A number of technological advances have been made in recent years to address the numerous challenges associated with MP structural biology. Among these are automated liquid handling and high throughput crystallization for X-ray crystallography (69) and methyl-TROSY methods, site-specific labeling, and the use of paramagnetic relaxation enhancements for solution NMR (70-72). Other structural biological techniques, such as electron paramagnetic resonance (EPR) (73-77), and ssNMR spectroscopy (78-81) have also been shown to be useful, but unlike X-ray crystallography and solution NMR, these methods do not currently allow for 3D structure elucidation. As a result, despite the improvements in technology, much of which has stemmed from structural genomics initiatives, MP structures are sorely underrepresented (82) (Figure 2). Computational methods for MP structure prediction have also been progressing at a slow but steady pace. The two main MP structure prediction approaches are comparative modeling and *de novo*, or *ab initio*, folding. However, even with an increasing number of template structures for comparative modeling and more sophisticated approaches to *de novo* folding, the use of experimental data in order to limit the conformational search space is often required (see below).

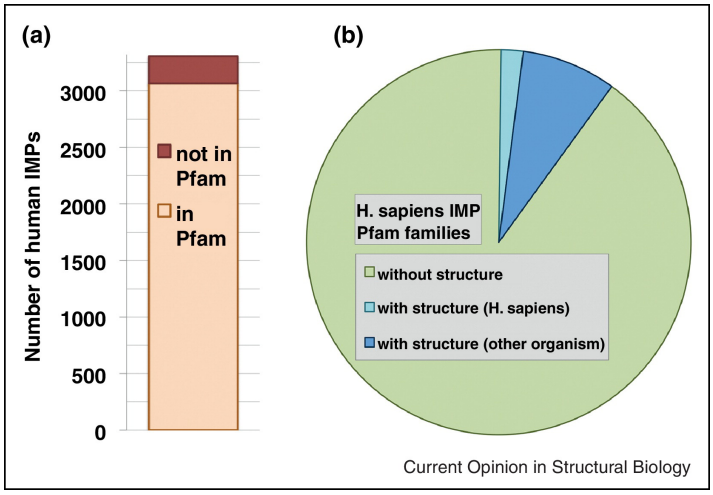

Figure 2: Human integral membrane proteins
Pfam (83) families and PDB structures. a) Mapping human MPs to Pfam families. Three thousand, three hundred and five (3,305) polytopic α-MPs were extracted from the 20,247 sequences part of the SwissProt *Homo sapiens* proteome (UniProt release February 22, 2012) using PolyPhobius (84). Assignment of proteins to Pfam families was done as described in (82) using the transmembrane assignment of PolyPhobius. Three thousand and sixty-three (3,063) MPs can be mapped to a Pfam family (orange); 242 MPs fall outside of the current Pfam collection of families (red). b) Human MP Pfam families covered by structure. Human MP Pfam families with no structural representative (green) and with at least one structural representative (blue: representative is a human protein; light blue: representative is not a human protein) (82).

Small peptides are highly flexible, making structure determination difficult

Similarly to MPs, the structure elucidation of small (< 50 amino acids) peptides also appears to be challenging. An advanced search of the PDB for molecules in the "peptide" class according to the Structural Classification of Proteins (SCOP) (85) having sequence homology of less than 90% and fewer than 50 residues returned fewer than 700 hits, which is less than 1% of all depositions.

Small peptides are often highly flexible and unstructured, which makes them difficult to crystallize. Furthermore, due to their lack of stable secondary structure, NMR peak assignment can become cumbersome, if not impossible. Even when chemical shifts (CSs) can be assigned, the population of peptide conformations can be highly heterogeneous. As a result, structural characterization of peptides is often limited to low-

resolution techniques, such as CD, fluorescence, and fourier transform infrared (FTIR) spectroscopy (86). These methods allow for the study of peptide secondary structure and conformations. MD simulations of peptides have also been performed in order to observe them on an atomic level, but these simulations are often not long enough to capture secondary structure transitions or larger-scale conformational changes (87-90).

Computational modeling can serve as an alternative approach to protein structure determination

There are two main means of computational protein structure prediction: comparative modeling, which is often referred to as homology modeling, and *de novo*, or *ab initio* (i.e., from the primary sequence) folding. These two approaches are methodologically distinct from one another, but they can both be used for generating 3D-models of proteins relatively quickly. Further, they can both be paired with experimental information to produce models that are consistent with empirical data.

Comparative modeling relies on the availability of structures of related proteins

Comparative modeling refers to the elucidation of the tertiary fold of a protein, guided by the known structure of another, often homologous, protein. The unknown structure is commonly called the "target," while the protein of known structure, upon which the primary sequence of the target is threaded, is termed the "template." The known template structure reduces the conformational search space by providing a protein backbone scaffold; areas where the template and target sequences diverge significantly are typically remodeled and refined via the loop building application. Although the

application is known as “loop building,” a “loop” is defined here as any area where the backbone is to be rebuilt *de novo*, which most often occurs in flexible regions but can also include secondary structural elements (SSEs). Comparative models have played a major role in aiding experimental design and the interpretation of experimental results. They can be employed to help predict structure-function relationships (91), predict binding pockets for ligands during structure-based drug design (92), and aid in the determination of target residues for site-directed mutagenesis (93, 94).

Modeller (95) is one of the most popular comparative modeling tools. Comparative modeling with Modeller is highly automated and, as with Rosetta, works best for cases in which the sequence identity between the target sequence and the template structure is 30% or greater. It works by optimizing the comparative model’s satisfaction of spatial restraints derived from one or multiple templates. Comparative modeling in Rosetta (96, 97) is a multiple-step process that requires more input from the user; specifically, user-defined alignment and loop definitions are taken into account throughout the process. These definitions can be provided to Modeller but are not necessary for the program to generate a model.

Sometimes, homologous, experimentally determined structures cannot be identified for use as templates, in which case homology modeling is not applicable. However, as structure is better conserved evolutionarily than sequence, proteins with low sequence identity can have similar folds. In this case, 3D-fold recognition meta-servers, such as Phyre (98) can be used. Phyre constructs a “fold library” via three steps: 1) combining a library of proteins of known structure from the SCOP database (85) with new entries from the PDB, 2) scanning the sequences against a non-redundant sequence

database, and 3) constructing a sequence profile from the previous step. When a query sequence is submitted to the server, Phyre produces a sequence profile of the query with potential homologs by running PSI-BLAST, generates a consensus secondary structure prediction of the query after running a plethora of secondary structure prediction methods, and performs a profile-profile alignment of the results from PSI-BLAST and secondary structure prediction by scanning these inputs against the fold library. The resulting alignments are scored and ranked (99). Once a suitable template has been identified, a sequence alignment should be performed between the target and template sequences.

De novo *folding with Rosetta*

When only the primary sequence of a protein is known, *de novo* folding can sometimes be used to predict the protein's tertiary structure. This method of protein structure determination is usually only considered when a suitable template structure for comparative modeling cannot be found or if the protein has a potentially novel fold. The Rosetta software suite is one of the most commonly used programs for *de novo* folding (100-103). However, to date, Rosetta has been shown to successfully fold only small, soluble proteins (fewer than 150 amino acids) and performs best if the proteins are mainly composed of SSEs (α-helices and β-strands) (104). Helical MPs between 51-145 residues were predicted within 4Å of the native structure (105). Accurate prediction of larger and/or more complex proteins can be achieved with the addition of experimental data, such as NMR CSs and distance data (106-108). Further, only sequences of very small proteins (up to 80 residues) have been predicted to atomic-detail accuracy in the

absence of experimental restraints (109-111). Therefore, whenever an experimental structure of a related protein is available, comparative modeling is the method of choice.

Other de novo *folding methods are also available*

While Rosetta was one of the earliest tools for *de novo* protein structure prediction, there are numerous other promising software tools as well, but most are primarily applicable to soluble proteins. The Zhang lab's QUARK (112), for example, was one of the top performers in the template-free modeling category in the most recent Critical Assessment of protein Structure Prediction (CASP), which is held biannually (113). QUARK, like Rosetta, combines fragment-based assembly with knowledge-based potentials. QUARK samples conformational space of protein folds via replica exchange Monte Carlo simulations, where the initial structure for each simulation is constructed by randomly connecting the peptide fragments, which can range from 1-20 residues. EVFold leverages the co-evolution of pairs of residues to generate initial 3D geometries of proteins before refining the models using simulated annealing molecular dynamics (SAMD) (114). The BioChemical Library *de novo* folding protocol, called BCL::Fold, was able to accurately predict topologies for 61 of 66 test proteins ranging in size from 83 to 293 amino acids via the movement of idealized SSEs in the presence of knowledge-based potentials (115, 116).

There are relatively few de novo *folding methods for membrane proteins*

Compared to the numerous soluble protein *de novo* folding tools available, there are relatively few MP-specific methods. RosettaMembrane, which was introduced in

2006, was initially tested on 12 helical MPs and was able to predict between 51 and 145 residues with an root mean square deviation (RMSD) of less than 4Å relative to the crystal structure (105). In 2009, an alternative version of RosettaMembrane was shown to be able to fold 9 of 12 MPs that were 190 to 300 residues in length with approximately the same level of accuracy (117). A MP-specific version of EVFold has been reported to fold MPs of up to 14 helices with impressive accuracy by employing the same concepts as the original EVFold (118). FILM3 employs fragment-based assembly, in which secondary structure prediction is used to select fragments, in combination with a scoring function that takes only correlated mutational information of the protein into account. This method achieved an RMSD of 5.7Å over an MP of 514 residues in length (119).

De novo *folding tools for peptides are also limited*

There are a number of tools for folding peptides *de novo* (120-124). Two of the best-performing methods are PEP-FOLD (120) and Rosetta FlexPepDock (121). However, PEP-FOLD can only fold peptides that are between 9 and 23 amino acids, and FlexPepDock is designed to fold peptides in the presence of the peptide binding site of a soluble protein receptor. MD simulations can be used to predict the 3D structures of peptides as well, but these tend to be computationally expensive and require non-equilibrium sampling strategies, such as Monte Carlo, replica exchange, or parallel replica dynamics (90, 125, 126).

The protein structure prediction sampling and scoring problems

The two main types of energy functions: physics-based and knowledge-based

In order to predict the structures of proteins computationally, computational structural biologists must address two main challenges when developing or improving prediction methods. These are often referred to as the scoring and sampling problems. During modeling, protein conformations are often evaluated via one or more energy, or scoring, terms. The set of scoring terms used during model assessment is called the scoring function, energy function, or force field. Broadly speaking, energy functions come in two primary categories: physics-based and knowledge-based. As implied by the name, physics-based scoring functions, or potentials, employ physical principles in their treatment of protein conformations. These are most often based on Newtonian's laws of motion, in which, for example, chemical bonds are treated as springs. The use of Newtonian mechanics-based force fields is commonly, but not always, used in MD simulations.

Knowledge-based potentials (KBPs) make the assumption that, in terms of protein structure, naturally occurring phenomena, such as torsion angles, hydrogen bonding propensities, etc., are common because they are energetically and biologically favorable. These potentials are generally derived by collecting statistics on proteins of known structure (e.g., from the PDB) and correlating statistical propensities with energies via the Boltzmann relationship. One advantage of KBPs is that they require relatively little computational power to generate. However, they are inherently limited in their accuracy because they are developed from available protein structures, which can be problematic for the evaluation of conformations of MPs and peptides.

The RosettaMembrane energy function

The energy function in Rosetta is derived empirically through analysis of observed geometries of a subset of proteins in the PDB. The measurements include, but are not limited to: radius of gyration, packing density, distance/angle between hydrogen bonds, and distance between two polar atoms. The measurements are converted into an energy function through Bayesian statistics (102, 127). The scoring function in Rosetta can be separated into two main categories: centroid-based scoring and all-atom scoring. The former is used for *de novo* structure prediction and initial rounds of loop building (102, 128, 129). The side-chains are represented as "super-atoms," or "centroids," which limit the degrees of freedom to be sampled while preserving some of the chemical and physical properties of the side-chain. When *de novo* folding MPs in Rosetta, MP-specific scoring terms are used (105). These scoring terms are similar in nature to those used for soluble protein folding, but they were derived while taking the RosettaMembrane implicit membrane environment account. This membrane environment is divided into 5 main sections: 1) inner hydrophobic, 2) outer hydrophobic, 3) interface, 4) polar, and 5) water. During folding of MPs, each amino acid's position in the membrane is determined and its contribution to the overall energy of the model computed accordingly.

The RosettaMembrane all-atom scoring function represents side-chains in atomic detail (130). Like the centroid-based scoring function, the all-atom scoring function is comprised of weighted individual terms that are summed to create a total energy for a protein. Most of the scoring terms are derived from statistics generated over proteins of known structure. The MP full-atom energy function assesses van der Waals attractive / repulsive forces based on the Lennard-Jones 6-12 potential. It also includes scoring terms

that evaluate backbone torsion angles, inter-residue pairing propensities, and solvation based on the Laaridis-Karplus model (131). There is also an orientation-dependent hydrogen bonding term (132). The solvation and hydrogen bonding terms were modified in order to account for the implicit membrane environment. While the atomic-level RosettaMembrane energy function is important for applications, such as small-molecule ligand docking, peptide docking, and comparative modeling, it is not employed during *de novo* folding and will therefore not be discussed in further detail. The development and implementation of the scoring function is reported in (130).

Means of limiting protein conformational sampling during de novo *folding*

Overcoming the sampling problem is an ever-present goal in the development of protein structure prediction methods. While approximations of conformational energies using Newtonian physics or KBPs can speed up simulations, sampling complex topologies, such as those seen in transporters and GPCRs, often requires the implementation of more clever folding algorithms. For example, Rosetta estimates local interactions via the use of 3- and 9-amino acid fragments generated from a database of proteins of known structure (100, 102). While this allows for efficient sampling for small proteins, such as ubiquitin, it is insufficient for folding large proteins with complex folds. To address this weakness, BCL::Fold can fold proteins with higher contact order by moving entire SSEs (115). EVFold (114, 118, 133) and FILM3 (119) confine the conformational search space by generating distance restraints based on co-evolutionary information. In the case of RosettaMembrane (105, 117), the implicit membrane environment itself imposes an additional constraint due to the fact that it favors the

placement of hydrophobic residues in the membrane core and requires helices, which alternate N- to C-terminus, to lie relatively orthogonal to the membrane plane.

Combining computational methods with experimental data can further reduce conformational search space

Incorporation of experimental data into structure prediction and analysis has also been shown to improve the quality of the final model or ensemble of models (106-108, 134-137). This is because, in addition to other sampling enhancements, such as fragment insertion, experimental restraints can further narrow down the conformational search space (Figure 3). Numerous types of experimental data have been incorporated into such protocols, including electron density from X-ray crystallography (138) and electron microscopy, NMR distance and orientation data (137, 139), EPR distance data (134, 135), crosslinking restraints (140), small angle X-ray scattering data (141), and deuterium exchange mass spectrometry data (142). While these types of information are more often applied to *de novo* protein structure elucidation, they can also be of some utility in the building of loops (19), reorientation of domains during comparative modeling, or identification of residues involved in ligand binding.

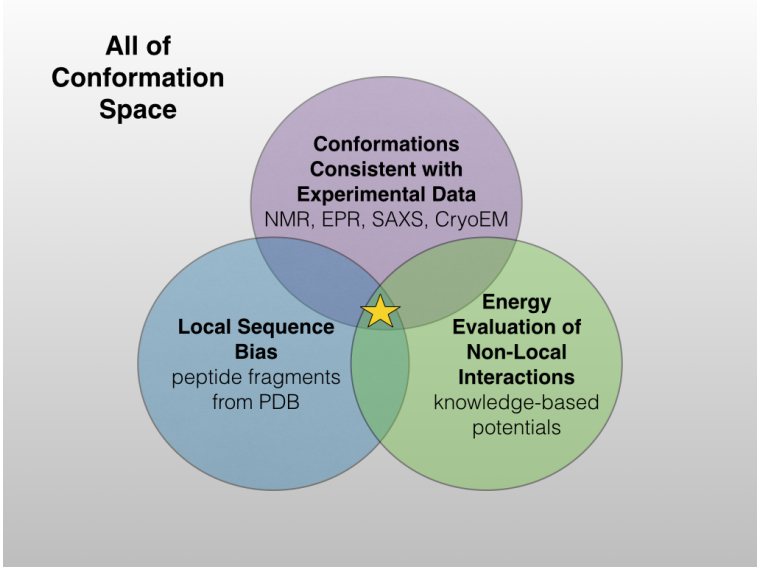

Figure 3: Rosetta approach to limiting conformational sampling
If all of a protein's conformational space is represented by the gray background, Rosetta enhances its sampling of that space by incorporating local sequence bias (blue), evaluating non-local interactions based on KBPs (green), and taking experimental data into account (purple). Ideally, the native structure or ensemble of structures (star) will be found at the intersection of all three approximations.

Protein structure determination by combining NMR and computational methods

Protein structure determination using NMR spectroscopy requires that computational methods be used to generate an ensemble of models that represents protein conformations consistent with restraints derived from the NMR data. The quality of an ensemble of models determined by NMR is often reported as RMSD, which is, in this case, a measure of precision. That is, the "tightness" of a structural ensemble will have a lower RMSD than a "looser" one. Traditionally, the restraints employed during the structure calculations are derived from inter-proton distance information arising from NOEs, as well as dihedral angle restrictions based on secondary structure definitions predicted from the CS values of the individual residues (143). It should be noted that other NMR experimental information, such as that resulting from residual dipolar

couplings (RDCs) and paramagnetic relaxation enhancements (PREs) can be used as well. Typically, the quality of the conformational ensemble is directly related to the richness of the NMR dataset. It is often difficult to fully assign NMR CSs, which makes the acquisition of such datasets difficult. This is especially the case for intrinsically disordered, or unordered, proteins (IDPs and IUPs, respectively), which the protein or peptide undergoes conformational changes and dynamics on timescales for which NMR is ill-suited. In these cases, additional information, such as from homologs of the protein for which 3D structures are available, can be used to generate an ensemble using sparse NMR data. Further, "bootstrapping" methods, in which ambiguous NMR structural restraints are used for modeling in an iterative fashion, can also be helpful (107, 108, 136, 144).

Protein structure determination by combining EPR and computational methods

The applicability of site-directed spin labeling (SDSL) EPR spectroscopy combined with X-ray crystallographic information has been demonstrated in the characterization of potassium ion channels (145, 146) and ABC transporters (74, 147). The secondary structural environment (148), burial state (149, 150), and position in the membrane of the spin label can be probed by measuring collision frequencies with NiEDDA (Ni(II) ethylenediaminediacetic acid) and molecular oxygen (O_2) (151). Global geometric restraints can be derived from distances between two spin labels 5-80Å apart. EPR distance measurements have been used extensively in studies of protein dynamics (152, 153). This structural biological technique requires relatively low amounts of protein

due to its sensitivity and is not restricted by protein size or native environment (154), making it an appealing tool for studying MP structure and conformational changes.

The combination of SDSL-EPR and computational modeling is steadily becoming a popular method of protein structure elucidation. In 1995, an automated method for modeling the 7 TMHs of GPCRs was presented. In addition to other experimental and sequence information, SDSL-EPR data provided information concerning labeled residues' orientation relative to the inside or outside of the 7-helix bundle (155). The structure of α-synuclein was built and refined using SAMD restrained by EPR-determined immersion depths and distances. It was found that α-synuclein forms an extended, curved α-helical structure that is over 90 amino acids in length (156). In another study, EPR-computational modeling hybrid methods were used to propose a novel closed conformation of MscS that includes the previously unresolved NH_2-terminus. The authors proposed that the MscS closed state is in a different and more compact conformation than the one trapped in the crystal structure (157).

EPR distances have a rather large uncertainty when translated into distances between C_αs or C_βs unless the conformation of the spin label is known at every site. To solve this problem, Alexander, *et al*. presented a low-resolution spin-label model in which spin label distances are converted into distance ranges between C_βs by using a "motion-on-a-cone" model. This approach was tested on T4-lysozyme and αA-crystallin with Rosetta, the results for which yielded 1.0Å and 2.6Å full-atom models, respectively (134). In a related approach, the restraint-driven Cartesian transformations (ReDCaT) method for calculating conformational changes in MPs employs a distance deviation factor, ∂, to define a range between the restraints' upper and lower limits. After using this

method for modeling, analysis of the structural basis of activation gating in the K^+ channel, KcsA, revealed a mechanism consistent with a scissoring-type motion of the TM2 segments (158). The aforementioned structural study of MscS also modeled spin labels using ReDCaT (157). More recently, RosettaEPR was introduced, in which an EPR distance KBP was derived based on the cone model and used to fold T4-lysozyme to atomic detail (135). EPR data from double electron-electron resonance (DEER) experiments were also used to guide the modeling of conformational changes seen upon GPCR activation (159, 160). Although a spin label rotamer library based on the crystal structure of T4-lysozyme labeled with methanethiosulfonate (MTS) has been developed (161), methods for modeling EPR spin labels in atomic detail are also desired. Sale, *et al*. described a method for enhancing the utility of dipolar EPR distances as constraints in modeling protein structures by explicit incorporation of the spin labels and showed that accounting for the probe conformation and tether length increased accuracy of distance measures 2-fold (162). Simulated scaling, which couples the random walk of a potential scaling parameter and MD in the framework of a Monte Carlo, proved to be an efficient means of mapping the MTS spin label to atomic detail in spin-labeled T4-lysozyme (163). Additionally, spin label rotamer libraries have been developed (164).

Membrane protein and peptide structure determination made possible by computational-experimental hybrid technologies

Computational modeling and experimental data from NMR and EPR spectroscopy has the potential to provide much insight into MP structure and function. While computational methods alone cannot currently sample conformational space

sufficiently enough to reliably predict MP structures, and flexible peptides and complex MPs continue to evade structure determination by more traditional methods, when taken together, these diverse methods have shown to be synergistic. The work presented in the following chapters describe a few examples of how the coupling of computational biology with experimental data can enable scientists to learn about the structural basis of protein function and presents a new method for MP structure prediction that can be used in conjunction with experimental data.

CHAPTER II

THE ACTIVITY OF PROLACTIN RELEASING PEPTIDE CORRELATES WITH ITS HELICITY

This work is based on publication (DeLuca*, Rathmann*, Beck-Sickinger, and Meiler, 2013). *These authors contributed equally.

Summary

The prolactin releasing peptide (PrRP) is involved in regulating food intake and body weight homeostasis, but molecular details on the activation of the PrRP receptor remain unclear. C-terminal segments of PrRP with 20 (PrRP20) and 13 (PrRP8-20) amino acids, respectively, have been suggested to be fully active. The data presented herein indicate this is true for the wildtype receptor only; a 5-10-fold loss of activity was found for PrRP8-20 compared to PrRP20 at two extracellular loop mutants of the receptor. To gain insight into the secondary structure of PrRP, we used CD spectroscopy performed in TFE and SDS. Additionally, previously reported NMR data, combined with RosettaNMR, were employed to determine the structure of amidated PrRP20. The structural ensemble agrees with the spectroscopic data for the full-length peptide, which exists in an equilibrium between α- and 3_{10}-helix. We demonstrate that PrRP8-20's reduced propensity to form an α-helix correlates with its reduced biological activity on mutant receptors. Further, distinct amino acid replacements in PrRP significantly decrease affinity and activity but have no influence on the secondary structure of the

peptide. We conclude that formation of a primarily α-helical C-terminal region of PrRP is critical for receptor activation.

Introduction

The prolactin releasing peptide, or PrRP, is a member of the RF-amide peptide family and is mainly expressed in the medulla oblongata, brainstem, and hypothalamus (30, 31, 165). It is the endogenous agonist of the PrRP receptor (also known as GPR10 or hGR3) and interacts with nanomolar binding affinities (28). Furthermore, it has some affinity for other RF-amide and FF-amide receptors, such as the hNPFF2 receptor (32). These receptors are integral membrane proteins that belong to the large family of G-protein coupled receptors, or GPCRs, which constitute about one-third of all major drug targets (25). While the original function of PrRP was proposed to be the stimulation of prolactin secretion (28, 166), it is now generally accepted that this is not the primary function of the peptide. Increasing evidence indicates that PrRP plays a significant role in food intake and body weight homeostasis (167). Indeed, intracerebroventricular administration of PrRP with leptin in rats resulted in body weight gain (33). In addition, both PrRP- and PrRP receptor-deficient mice were shown to develop late-onset obesity (29).

PrRP exists in two isoforms: PrRP20 and PrRP31, which consist of 20 and 31 residues, respectively. The C-terminal residues of both isoforms are identical, and both isoforms are biologically equipotent in the activation of the PrRP receptor. It has been demonstrated that PrRP can be shortened to PrRP8-20 without any loss of activity at the

wildtype (wt) receptor and that these thirteen C-terminal residues are the minimum number of amino acids essential for full activation of the PrRP receptor (34).

Little is known about the mode of binding and activation of the PrRP receptor by PrRP, especially on a structural level. This is likely due to the lack of functional antagonists of the PrRP receptor and difficulties in structure determination of GPCRs. Here, we investigate the importance of the peptide's secondary structure for receptor activation. Nuclear magnetic resonance (NMR) spectroscopy had previously been used to determine the structure of PrRP20 in micelles (35). A second study reported an image of a PrRP20 structural model without revealing experimental details, such as solvent conditions or a list of NMR restraints (36). *Neither study made the models publicly available.* However, D'Ursi *et al*. provided a list of sparse chemical shifts and nuclear Overhauser effect distance restraints (NOEs) (35). We employed RosettaNMR (104, 106, 168) to generate an ensemble of peptide conformations that is consistent with newly obtained circular dichroism (CD) spectroscopy data and this set of NMR restraints. Further, we identified receptor mutants for which PrRP8-20 displays a significant loss in activation compared to PrRP20. By comparing the activation ability of four PrRP analogs on two receptor mutants, we can distinguish direct effects on ligand-receptor interaction and indirect effects that result from alteration of peptide helicity. This combined computational-experimental approach allows us to understand the interaction of PrRP and its receptor on a molecular level.

Materials and methods

Structure determination using RosettaNMR

Details of the RosettaNMR protocol have been described elsewhere (104, 106, 137, 168). Briefly, torsion angle restraints were derived from 13 H_α chemical shift values using TALOS (169) (Table 1). Further, 28 distance restraints obtained from NOEs between backbone hydrogen atoms were used and were classified as either "strong" (proton-proton distance ≤ 3Å) or "weak" (proton-proton distance ≤ 5Å) **(**Table 2). A library of overlapping 3- and 9-amino acid peptides spanning residues 8-20 of PrRP20 were generated from coordinates found in the PDB. During folding, an additional 10 NOEs resulting from resonances between side-chain protons--again, classified as "strong" (≤ 3Å) or "weak" (≤ 5Å)–were included as distance restraints (Table 2).

Ten thousand backbone-only structural models were generated using RosettaNMR's *de novo* folding algorithm(100, 102). From these original models, the 10% most energetically favorable models (according to the RosettaNMR scoring function) were refined to atomic detail, including the addition of the functionally obligatory C-terminal amide functional group. The RosettaNMR energy function includes solvation, electrostatic interactions, van der Waals attraction/repulsion, and hydrogen bonding, all of which were included in the assessment of overall structural quality(100, 170). The 20 conformations that fulfill the distance restraints with deviations smaller than 1Å and have the lowest RosettaNMR energies constitute a conformational ensemble that is consistent with the published NMR data and is physically plausible according to the RosettaNMR energy function.

Table 1: Chemical shifts (35) used to generate 3- and 9-amino acid fragments

Residue	Δ in H_α Chemical Shift[a]
W^{8}	4.34
Y^{9}	4.03
A^{10}	4.14
S^{11}	4.31
R^{12}	4.18
G^{13}	3.86
I^{14}	4.12
R^{15}	4.65
P^{16}	4.46
V^{17}	4.14
G^{18}	3.94
R^{19}	4.05
F^{20}	4.60

[a] For G^{13} and G^{18}, took the first value reported for the H_α chemical shift

Table 2: NOEs (35) used to generate 3- and 9-amino acid fragments and to *de novo* fold and refine PrRP models

Residue Pair	NOE Type	NOE Strength[a]
W^{8}-Y^{9}	H_N-H_N	weak [b]
Y^{9}-A^{10}	H_N-H_N	weak
A^{10}-S^{11}	H_N-H_N	weak
S^{11}-R^{12}	H_N-H_N	weak
G^{13}-I^{14}	H_N-H_N	weak
I^{14}-R^{15}	H_N-H_N	weak
V^{17}-G^{18}	H_N-H_N	strong
G^{18}-R^{19}	H_N-H_N	strong
W^{8}-Y^{9}	H_α-H_N	weak
Y^{9}-A^{10}	H_α-H_N	weak
A^{10}-S^{11}	H_α-H_N	weak
S^{11}-R^{12}	H_α-H_N	weak
R^{12}-G^{13}	H_α-H_N	weak
G^{13}-I^{14}	H_α-H_N	weak
I^{14}-R^{15}	H_α-H_N	weak
P^{16}-V^{17}	H_α-H_N	weak
V^{17}-G^{18}	H_α-H_N	strong
G^{18}-R^{19}	H_α-H_N	weak
R^{19}-F^{20}	H_α-H_N	strong
W^{8}-Y^{9}	**H_β-H_N [c]**	**weak**
Y^{9}-A^{10}	**H_β-H_N**	**weak**
S^{11}-R^{12}	**H_β-H_N**	**weak**
R^{12}-G^{13}	**H_β-H_N**	**weak**
I^{14}-R^{15}	**H_β-H_N**	**strong**
P^{16}-V^{17}	**H_β-H_N**	**weak**
V^{17}-G^{18}	**H_β-H_N**	**weak**
R^{19}-F^{20}	**H_β-H_N**	**strong**
Y^{9}-S^{11}	H_N-H_N	weak
A^{10}-R^{12}	H_N-H_N	weak
S^{11}-G^{13}	H_N-H_N	weak
Y^{9}-S^{11}	H_α-H_N	weak
A^{10}-R^{12}	H_α-H_N	weak
S^{11}-G^{13}	H_α-H_N	weak
P^{16}-G^{18}	H_α-H_N	weak
I^{14}-V^{17}	H_α-H_N	weak
P^{16}-R^{19}	H_α-H_N	weak
R^{12}-R^{15}	**H_α- H_β**	**weak**
I^{14}-V^{17}	**H_α- H_β**	**weak**

[a] NOEs were classified as either "weak" (proton-proton distance ≤ 5Å) or "strong" (proton-proton distance ≤ 3Å); [b] Color key: white=d(i,i+1); light gray=d(i,i+2); dark gray=d(i,i+3); [c] Bolded text indicates that the NOE occurred between side-chain protons and was only used during folding (not fragment generation).

Peptide synthesis

PrRP20, PrRP14-20, PrRP8-20, PrRP4-20, A^{15}PrRP20, A^{19}PrRP20, and A^{20}PrRP20 were synthesized by automated multiple solid-phase peptide synthesis on the multiple peptide synthesizer Syro II (MultiSynTech GmbH, Witten, Germany) using the orthogonal Fmoc/tBu strategy.59 Rink amide resin (30 mg, resin loading 0.6 mmol·g-1), obtained from Iris Biotech GmbH (Marktredwitz, Germany), was used to produce the C-terminally amidated peptides. Nα-Fmoc (N-(9-fluorenyl)methoxycarbonyl)-protected amino acids were purchased from Iris Biotech GmbH (Marktredwitz, Germany). The protected amino acids (10eq) were dissolved in 0.5 M tert-butyl alcohol in dimethylformamide and activated in situ by diisopropylcarbodiimide (DIC) (10eq). Removal of protection groups and final cleavage of the peptide from the resin was accomplished simultaneously using a cleavage cocktail consisting of either trifluoroacetic acid (TFA)/thioanisole/1,2-ethanedithiol (90:7:3 v/v/v) for tryptophan-containing peptides or TFA/thioanisole/*p*-thiocresol (90:5:5 v/v/v) within 3 hours.

Peptide purification was achieved by preparative reversed-phase HPLC (Vydac RP18-column, 22 × 250 mm, 10 μm/300Å, Grace, Deerfield, IL, USA or Phenomenex Jupiter 10 U Proteo column, 250 × 21.20 mm, 90Å, Aschaffenburg, Germany) using 0.08% TFA in either acetonitrile or methanol (MeOH) and 0.1% TFA in water as the eluting system to yield homogenous peptides of > 90% purity. The peptides were characterized by mass spectrometry using matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry on an Ultraflex III MALDI-TOF/TOF mass spectrometer (Bruker Daltonics, Bremen, Germany). Analytical reversed-phase HPLC was performed on a Vydac RP18-column (4.6 × 250 mm; 5 μm/300 Å; Grace,

Deerfield, IL, USA) by using two different linear gradient systems of 0.1% (v/v) TFA in water and 0.08% (v/v) TFA in either acetonitrile (ACN) or methanol. Analytical data are summarized in Table 3.

Table 3: Analytics of PrRP20 used for structural and biological investigations

Peptide	Sequence	Mass $[M+H]^+$		HPLC		
		Calc.	Exp.	ACN [%]	MeOH [%]	Purity [%]
PrRP20	TPDINPAWYASRGIRPVGRF-NH_2	2272.6	2273.7	40.3[a]	65.5[b]	>99
PrRP4-20	INPAWYASRGIRPVGRF-NH_2	1959.3	1960.4	40.5[a]	66.9[b]	>99
PrRP8-20	WYASRGIRPVGRF-NH_2	1562.9	1563.9	38.3[a]	61.6[b]	>92
PrRP14-20	IRPVGRF-NH_2	842.5	843.5	33.8[a]	52.6[b]	>96
A^{15}PrRP20	TPDINPAWYASRGI<u>A</u>PVGRF-NH_2	2186.1	2187.2	42.9[a]	71.7[b]	>96
A^{19}PrRP20	TPDINPAWYASRGIRPVG<u>A</u>F-NH_2	2187.5	2188.4	41.6[a]	70.8[c]	>99
A^{20}PrRP20	TPDINPAWYASRGIRPVGR<u>A</u>-NH_2	2196.5	2196.2	37.7[a]	61.6[b]	>99

[a] 10% to 60% ACN (0.08% TFA) in water (0.1% TFA) over 30 min. [b] 20% to 100% MeOH (0.08% TFA) in water (0.1% TFA) over 40 min. [c] 30% to 100% MeOH (0.08% TFA) in water (0.1% TFA) over 30 min.

Cloning of the RF-amide peptide receptors in eukaryotic expression vectors

To obtain genomic DNA from SMS-KAN cells, approximately 1 million cells were digested overnight at 55°C with 500 µL lysis buffer (1 M NaCl, 20% SDS, 0,5 M EDTA, 1 M Tris, pH 8.5) containing 50 µg proteinase K (Promega, Mannheim, Germany). Genomic DNA was extracted using phenol/chloroform and precipitated from the aqueous phase with isopropanol, washed with ethanol, and then dissolved in water. The coding sequence of the human PrRP receptor was obtained by PCR amplification from the genomic DNA of SMS-KAN cells. Cloning of cDNA into the eukaryotic expression vector pEYFP-N1 (Clontech, Heidelberg, Germany) C-terminally fused to EYFP was performed, using the XhoI and BamHI site to result in the constructs phPrRP receptor_EYFP-N1. Mutations were introduced with the QuikChange™ site-directed

mutagenesis method (Stratagene). The residues are numbered according to the system of Ballesteros and Weinstein (38). The correctness of all constructs was confirmed by sequencing of the entire coding sequence.

Cell culture

Cell culture material was supplied by PAA Laboratories GmbH (Pasching, Austria). COS-7 cells (African green monkey, kidney) were cultured in Dulbecco's Modified Eagle's Medium containing 10% (v/v) heat-inactivated fetal calf serum (FCS), 100 units/mL penicillin and 100 µg/mL streptomycin. SMS-KAN cells (human neuroblastoma cells) were maintained in nutrient mixture Ham's F12/Dulbecco's modified Eagle medium (1:1) with 15% (v/v) FCS, 4 mM glutamine, 0.2 mM non essential amino acids, 10 units/mL penicillin, and 10 µg/mL streptomycin. Cells were grown as monolayers at 37°C in a humidified atmosphere of 5% CO_2 and 95% air.

Signal transduction assay

For signal transduction (inositol phosphate accumulation) assays, COS-7 cells were seeded into 24-well (1.0 × 105 cells/well) or 48-well plates (6.0 × 104 cells/well) and transiently transfected with 0.4 µg plasmid DNA using 1.2 µL metafectene (Biontex Laboratories GmbH, Martinsried/Planegg, Germany). Incubation with 2 µCi/mL [^{3}H]myo-inositol (GE Healthcare Europe GmbH, Braunschweig, Germany) in DMEM supplemented with 10% (v/v) FCS was performed one day after transfection and 16 h before stimulation. Labeled cells were washed once and stimulated with increasing concentrations of each peptide for 1 h at 37°C in DMEM containing 10 mM LiCl (Sigma-

Aldrich, Taufkirchen, Germany) as described previously (171, 172). Receptor stimulation and IP accumulation were stopped by aspiration of medium, and cell lysis was performed with 0.1 M NaOH (24-well plate: 150 µL/well; 48-well plate: 100 µL/well) for 5 minutes. After neutralizing with 0.2 M (24-well plate: 50 µL/well) or 0.13 M (48-well plate: 50 µL/well) formic acid, IP dilution buffer (5.0 mM Na-borate + 0.5 mM Na-EDTA; 24-well plate: 1 mL/well; 48-well plate: 750 µL/well) was added to each well.

Intracellular IP levels were determined by anion-exchange chromatography on Bio-Rad AG 1-X8 resin either by manual pipetting or using an automated pipetting robot system (USK-UTZ GmbH, Limbach-Oberfrohna, Germany). Radioactivity was measured by a scintillation counter (Win Spectral 1414 Liquid Scintillations Counter Wallac) (173, 174). Data were analyzed with GraphPad Prism 3.0 program (GraphPad Software, San Diego, USA) and EC_{50}-values were obtained from concentration response curves. The EC_{50}-determinations were performed in duplicate and signal transduction assays were repeated at least twice independently.

Radioligand binding studies

For radioligand binding studies, 1.5×10^6 COS-7 cell were seeded into 25 cm^2 flasks. At 60-70% confluency, cells were transiently transfected using 4 µg vector DNA and 15 µL of Metafectene™ (Biontex Laboratories GmbH, Martinsried/Planegg, Germany). Approximately 24 hours after transfection, binding assays were performed on intact cells using N[propiony^{l3}H]hPrRP20. Binding was determined with 1 nM N[propiony^{l3}H]hPrRP20 in the absence (total binding) or in the presence (non-specific binding) of 1 µM unlabeled hPrRP20, respectively, as described previously(172, 175).

N[propiony^{l3}H]hPrRP20 was obtained by selective labeling as described previously and resulted in a K_D-value of 0.58 nM (176). Specific binding of each PrRP receptor mutant was compared to specific binding of the PrRP wt receptor. IC_{50}-values and K_D-values were calculated with GraphPad Prism 3.0 (GraphPad Software, San Diego, USA), fitted to a one-site competition or a one-site binding model, respectively. Each experiment was performed in triplicate.

CD spectroscopy

CD measurements of 40 µM peptide solutions buffered with 10 mM phosphate buffer at pH 5.5 or 7 were performed in the far ultraviolet region from 250 to 190 nm using a Jasco J-715 spectropolarimeter. Additionally, CD spectra of 10 mM phosphate buffered peptide solutions were measured in either 25% TFE or 100 mM SDS-containing solutions. Cuvettes with 2 mm path length (quartz cuvette; Hellma, Jena, Germany), as well as the following parameters, were used: 50 nm·min^{-1} scanning speed, 4 s response, 0.2 nm step resolution, 2 nm bandwidth, temperature of 22°C. Peptide concentration was determined from the aromatic spectrum determined in aqueous solution and calculated using the molar extinction coefficient of the peptides at 280 nm (6990 M^{-1} cm^{-1}). For PrRP14-20, the pure lyophilized peptide was weighed and diluted to 40 µM, considering that the final peptide mass results from the salt with TFA as counterion for both arginine residues. Spectra were measured in a constant nitrogen stream of 15 L·min^{-1}. The final spectra were averaged from 6 to 9 baseline-corrected scans without any smoothing. The raw CD signal [mdeg] was converted to mean residue ellipticity, [Θ], by $[\Theta] = [\Theta]_{observed}(MRW/l \cdot c \cdot 10)$, where *MRW* is the mean residue weight (molecular mass divided

by number of peptide bonds), *l* is path length [cm] and *c* is the concentration of peptide in mg/mL. Graphs were processed using GraphPad Prism 3.0 program (GraphPad Software, San Diego, USA), Microsoft Excel 2011™, as well as with the Jasco-715 spectropolarimeter-related Jasco software.

Results

Previous NMR studies of C-terminally amidated PrRP20 reveal a helical C-terminal region.

RosettaNMR (106, 136, 168) was employed to construct a model of C-terminally amidated PrRP from 38 previously reported inter-proton distances (NOEs) and 13 H_α chemical shifts, which were collected at pH 5.5 in 100 mM sodium dodecyl sulfate, or SDS (Table 1 and Table 2) (35). These NMR data were obtained for PrRP20. We chose to construct structural models for residues 8-20 of PrRP20 because the structural restraints cover mainly these residues, implying that residues 1-7 are conformationally flexible. However, because only a partial dataset was available, the herein discussed peptide model ensemble serves only as a starting point for further structural characterization of the PrRP/PrRP receptor interaction. The generated models were further confirmed with CD spectroscopy (see *Structural investigations of PrRP by CD spectroscopy studies indicate a decreased helical propensity for PrRP8-20*).

The NOEs and chemical shifts occurring within residues 8-20 are indicative of a combination of α- and 3_{10}-helical secondary structure. The presence of αN(i,i+2) NOEs is often associated with i(i+3) hydrogen bonding characteristic of 3_{10}-helices. Further, the ratio of αβ(I,i+3) to αN(i,i+3) NOEs, as well as the lack of αN(i,i+4) NOEs, support the

idea that the peptide exists in an equilibrium of α- and 3_{10}-helix in SDS micelles (177-179) (see D'Ursi *et al*. for original figures). An ensemble of twenty low-energy models of the PrRP20 residues 8-20 consistent with the NMR data obtained for the full-length peptide was generated and deposited in the Protein Model Database (180) (Figure 4 PMID: 0078404).

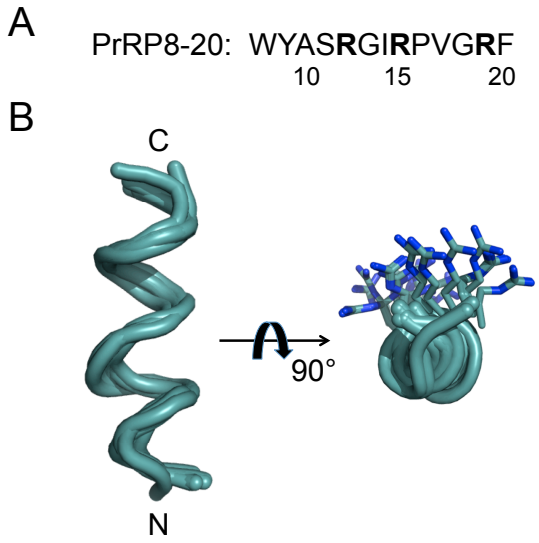

Figure 4: The conformational ensemble of PrRP8-20 generated using RosettaNMR
A) The primary sequence of PrRP8-20. The three arginines are in bold. B) The twenty lowest-energy models resulting from full-atom refinement that had a RosettaNMR restraint score ≤ 1.0 Rosetta Energy Unit (REU). Briefly, ten thousand models were *de novo* folded in the presence of 38 distance restraints. Energetically favorable models that satisfied the NMR data were then refined to atomic detail using the same 38 restraints. Notice that all three arginine residues are on one side of the amphipathic helix.

Secondary structural analysis of PrRP20 models implies a conformational equilibrium.

The final ensemble of PrRP models was chosen based on the models' overall energy according to the RosettaNMR full-atom soluble protein scoring function (110), as well as their agreement with the NMR distance restraints for the full-length peptide (Table 4 and Table 5) (35). Define Secondary Structure of Proteins (DSSP)(181, 182) analysis indicates that these models are mainly α-helical, especially between residues 10-

13 and 15-19, with the other residues being coil or bend/turn (Figure 5A). Note the often-observed non-ideal helical character around residue I^{14}. This is likely due to the inability of the nitrogen of P^{16} to hydrogen bond with the carbonyl oxygen on R^{12} (distance = 4.98 ± 0.27Å), thus disrupting the hydrogen bond between G^{13} and V^{17} (distance = 5.00 ± 0.26Å) (Figure 5B). The models exhibit ϕ and ψ angles (torsion angles around the N-C_α bond and the C_α-C bond, respectively) characteristic of both α- and 3_{10}-helix, where α-helices have an average ϕ angle of −57° and an average ψ angle of −70°. 3_{10}-helicies typically have average ϕ angles of approximately −49° and average ψ angles of −26° (Figure 5C) (183-185). Interestingly, residues 10-13 appear to usually form an α-helical turn, but they can also adopt a 3_{10}-helical structure (Table 6, Models 10 and 11). Furthermore, the DSSP secondary structure analysis reveals that approximately 15% of all models *de novo* folded and refined with RosettaNMR contained both α- and 3_{10}-helical conformation, but the majority of models were primarily α-helical (Figure 6). These results match D'Ursi *et al*.'s NOE data, which support an unambiguously α-helical C-terminal region (residues 15-19), whereas the N-terminus of PrRP20 appeared to be in a conformational equilibrium, fluctuating between α-helix, 3_{10}-helix, and nascent helix or coil. It is noteworthy that the new ensemble agrees well with D'Ursi *et al*. considering the sparseness of the available data, which recapitulates RosettaNMR's sampling efficiency.

Table 4: Statistics for restraints, structural calculations, and structural quality for final ensemble of PrRP models

NMR distance restraints used during folding and refinement	
Total restraints	51
Chemical shifts[a]	13
Distance restraints	
Total NOE	38
Intra-residue	0
Inter-residue	38
Sequential ($\|i-j\| = 1$)	27
Medium-range ($\|i-j\| < 5$)	11
Long-range ($\|i-j\| \geq 5$)	0
Structural statistics	
Violations of distance restraints (Å)[b]	0.08 ± 0.11
Deviations from idealized geometry	
Bond lengths (Å)	0.024
Bond angles (°)	0.8
Main chain RMSD to the mean structure (Å)	0.83
Average main chain pairwise RMSD (Å)	1.10
Ramachandran plot statistics (%)	
Most favored regions	85.6
Additionally allowed regions	14.4

[a] Chemical shifts were used only during fragment generation and were not used during *de novo* folding and refinement. [b] For analysis, a violation of a restraint was counted if the inter-proton distance was > 5.5Å (weak) or 3.5Å (strong). There are no distance restraint violations greater than 0.4Å.

Table 5: NMR distance restraints violated by final ensemble of PrRP models

Residue Pair	**NOE Type**	**NOE Strength[a]**	**Number of Models Violating[b] This Restraint**	**Average Violation Distance** (Å)
17-18	H_α-H_N	strong	14	0.06 ± 0.002
19-20	H_β-H_N	strong	6	0.30 ± 0.11

[a] During folding and refinement, NOEs were classified as either "weak" (proton-proton distance ≤ 5Å) or "strong" (proton-proton distance ≤ 3Å). [b] For analysis, a violation of a restraint was counted if the inter-proton distance was > 5.5Å or 3.5Å, respectively).

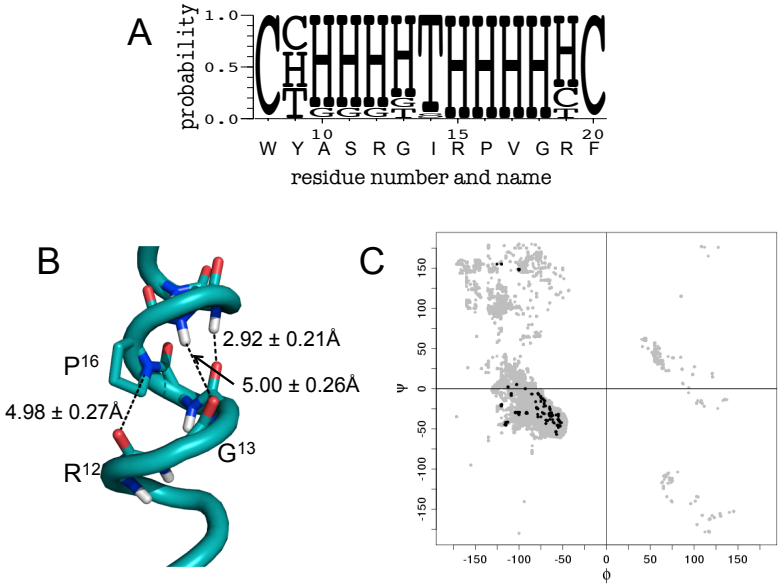

Figure 5: Evidence of helical secondary structure in the PrRP ensemble of models
A) WebLogo (http://weblogo.berkeley.edu/) summarizing the consensus secondary structure information obtained by DSSP; C=coil, T=turn, H=α-helix. B) Close-up view of backbone interactions between residues 12-18. Oxygens are colored in red, nitrogens in blue, and hydrogens in white. C) Ramachandran plot of ϕ/ψ angles of the models as computed by DSSP; gray=angles obtained for all models generated; black=angles for final ensemble.

Table 6: Secondary structure analysis of PrRP8-20 models by DSSP[a] analysis

State	Secondary Structure	# Residues α-Helix (H)	# Residues 3_{10}-Helix (G)	# Residues Turn (T)	# Residues Bend (S)	# Residues Coil (-)
1	--HHHHSHHHHH-	9	0	0	1	3
2	-HHHHHTHHHHH-	10	0	1	0	2
3	-HHHHHTHHHHH	10	0	1	0	1
4	-HHHHHTHHHHH-	10	0	1	0	2
5	-HHHHHTHHHHH-	10	0	1	0	2
6	-THHHHTHHHHH-	9	0	2	0	2
7	-THHHHTHHHHH-	9	0	2	0	2
8	-THHHHTHHHHH-	9	0	2	0	2
9	-THHHHTHHHHH-	9	0	2	0	2
10	--GGGGTHHHHT-	4	4	2	0	3
11	--GGGGTHHHHT-	4	4	2	0	3
12	-THHHHTHHHHH-	9	0	2	0	2
13	-HHHHHTHHHHH-	10	0	1	0	2
14	--HHHHTHHHH--	8	0	1	0	4
15	--HHHHTHHHH--	8	0	1	0	4
16	--HHHHTHHHH--	8	0	1	0	4
17	--HHHHTHHHH--	8	0	1	0	4
18	-THHHHTHHHHH-	9	0	2	0	2
19	-HHHHTTHHHHH-	9	0	2	0	2
20	-HHHHTTHHHHH-	9	0	2	0	2

[a] For more information, go to http://swift.cmbi.ru.nl/gv/dssp/. In this case, H=alpha-helix, G=3-10 helix, S=bend, T=hydrogen-bonded turn, (-)=random coil.

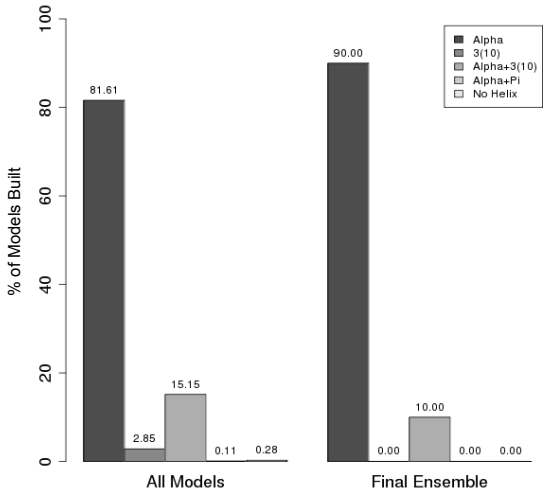

Figure 6: Secondary structure of PrRP models generated with RosettaNMR
DSSP was used to determine the secondary structural make up of all models (left) generated, as well as the final ensemble submitted to the Protein Model Database (right). All models contained at least 1-2 residues having non-helical character (i.e., unstructured, turn, bend, etc.).

Structural investigations of PrRP by CD spectroscopy indicate a decreased helical propensity for PrRP8-20.

To elucidate the structural and functional requirements for PrRP20 binding and receptor activation, a set of PrRP analogs was synthesized and characterized (Table 3). Because the C-terminal region of the peptide is presumably responsible for receptor binding and activation (28, 30, 34, 36), we focused primarily on N-terminal truncation of PrRP20 to PrRP4-20, PrRP8-20, and the shortest reported full agonist, PrRP14-20 (30). CD spectra of PrRP20 and PrRP4-20 recorded in aqueous phosphate buffered solution at pH 7.0 and 22°C show significantly more intense signal between 200-230 nm in comparison to PrRP14-20, which is expected to be flexible and mostly disordered. Further, the CD spectrum of PrRP8-20 in phosphate buffer also suggests a primarily disordered peptide; the slight maximum at approximately 228 nm suggests the presence of some poly-proline II helix conformation as well (186, 187) (Figure 7, left panel).

Interestingly, according to the spectra of PrRP20 and PrRP4-20, the peptides may contain some ordered secondary structural character, including 3_{10}-helix (Table 7); note the deep minima at ~205 nm and the shoulder at ~222 nm. This is also supported by the peptides' $R_{222/208}$ values of 0.46 ± 0.01 and 0.37 ± 0.02, respectively. According to Toniolo *et* al., this ratio is expected to be between 0.15 and 0.40 for 3_{10}-helical peptides and ~1.0 for α-helical peptides (188, 189).

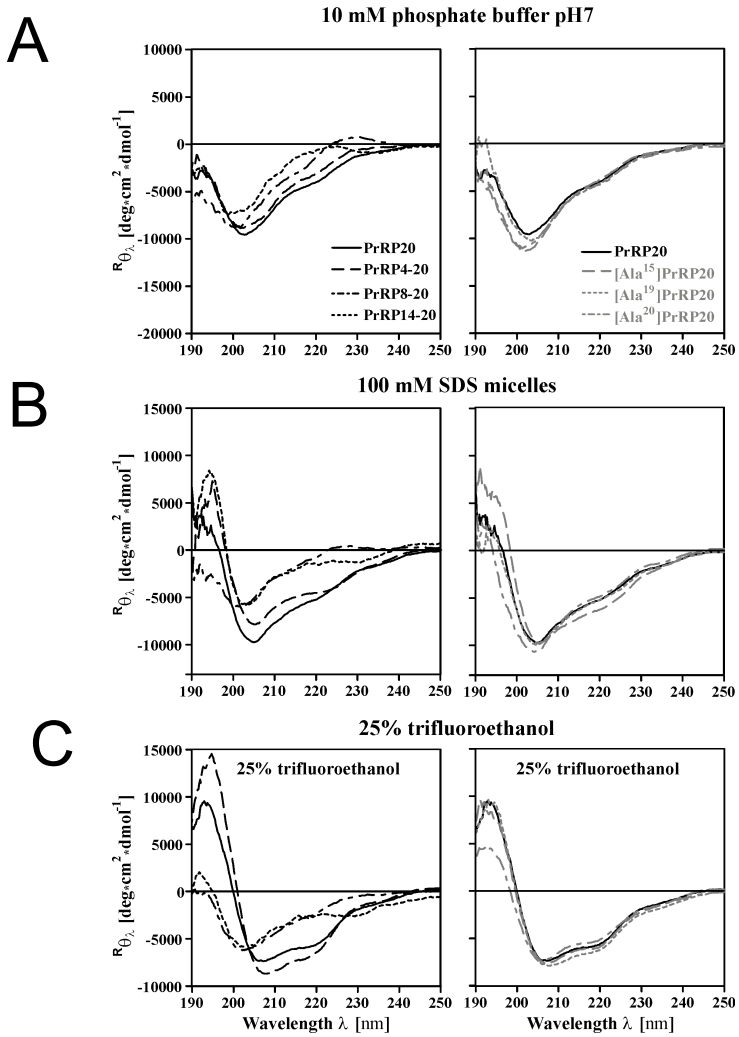

Figure 7: Influence of different solvents on the structure of wildtype and mutant PrRP
Left panel: Truncation mutants of PrRP20 (PrRP4-20, PrRP8-20, and PrRP14-20). Right panel: Single-mutant PrRP20 analogs (A^{15}PrRP20, A^{19}PrRP20, and A^{20}PrRP20). CD spectra are represented in mean residue ellipticity, measured in 40 µM peptide in 10 mM phosphate buffered solution at pH 7 and 22°C. (A) CD spectra measured without additives, (B) in 100 mM micellar SDS solution, and (C) 25% TFE-containing solution. All curves were calculated with the baseline corrected for buffer effects.

Table 7: Characterization of CD data

Peptide	Condition	Ratio [R] ± SD
	pH 7, 10 mM pb	$\frac{[\theta]222}{[\theta]208}$
PrRP20	aqueous	0.46 ± 0.01
PrRP4-20	aqueous	0.37 ± 0.02
PrRP8-20	aqueous	NC
PrRP14-20	aqueous	NC
PrRP20	100 mM SDS	0.54 ± 0.01
PrRP4-20	100 mM SDS	0.63 ± 0.01
PrRP8-20	100 mM SDS	NC
PrRP14-20	100 mM SDS	0.40 ± 0.05
PrRP20	25% TFE	0.68 ± 0.01
PrRP4-20	25% TFE	0.65 ± 0.01
PrRP8-20	25% TFE	0.45 ± 0.08
PrRP14-20	25% TFE	0.54 ± 0.03

pb = phosphate buffered; SD = standard deviation; NC = not considered for reasons of missing characteristic helical CD spectra.

Next, we investigated the peptide in solvents mimicking the partially apolar membrane environment while retaining a certain biocompatibility. We will label the three experimental conditions as "aqueous," "SDS," and "TFE" throughout the remainder of the manuscript. For PrRP20 tested in 100 mM SDS solution, a well-known membrane mimicking detergent, we observe a maximum at 195 nm, a minimum at 205 nm, and a shoulder around 222 nm (Figure 7B, left panel); the latter two spectral features are indicative of a 3_{10}-helical component to the conformational ensemble. The characteristic minima for solely α-helically structured peptides are at 208 nm and 222 nm (190). However, the $R_{222/208}$ value of 0.54 ± 0.01 is higher than expected for a pure 3_{10}-helix. We therefore conclude that, in SDS, PrRP20 adopts a partially α-helical conformation, with 3_{10}-helix and other secondary structural components also being present. Similar observations were observed for PrRP4-20 ($R_{222/208}$ = 0.63 ± 0.01). The CD spectra of

PrRP14-20 has 3_{10}-helix character, ($R_{222/208}$ = 0.40 ± 0.05), whereas PrRP8-20 appears to remain primarily coil/poly-proline II helix under these conditions (Figure 7B, left panel; Table 7).

Fluorinated alcohols, such as trifluoroethethanol, or TFE, are organic solvents that induce environmental constrains; TFE/water mixtures exhibit helix-inducing biocompatible conditions. For CD spectroscopy of PrRP20 and PrRP4-20 measured in TFE/water, $R_{222/208}$ values of 0.68 ± 0.01 and 0.65 ± 0.01, respectively, were calculated. These values support the assumption that the peptides are primarily α-helical (Table 7). Indeed, in TFE/water, the helical content of the full-length peptide increased, with the spectrum exhibiting deep minima at 208 nm and 222 nm. These minima are more pronounced than those seen in the CD spectra obtained in SDS micelles. In contrast, the spectra of PrRP8-20 and PrRP14-20 in TFE are more reminiscent of that of a mixture of helices with a strong 3_{10}-helix component. Both peptides exhibit a minimum at approximately 202 nm and a shoulder at about 220 nm (Figure 7C, left panel). Further, the $R_{222/208}$ values for these peptides were 0.45 ± 0.08 and 0.54 ± 0.03, respectively (Table 7). The experiments in TFE were repeated at various pH-values and temperatures of 25°, 37°, and 50° for both PrRP20 and PrRP8-20 in order to confirm that the spectra were largely independent of these parameters (Figure 8).

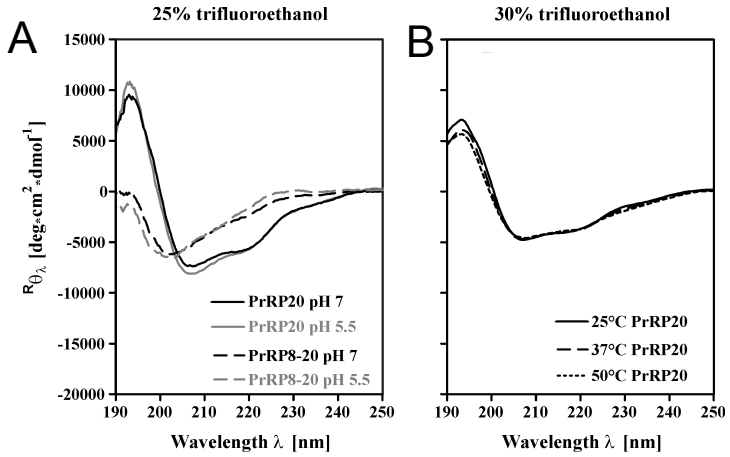

Figure 8: Structural effects of pH and temperature
CD spectra were recorded from 190–250 nm with 40 µM PrRP8-20 and PrRP20 in 10 mM phosphate buffered solution (*Materials and methods*), and mean residue ellipticity was calculated. (A) Measurement performed at pH 7 and 5.5. (B) PrRP20 was tested at different temperatures and showed no change.

Single-substituted PrRP20 analogs do not exhibit different secondary structure from wt PrRP20.

Single alanine mutants of PrRP20 at R^{15}, R^{19}, and F^{20} positions have been previously implicated with peptide activity (30, 34, 36) and were also tested here. Note that the highly conserved C-terminal residues, R^{19} and F^{20}, make PrRP a member of the RF-amide peptide family. To study the influence of the conserved RF-amide motif and the impact of charged amino acids at the hydrophilic side of the helix on the overall peptide structure, we performed CD spectroscopy on A^{15}PrRP20, A^{19}PrRP20, and A^{20}PrRP20 compared to wt PrRP20 (Figure 7, right panel; Table 7). Interestingly, all tested conditions (aqueous, SDS, TFE) resulted in almost identical CD spectra for PrRP20 and all alanine mutants. Although CD spectroscopy is not sensitive to identify small, local rearrangements in the peptide, we conclude that the modified single side-chains at positions 15, 19, and 20 have no impact on the *overall* secondary structure of

the peptide. Therefore, any loss of activity when interacting with the receptor results from a change in the interaction with the receptor rather than a change in structure or dynamics of the peptide (see *Binding to and activation of the wt PrRP receptor is primarily mediated by direct interactions with PrRP*).

Binding to and activation of the wt PrRP receptor is primarily mediated by direct interactions with PrRP.

To evaluate the biological relevance of the PrRP20 analogs, binding and signal transduction capabilities were investigated in COS-7 cells transiently transfected with the PrRP receptor. In a displacement assay with the wt PrRP receptor using 1 nM N[propionyl^{3}H]hPrRP20, an IC_{50}-value of 4.1 ± 0.7 nM was obtained, where IC_{50} is the inhibition concentration of the ligand at half maximum biological activity of the receptor. A dissociation constant, or K_D, value of 0.58 nM was computed using established methods(191). The activity of PrRP20 was determined using an IP, or inositol phosphate, accumulation assay (see *Materials and methods*) and resulted in an EC_{50}-value of 2.2 ± 0.3 nM (Table 8). The EC_{50} is the effective concentration of the ligand at half maximum biological activity.

Table 8: Effects of mutation of PrRP on binding and signaling

Peptide	Binding Assay		Signal Transduction Assay	
	IC_{50} [nM][a]	x-fold[b]	EC_{50} [nM][c]	x-fold[d]
PrRP20	4.1 ± 0.7	1	2.2 ± 0.3	1
PrRP4-20	1.2 ± 0.1	0.3	1 ± 0.2	0.5
PrRP8-20	7 ± 1.8	1.7	2.3 ± 0.5	1
PrRP14-20	430 ± 16	105	14 ± 2	6
A^{15}PrRP20	882 ± 376	215	49 ± 12	22
A^{19}PrRP20	> 10000	> 2440	1198 ± 231	545
A^{20}PrRP20	870 ± 288	212	20 ± 5	9

Values are the standard deviation (± SD) of parameters deduced by using GraphPad Prism 3.0 software. IC_{50} and EC_{50} values were obtained from resulting concentration-response curves. All signal transduction assays were performed in duplicates and repeated at least twice independently. [a] COS-7 cells were transiently transfected with PrRP receptor. The IC_{50} value was determined by competition assays using N[propionyl-^{3}H]hPrRP20. [b] Ratios with respect to the IC_{50} values of wt peptide: IC_{50} (peptide)/IC_{50} (PrRP20). [c] COS-7 cells were transiently transfected with wt hPrRP receptor. EC_{50} values were obtained from IP accumulation assay. [d] Ratios with respect to the EC_{50} values of wt peptide: EC_{50} (peptide)/ EC_{50} (PrRP20).

The radioligand binding assays revealed IC_{50}-values of 1.2 ± 0.1 nM and 7 ± 1.8 nM for PrRP4-20 and PrRP8-20, respectively. These values are comparable to PrRP20 (4.1 ± 0.7 nM, The heptapeptide, PrPR14-20, exhibited a 105-fold reduction in binding compared to PrRP20. Loss of binding was even more dramatic in the single mutant analogs: an IC_{50}-value of 870 ± 288 nM was obtained for A^{20}PrRP20, whereas for A^{19}PrRP20, no IC_{50}-value could be determined for concentrations of up to 10 µM of the ligand. A^{15}PrRP20 behaved similarly to A^{20}PrRP20, resulting in a 215-fold decrease in binding (Table 8).

In the signal transduction assays with the wt receptor, A^{19}PrRP20 revealed a 545-fold increase in EC_{50}-values (1198 ± 231 nM) over unmodified PrRP20 (2.2 ± 0.3 nM). A^{20}PrRP20 and A^{15}PrRP20 had a lower impact in IP accumulation. The EC_{50}-values were only 9- and 22-fold increased compared to the unmodified PrRP20, respectively. Apart

from PrRP14-20, which exhibited a 6-fold increased EC_{50}-value of 14 ± 2 nM, the truncated analogs, PrRP4-20 and PrRP8-20, showed wildtype-like signaling properties (Table 8).

PrRP8-20's is unable to activate extracellular loop 1 PrRP receptor mutants.

Next, we investigated the interaction of PrRP8-20 and PrRP20 with different receptor mutants. Because extracellular loop 1, referred to as EL1 for the remainder of this discussion, of other peptide receptors is known to be important for interactions with the ligands (192, 193), we assumed that charged or aromatic amino acids of the EL1 region may be involved in ligand recognition via hydrophobic, ionic, or π-cationic interactions. Therefore, we substituted all such residues between position 2.64 and 2.73 to alanine (Table 9). The single-substituted $F^{2.66}A$, $E^{2.67}A$, $R^{2.69}A$, and $F^{2.73}A$ receptor mutants behaved like wt PrRP receptor after treatment with PrRP20 in an IP accumulation assay. However, $Y^{2.64}A$ and $W^{2.71}A$ PrRP receptor variants resulted in significantly increased EC_{50}-values when stimulated with PrRP20 (50 ± 7.5 nM and 593 ± 78 nM, respectively). Stimulation of receptor mutants $Y^{2.64}A$ and $W^{2.71}A$ with PrRP8-20 revealed a further right-shifted concentration-response curve when compared to activation with PrRP20 (Figure 9) and hence elevated EC_{50}-values (434 ± 96 nM and 2119 ± 390 nM, respectively, Table 9). We hypothesized that changes in structure or dynamics of the ligand might cause this difference in receptor activation, as mutation/deletion studies of residues 1-7 did not suggest a direct contact point between this part of the ligand and the receptor.

Table 9: Signaling properties of PrRP8-20 with respect to PrRP receptor

Receptor Mutants	PrRP20		PrRP8-20	
	EC_{50} [nM][a]	x-fold[b]	EC_{50} [nM][a]	x-fold[b]
wt PrRP receptor	2.2 ± 0.3	1	2.3 ± 0.5	1
$Y^{2.64}A$	**50 ± 7.5**	**23**	**434 ± 96**	**197**
$F^{2.66}A$	6.2 ± 3.3	3	NT	-
$E^{2.67}A$	7.2 ± 3.4	3	NT	-
$R^{2.69}A$	4.2 ± 2.5	2	NT	-
$W^{2.71}A$	**593 ± 78**	**270**	**2119 ± 390**	**963**
$F^{2.73}A$	4.4 ± 2	2	NT	-

NT = not tested; Values are the standard deviation (± SD) of parameters deduced by using GraphPad Prism 3.0 software. EC_{50} values were obtained from resulting concentration-response curves. All signal transduction assays were performed in duplicate and repeated at least twice independently. [a] COS-7 cells were transiently transfected with wt hPrRP receptor. EC_{50}-values were obtained from IP accumulation assay. [b] Ratios with respect to the EC_{50} values of wt peptide: EC_{50} (peptide)/ EC_{50} (PrRP20).

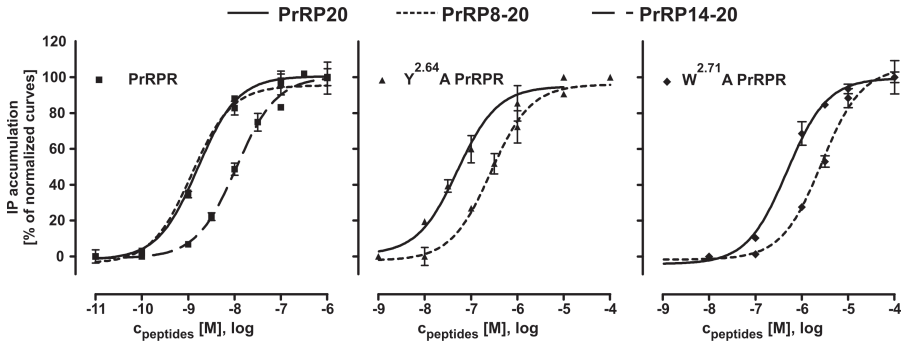

Figure 9: IP accumulation of PrRP and truncated analogs test at PrRP receptor mutants
COS-7 cells were transiently transfected with DNA coding for the wt, $Y^{2.64}A$, or $W^{2.71}A$ receptor. The signal transduction assay was performed with PrRP20, PrRP8-20, as well as with PrRP14-20 for wt PrRP receptor. All experiments performed with PrRP8-20 lead to a significantly right shifted curve, whereas PrRP8-20 behaves like PrRP20 with respect to wt receptor.

Discussion

Structure-activity/affinity studies are needed to understand PrRP receptor activation.

The objective of this study is to better understand the structural determinants of PrRP receptor activation, an important milestone towards the development of potent small-molecule agonists given the increasing prevalence for the physiological role of PrRP20 and its receptor (171). This is a formidable challenge, as structure-activity relationship studies of PrRP/PrRP receptor system are rare. Initial investigations of the truncated PrRP20 analogs, PrRP4-20 and PrRP8-20, exhibited wildtype-like binding and IP accumulation behavior. Further, in our assay system, a reduced affinity of the full agonist, PrRP14-20, is in accordance with recent studies (30, 34). We hypothesized that the structure and dynamics of PrRP's interaction with the receptor is altered through the truncation, rather than single point mutation, of the peptide. This hypothesis was tested through CD and NMR spectroscopic studies that assert the secondary structure of the peptide. To mimic the amphipathic environment of the peptide when it is interacting with the receptor, the additives SDS and TFE were used (194, 195).

CD and NMR spectroscopic studies support a mainly helical peptide conformation.

While SDS is an accepted membrane mimic, TFE mainly induces secondary structure (196). SDS micelles provide a non-isotropic, apolar environment in which the membrane interactions of the biomolecules can be investigated. A molecular dynamics study has shown that, in a TFE/water mixture, the organic co-solvent aggregates around the peptide, forming a matrix that partly excludes water. This process stabilizes the secondary structure, as the formation of proximate interactions is assisted.[38] We suggest

that, to some extent, both solvents mimic the membrane surface thought to contribute to the transition of the peptide from a random coil to a helical conformation that is recognized by the receptor (197). Accordingly, we assume that PrRP20 will adopt a conformation more similar to the bioactive form when interacting with these solvents.

According to our CD spectroscopic studies, the single mutant PrPR20 analogs, A^{15}PrRP20, A^{19}PrRP20, and A^{20}PrRP20, fully maintained their PrRP20-like $\alpha/3_{10}$-helical conformation in SDS and TFE. This is especially remarkable because all of them display significantly reduced binding and signaling properties with respect to the wt receptor. It is noteworthy that the binding and signaling studies herein are in agreement with recently published structure-activity studies that describe the importance of R^{15}, as well as the RF-amide motif (30, 34). PrRP20 and PrRP4-20, while exhibiting some 3_{10}-helical character in phosphate buffer, became increasingly α-helical in SDS and TFE. In contrast, PrRP8-20 appears to be primarily disordered, or nascent helix at most, in SDS. Its 3_{10}-helix component does increase in TFE, but it is almost undoubtedly not an α-helix, unlike the full-length peptide. Our results indicate that the peptide length of PrRP is a significant determinant in its ability to form an α-helix. It appears that the N-terminus, which exhibits increased flexibility, is nevertheless involved in stabilizing the C-terminal helical segment. Even though PrRP8-20 fully activates the wt receptor (Figure 9), it shows little α-helical propensity in SDS and TFE when compared to PrRP20.

Earlier CD studies could not clearly distinguish between 3_{10}- and α-helical peptide structures, which were investigated using a set of seven peptides ranging in length from 10 to 21 amino acid residues (198). In contrast, a recent report describes the standard CD spectrum of a 3_{10}-helial octapeptide (188). Indeed, evidence of this combination of coil

and helical secondary structure can be seen in the CD spectra of the PrRP analogs, which were collected in SDS micelles or TFE (Figure 7, left panel). The shape of PrRP20 and PrRP4-20 in TFE fits to the former described spectrum for an α-helical peptide (188, 199). In the case of PrRP8-20, the membrane-mimicking SDS micelles are not capable of inducing α- or 3_{10}-helical conformation, in contrast to the longer peptides. For the analogs PrRP8-20 and PrRP14-20, the shape of the curves is altered, having a lower Cotton effect and different minima.

The combination of CD and computational modeling results, as well as analysis of the 13 C-terminal residues of PrRP20, imply a structural model for the full-length peptide, in which the peptide forms an extended helix. According to a secondary structure analysis of the final ensemble with DSSP (181, 182), it appears that, in most low-energy RosettaNMR models, 8-9 of the 13 residues tend to be α-helical. The ideal helical geometry is broken around residue I^{14}. This is expected due to the lack of ideal α-helix hydrogen bonding between R^{12} and P^{16}, as well as between G^{13} and V^{17}. In our model, the helix bulges and bends in this area. Interestingly, the helical character of the models can either consist of all α-helix or a combination of approximately half α-helix and half 3_{10}-helix (Table 6, Models 10 and 11); this matches observations from CD investigations of PrRP20 and PrRP4-20, as well as the combination of i(i+2) and i(i+3) NOEs obtained by D'Ursi *et al*. on PrRP20 (Table 2). According to these data, PrRP20 in solution is not solely α-helical, nor is it completely random coil.

The presence of potential 3_{10}-helical character in the PrRP20 models may be a result of its amphipathic nature and the fact that the NMR data were also collected in SDS micelles at high PrRP20 concentration (0.5−15 mM) (35). Indeed, there is evidence

that amphipathic helices can assume extended (often 3_{10}) helical conformations in certain mediums, such as in detergent micelles (200, 201). Remarkably, it has been proposed that the $R_1xxR_2xxR_3xxR_4xxR_5xxR_6$ motif in the Kv1.2- and Kv2.1-chimeric potassium ion channel structures form an extended 3_{10}-helix, which allows the arginine residues to sit on the same side of the helix (202, 203). This is also often observed in our models of PrRP, which contains three arginine residues in an $R_1xxR_2xxxR_3$ motif. Further, the conformational equilibrium between nascent, α-, and 3_{10}-helix is seen in other systems. Another neuropeptide, the galanin-like peptide (GALP) has been shown to be only loosely ordered in solution, but in TFE, it forms stable helical structures. Indeed, its CD spectrum resembles that of a 3_{10}-helix and is similar to our CD spectra obtained for PrRP20 in buffer and SDS and PrRP8-20 in TFE (204). The 16 amino acid sequences of the C-terminal helices of two bacterial cytochromes were synthesized and characterized by CD and NMR spectroscopy. These peptides' spectra also imply a dynamic equilibrium between α- and 3_{10}-helix (205). It is possible that this conformational equilibrium is due to folding and unfolding of the free (as opposed to receptor-bound) peptide in solution; the 3_{10}-helix is often considered to be a kinetic intermediate when forming an α-helix from coil (184, 206, 207).

Receptor residues $Y^{2.64}$ and $W^{2.71}$ may induce ligand helicity and facilitate binding and activation.

To further elucidate the role of N-terminal PrRP20 truncations with respect to ligand binding, we chose to study the EL1 of the receptor because this region is known to be a prominent agonistic binding region in GPCRs. With respect to receptor activation, the alanine scan of selected amino residues within EL1 of the PrRP receptor identified the

aromatic residues $Y^{2.64}$ and $W^{2.71}$ to be important. Both residues might contribute to a hydrophobic cluster, as described for the neurotensin receptor 1, where EL1 is described to be stabilized by π-stacking clusters and was proved to be important for agonist binding (208). In addition, $Y^{2.64}$ in particular has already been identified to participate in ligand binding in the Y1 receptor (209) and is thought to be part of a formed cluster in the binding-site crevice at the aminergic GPCR (210). PrRP20 stimulation resulted in increased EC_{50}-values in $Y^{2.64}$A and $W^{2.71}$A PrRP receptor mutants. This fits to the reported ligand-binding and receptor-activating role of EL1 in GPCRs (192, 193, 211). In particular, $W^{2.71}$ is located in the previously described WxGF-motif (212), which is necessary for receptor activation. Activation by a ligand occurs most likely by inducing movement of the transmembrane helices. While PrRP20 and PrRP8-20 exhibit identical potency for the wt receptor, PrRP8-20 was less potent at the$Y^{2.64}$A or $W^{2.71}$A PrRP receptor.

Combining these findings, we expect that the receptor assists PrRP in forming its bioactive α-helical conformation. This conformation is induced by the wt receptor for PrRP20, as well as PrRP8-20, even though its α-helical propensity is reduced due to the missing residues 1-7. However, the mutations $Y^{2.64}$A and $W^{2.71}$A partially impair the helix-inducing capabilities of the receptor. This leads to a reduced activity for both peptides PrRP20 and PrRP8-20. The reduced helical propensity of PrRP8-20 results in a more dramatic loss of activity for its interaction with the mutant receptors. The results obtained from our structure-activity and spectroscopic studies suggest that $Y^{2.64}$ and $W^{2.71}$ provide part of the hydrophobic framework that induces helicity in the ligand.

Conclusion

The C-terminal segment of PrRP20 was shown by NMR and CD spectroscopy to adopt a combination of α- and 3_{10}-helical conformation in SDS micelles and becomes primarily α-helical in TFE. Moreover, the decreased stability of the helical segment generated by shorter PrRP20 analogs resulted in reduced biological activity. In contrast, single amino acid replacement of crucial residues led to significantly decreased binding and activity, while the overall peptide structure was maintained. With respect to future structure/activity studies, we disclose that a stable C-terminal α-helix facilitates the ligand recognition by its receptor. By making a three-dimensional structure of PrRP publicly available, the structure-function studies can now be performed more effectively with the ability to look at the structure of the peptide itself. Additionally, the identification of the important residues $Y^{2.64}$ and $W^{2.71}$ with respect to ligand binding and receptor activation offers an initial step, as comprehensive structure/activity studies are rare and no antagonist of the PrRP receptor is known. Due to the involvement of PrRP20 in energy and body weight homeostasis and food intake, it provides a remarkable target for future drugs (171). The Cartesian coordinates of the ensemble of structures of the PrRP20 C-terminal segment discussed herein has been included in the Supplementary Information, as well as deposited in the Protein Model Database (PMID: PM0078404) for other researchers to use to further their own studies.

Acknowledgements

We thank D'Ursi *et al*., the authors of the original report of the structure of PrRP, the NMR data without which we could not determine the structure with RosettaNMR. We

also thank the members of the RosettaCommons for their assistance, especially Oliver Lange (folding with NMR restraints) and Steven Combs (addition of C-terminal amidation capability). Financial support of the DFG to ABS (SFB 610, BE 1264-11) and NIH to JM (R01 MH090192, R01 GM GM080403) is kindly acknowledged.

CHAPTER III

INTEGRATING SOLID STATE NMR AND COMPUTATIONAL MODELING TO INVESTIGATE THE STRUCTURE AND DYNAMICS OF MEMBRANE-ASSOCIATED GHRELIN

This work is based on the manuscript submitted to PLoS ONE of the same title by Gerrit Vortmeier*, Stephanie H. DeLuca*, Sylvia Els-Heind, Constance Chollet, Holger A. Scheidt, Annette G. Beck-Sickinger, Jens Meiler, and Daniel Huster. *These authors contributed equally.

Summary

The peptide hormone ghrelin activates the growth hormone secretagogue receptor 1a (GHS-R1a), also known as the ghrelin receptor. This 28-residue peptide is acylated at Ser^3 and is the only peptide hormone in the human body that is lipid-modified. Little is known about the structure and dynamics of membrane-associated ghrelin. We carried out solid-state NMR studies of ghrelin in lipid vesicles, followed by computational modeling of the peptide using Rosetta. Spin diffusion experiments of ghrelin indicate that the peptide binds to membranes via its lipidated Ser^3. Further, Phe^4, as well as electrostatics involving the peptide's positively charged residues and lipid polar headgroups, may contribute to the binding energy. Other than the lipid anchor, ghrelin is highly flexible and mobile in solution. This observation is supported by our model, which is in good agreement with experimentally determined chemical shifts. In the final ensemble of models, residues 8-17 form an α-helix, while residues 21-23 and 26-27 often adopt a

polyproline II helical conformation. These helices appear to assist the peptide in forming an amphipathic conformation so that it can bind to the membrane.

Introduction

Ghrelin, a 28-amino acid peptide hormone, is the endogenous ligand of the growth hormone secretagogue receptor 1a (GHS-R1a or GHSR), a G protein-coupled receptor (GPCR) (41, 213, 214). In addition to stimulating the release of growth hormone from the pituitary (41, 213, 214), it has been implicated in appetite stimulation (215), insulin and glucagon secretion levels (216), decreased blood pressure (217), inhibition of apoptosis in cardiomyocytes and endothelial cells, and cell proliferation and differentiation (218). Further, circulating ghrelin levels have been found to change in patients with diseases involving perturbed energy balance, such as obesity (54, 219-222) and diabetes (223). See reference (224) for thorough review. Given the current prevalence and rapidly increasing rates of obesity and related conditions, it is of importance to understand the mechanism of action of ghrelin in order to eventually contribute to the understanding of the molecular basis of these diseases.

Ghrelin carries a fatty acid (FA) modification at position Ser^3 and represents the only hormone in the human body that is lipid modified. Although the desacylated form of ghrelin is the most abundant in the bloodstream, the FA modification proves necessary for receptor binding and activation. The initial identification of acylated ghrelin revealed an octanoyl group at Ser^3 (41), but ghrelin O-acyltransferase can add FA groups of varying lengths to the peptide (58-60). Remarkably, the length of the lipid side-chain has

a demonstrated effect on the ability of ghrelin to activate GHSR and on levels of adiposity in mice (225).

Bednarek, *et al*. identified a short N-terminal segment, spanning from Gly^1 to Phe^4, including the octanoylated Ser^3, that is able to activate the GHSR *in vitro* (226), but this active core neither displaces ghrelin from its receptor nor stimulates growth hormone release *in vivo* (227). This may be due to the influence of the membrane surface on transport and receptor binding. Membrane binding of a ligand is a crucial step for membrane-receptor activation. The limitation of ligand diffusion to two dimensions, as well as structural pre-orientation and pre-organization of the ligand, may lead to enhanced peptide-receptor interaction probability (228). However, more structural and dynamic information of the peptide in solution, membrane-bound, and receptor-bound states is needed to further examine this so-called "membrane catalysis".

Our current understanding of the structure of membrane-bound ghrelin is fragmentary at best. Spectroscopic studies from proton nuclear magnetic resonance (^{1}H NMR) and circular dichrosim (CD) of ghrelin in solution revealed a highly flexible peptide without a distinct structure, regardless of whether or not Ser^3 was acylated (61). CD experiments conducted in the membrane mimics, sodium dodecyl sulfate (SDS) and trifluoroethanol (TFE), showed formation of an α-helix with increasing TFE content (62), and molecular dynamics (MD) simulations in water and in 1,2-dihexanoyl-*sn*-glycero-3-phosphocholine (DMPC)-lipid bilayer/water systems suggest that this helix extends from Pro^7 to Gln^{13} (63). Chemical shift (CS) data from ^{1}H NMR experiments performed in phosphate buffered saline and in live cells also indicated a putative α-helix between residues Glu^8 and Lys^{20}, while the peptide remained seemingly unstructured in water

(64). A structure of desacyl-ghrelin solved with CS data from ^{1}H NMR experiments performed in a water/hexafluoroacetone (HFA) mixture supports the presence of a stable α-helix spanning from Pro^{7} to Gln^{14} (229). Furthermore, controversial results were published about the membrane binding segment. While simulations propose a C-terminal loop that mediates binding, with the octanoyl moiety pointing towards the aqueous phase (63), solution NMR experiments suggested that the peptide binds to detergent micelles via Phe^{4} and the lipid-modified Ser^{3} (230).

Lipid modifications typically serve as membrane anchors (231, 232). However, a short octanoyl chain is only weakly hydrophobic, and the strength of its interaction with the membrane has yet to be determined. In order to characterize the structure and dynamics of octanoylated ghrelin and how it interacts with the membrane, we employed solid-state NMR spectroscopy (ssNMR), which has been demonstrated to be a useful and versatile tool for studying membrane-associated proteins and peptides (233-235). We show that ghrelin binds to large unilamellar vesicles (LUVs) via its octanoyl chain and assumes a highly mobile structure at the membrane surface.

Previous research indicates that ghrelin is highly flexible, even in the presence of membranes, and its secondary structure propensities in LUVs remain unknown. Therefore, the CSs obtained from ssNMR were used in combination with the Rosetta molecular modeling software (100, 101, 103), which has been widely used for protein structure prediction. NMR CSs can be used to enhance Rosetta's ability to sample native-like structures (106, 107, 236, 237), with modeling of membrane and membrane-associated proteins becoming more feasible. Recently, the structure of hepatitis C virus protein p7, a small, helical membrane protein, was determined using the

RosettaMembrane environment (105, 156) with NMR CS, residual dipolar coupling (RDC), and paramagnetic relaxation enhancement (PRE) structural data (238). In addition to the extensive ssNMR studies mentioned above, we present a new, detailed protocol for elucidating the structural ensemble of membrane-associated ghrelin that is consistent with sparse CS data.

Materials and methods

Materials

1,2-Dimyristoyl-*sn*-glycero–3-phosphocholine (DMPC), 1,2-dimyristoyl(d_{54})-*sn*-glycero–3-phosphocholine (DMPC-d_{54}), 1,2-dimyristoyl(d_{54})-*sn*-glycero–3-phosphocholine–1,1,2,2-d_4-N,N,N-trimethyl-d_9 (DMPC-d_{67}), 1,2-dimyristoyl-*sn*-glycero–3-phosphatidylserine (DMPS), 1,2-dimyristoyl(d_{54})-*sn*-glycero–3-phosphatidylserine (DMPS-d_{54}), 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC) and 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphoglycerol (POPG) were purchased from Avanti Polar Lipids, Inc. (Alabaster, AL) and used without further purification. $^{13}C/^{15}N$ Fmoc-protected amino acids and deuterated octanoic acid were obtained from Euriso-Top GmbH, Saarbrücken, Germany. All other materials were purchased from Sigma, Deisenhofen, Germany.

Peptide synthesis

Ghrelin analogs were synthesized automatically on a Wang resin by solid-phase peptide synthesis (SPPS) using Fmoc/tBu protection group strategy on a robot system (SyroI, MultiSynTech, Bochum, Germany) as described previously.(239) $^{13}C/^{15}N$-labeled

amino acids were introduced via manual peptide coupling using 5 equiv Fmoc-amino acid, 5 equiv DIC and 5 equiv HOBt in DMF. To enable the incorporation of octanoic acid or perdeuterated octanoic acid, Ser^3 was introduced with the labile Trt side-chain protecting group. The ester bond was formed by incubation of 5 equiv octanoic acid, 5 equiv DMAP, and 5 equiv DCC in NMP with the resin. The final peptides were cleaved from the resin in one step, and purification was achieved by preparative HPLC on a reversed-phase C18 column (Phenomenex Jupiter 10u Proteo 90 Å: 250 × 21.2 mm; 7.8 µm; 90 Å). Peptides were analyzed by MALDI-TOF MS (UltraflexII, Bruker, Bremen, Germany) and by analytical reversed-phase HPLC on columns VariTide RPC (Varian: 250 × 4.6 mm; 6 µm; 200 Å) and Phenomenex Jupiter 4u Proteo 300 Å (Phenomenex: 250 × 4.6 mm; 4 µm; 300 Å)_ENREF_20. The observed masses were in full agreement with the calculated masses, and peptide purity ≥ 95% could be obtained, according to the analytical RP-HPLC.

Sample preparation

Aliquots of lipids were co-dissolved in chloroform; the solvent was evaporated, and the lipid film was suspended in 10 mM MES buffer (100 mM NaCl, pH 6) to reach a final concentration of 20 mM. After freeze-thaw cycles, the suspension was extruded across 100 nM polycarbonate membranes to produce LUVs (240). Aliquots of ghrelin were added to the LUVs to reach the desired peptide/lipid ratio. Samples were incubated for 2 h while shaking it at 190 rpm at 37 °C. Binding to the inner membrane leaflet was achieved after performing another five freeze-thaw cycles. The sample was ultracentrifuged at ~90.000 g for 8 h. After lyophilization, the precipitate was hydrated to

35 WT% water content, mixed with 5 freeze-thaw cycles, and transferred into 4 mm MAS rotors with Teflon inserts.

Membrane binding assay

For membrane binding analysis of ghrelin, 5 µM peptide solutions were ultracentrifuged with various amounts of 176 mM sucrose-loaded POPC/POPG vesicles (5:1, mol/mol). For each vesicle concentration, 10 µL of a 50 µM peptide solution in H_20 were added to 740 µL of iso-osmolar 1 mM MOPS buffer at pH 7, containing 100 mM KCl. Vesicle solutions of various lipid concentrations ranging from 0 mM to 10 mM were added to reach a final volume of 1 mL. Each concentration was prepared in duplicate, and lipid-only samples were taken to determine background signals. After vortexing and 30 min incubation at room temperature, samples were ultracentrifuged overnight at ~90.000 g and 4 °C. Immediately after centrifugation, ~900 µL supernatant were transferred into Eppendorf tubes. Pellets were resuspended in the remaining solution (~100 µL) and diluted by adding 900 µL buffer. 600 µl of both the supernatant and the pellet solutions were used to determine peptide concentration using a fluorescamine assay (241). The remaining volumes were used to measure the lipid concentration.

The pH of the samples was elevated to 10 using 5 µL 0.1 M KOH. 250 µL of a fluorescamine stock solution in acetone (1 mg/mL) was added to the samples, and the fluorescence was measured after ~5 min with excitation at 390 nm and emission at 475 nm. Background fluorescence determined from the lipid-only samples was subtracted, and the percentage of bound peptide was calculated according to the equation:

$$\%\ peptide\ bound = \frac{1 - I_{supernatant}}{I_{supernatant} + I_{pellet}} \cdot 100. \qquad (1)$$

The final lipid concentration was determined by phosphate analysis. Approximately 90% of the lipids were in the pellet fraction. Further, half of the lipids are not accessible for the peptide because the molecules are on the inside of the vesicles. Accordingly, the lipid concentrations were corrected with the factor 0.45 to deliver the effective lipid concentration $[L]_{eff}$ (242).

Solid-state NMR Spectroscopy

^{2}H NMR spectra were acquired using an Avance 750 MHz NMR spectrometer (Bruker Biospin, Rheinstetten, Germany) operating at a resonance frequency of 115.0 MHz for ^{2}H using a quadrupolar-echo pulse sequence, a 90°-pulse length of 2.8 μs, an echo time of 60 μs, and a relaxation delay of 0.75 s. Smoothed chain order parameter profiles were calculated from the quadrupolar splittings after dePaking, as described in reference (243). Standard ^{31}P NMR spectra were acquired on a Bruker DRX300 NMR spectrometer operating at a resonance frequency of 121.4 MHz using a standard Hahn echo pulse sequence with a 90° pulse length of 10.75 μs, a delay between pulses of 50 μs, and a relaxation delay of 2.5 s. The ^{13}C magic angle spinning (MAS) NMR spectra were acquired using a Bruker Avance III 600 NMR spectrometer at resonance frequencies of 600.1 MHz and 150.9 MHz for ^{1}H and ^{13}C, respectively. Typical ^{1}H and ^{13}C 90° pulse lengths were 4 and 5 μs, respectively, while the decoupling filed during acquisition was ~65 kHz using Spinal64. Standard CP (contact time 700 μs), directly excited, and INEPT excitation schemes were used. All CSs were referenced to external crystalline glycine at

176.46 ppm (equivalent to TMS). Standard 2D HetCor (244) and PDSP (245) spectra were acquired, with a total evolution time of 7.1 ms and 1.7 ms in the ^{1}H and ^{13}C indirect dimensions, respectively. Constant time DIPSHIFT experiments (246) were carried out at a MAS frequency of 4 kHz with FSLG homonuclear decoupling. Dipolar dephasing curves were simulated as described in the literature (247). The ratio of the motional averaged and full dipolar coupling (248) defined the molecular order parameter, *S*.

Spin diffusion experiments from the lipid into ghrelin were carried out using the pulse sequence from the literature (249). A T_2 filter of 6 ms and spin diffusion times from 0.01 to 900 ms were used. Peak intensities were corrected for relaxation using measured T_1 relaxation times. Intensities were normalized to 1 for the longest spin diffusion time of 900 ms. Spin diffusion build-up curves were simulated as a function of mixing time using a one-dimensional lattice model (250). In this model, the magnetization of a given spin (M_i) is transferred to the neighboring spins (M_{i-1} and M_{i+1}) according to:

$$\Delta M_i / \Delta t_m = -2\Omega M_i + \Omega M_{i+1} + \Omega M_{i-1} \qquad (2)$$

The rate of magnetization transfer, $\Omega = D / a^2$ depends on the spin diffusion coefficient, *D*, and the distance between spins, *a*. Simulations were carried out using D = 0.001 nm^2/s and a = 2 Å.

Overview of structure determination using Rosetta

The Rosetta Topology Broker framework (107, 108, 251) was employed to fold ghrelin *de novo*, or from the sequence, in the presence of the implicit RosettaMembrane environment (105, 156). The traditional Rosetta fragment-based assembly algorithm for

soluble proteins was employed (100, 102). The modeling and analysis protocol is summarized in Figure 10, and full details are available in Appendix C.

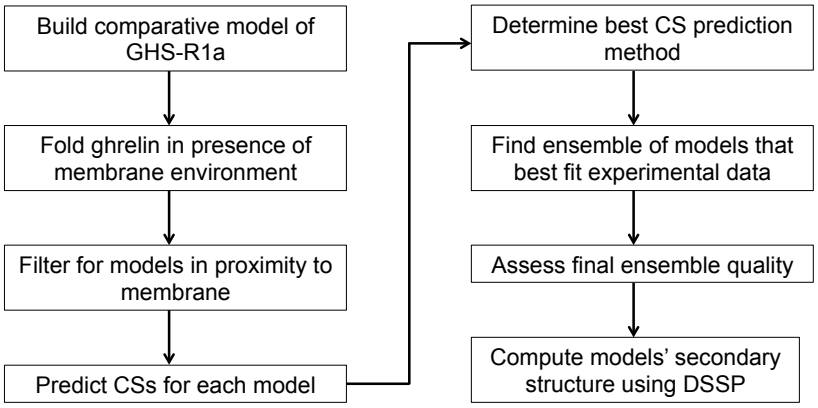

Figure 10: Flowchart of computational modeling and analysis protocol
The above flowchart outlines the protocol used to elucidate the structure of ghrelin based on ssNMR CS data.

Definition of membrane location in Rosetta

In order to fold membrane-associated proteins using Rosetta, transmembrane helical (TMH) regions must be specified. Therefore, because the modeling objective was to fold ghrelin at the membrane interface, a comparative model of GHS-R1a was created based on an alignment of nineteen different GHSR sequences and the sequences of twenty GPCRs of known structure (Figure 11). This receptor model was only used as a proxy to define the membrane location; that is, no interaction between receptor and peptide occurs. In the starting conformation for peptide folding, the receptor was placed more than 50 Å away from the peptide. During selection of the final ensemble, only models having a minimum interatomic receptor-to-peptide distance of 5 Å were analyzed and compared to experimental CSs.

Generation of GHSR comparative model to define membrane location in Rosetta

In order to fold a peptide at the membrane surface, for technical reasons, Rosetta requires at least one transmembrane span to define the location of the membrane. We decided to construct a comparative model of growth hormone secretagogue receptor 1a (GHSR) as we expect to leverage it in future studies. We then used this model to define the membrane location but ensured that no interaction between receptor and peptide occurs for the present study.

The comparative model was based on the sequence alignment in Figure 11 and generated according to the protocol described previously (19, 252-254). Briefly, GHSR amino acid sequences from nineteen species were aligned using ClustalW (19, 252-257), resulting in a sequence alignment profile. Next, twenty GPCRs of known structure, henceforth referred to as templates, were structurally aligned in Mustang (19, 254, 257-259), which resulted in a structural alignment profile. Then, a profile-profile alignment was performed in ClustalW, and the resulting alignment was manually adjusted to minimize gaps in TMH regions and maximize alignment of regions conserved across GPCRs (Figure 11).

The sequence of human GHSR was isolated from the final profile-profile alignment and threaded onto the backbone of the bovine rhodopsin structure (PDB: 1U19 (259-261)). Next, all loops and areas of missing electron density (from alignment gaps) were built in for one hundred models using the Rosetta cyclic coordinate descent (CCD) loop modeling algorithm (255, 257). The five lowest-energy models that did not contain chainbreaks were used as starting models for constructing extracellular loops (ECLs). For each starting structure, ECLs 1–3 were constructed for 185–200 models, resulting in a

total of approximately 985 complete comparative models. Finally, the lowest energy model after building the ECLs was selected to define the membrane in the ghrelin folding protocol.

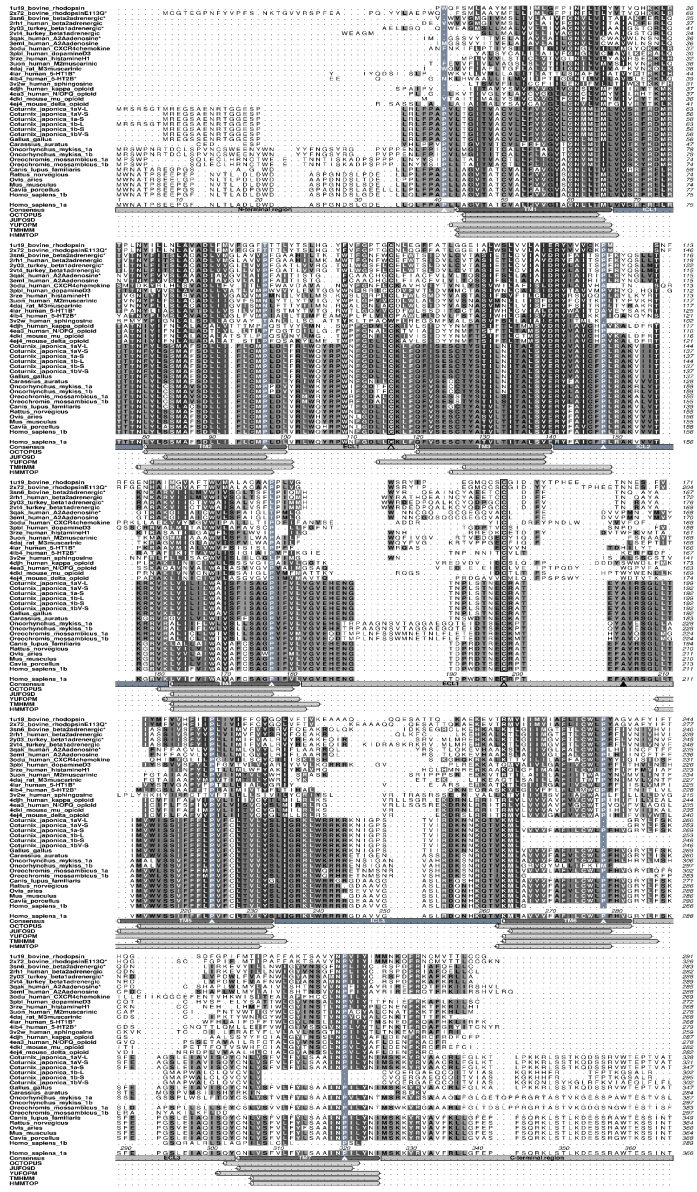

Figure 11: Sequence alignment of GHSR and GPCRs of known structure
The sequences of twenty GPCR templates of known structure and nineteen GHSR sequences were used manually aligned in Aline (260, 262) (http://crystal.scb.uwa.edu.au/charlie/software/aline/) such that gaps in the predicted TMH consensus ranges (dark gray helices) were minimized and conserved prolines (white triangles) and cysteines (open gray triangles) remained in alignment.

Fragment selection of ghrelin in Rosetta

A complete set of CSs can greatly increase the quality of fragments selected for Rosetta *de novo* structure prediction (106, 263). Fragment selection for *de novo folding* in Rosetta heavily prioritizes peptide fragment conformations that have the same secondary structure as that indicated by CS analysis (106, 264, 265). In the case of ghrelin, however, the CS dataset is incomplete, i.e. CS assignments are not available for every residue. This leads to inconsistencies in fragment selection, where CSs of a few residues can determine the secondary structure of the entire fragment. In the present case, the CS data suggest that residues 2-5 have β-strand torsion angles. Accordingly, these residues are often constructed from fragments that stem from β-hairpins (Figure 12). In result, even though the fewer CSs obtained for residues 8-28 are indicative of a random coil region with a slight helical tendency, the vast majority of Rosetta models generated from fragments selected based on the sparse CS dataset exhibited a β-hairpin fold (data not shown). Therefore, we elected to fold with fragments not generated using CS data, thereby sampling the complete conformational space reasonable for a peptide of this sequence. We then employed CS data to filter, from a large pool of models, an ensemble that agreed best with the CS data. This approach has another advantage in the case of highly flexible peptides in that the ensemble average CSs, not the CS of a single model, must conform to the experimental data.

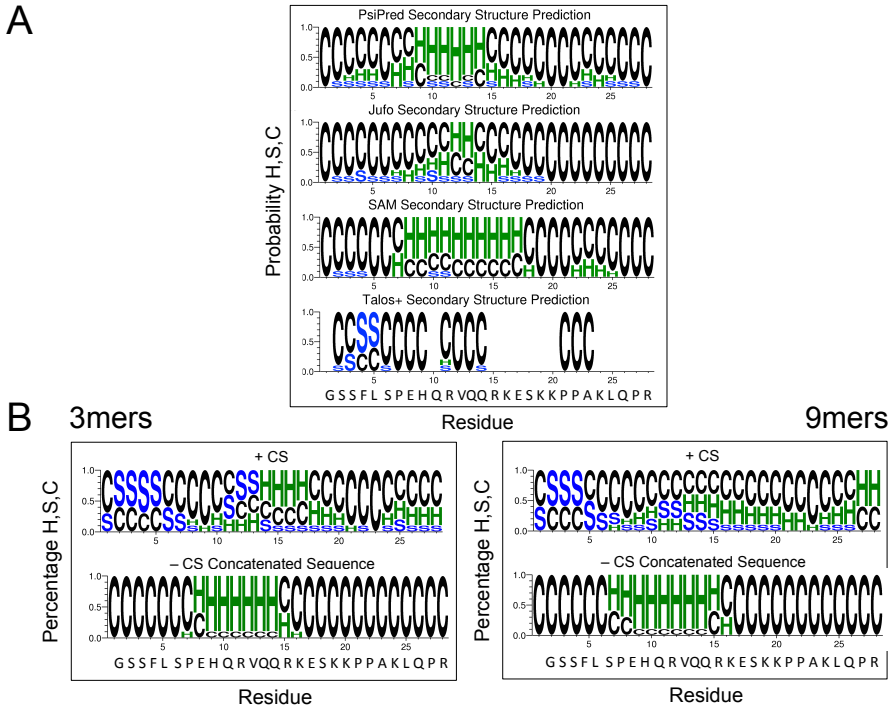

Figure 12: Secondary structure prediction of ghrelin
A) Secondary structure prediction for the primary sequence of ghrelin. B) Secondary structure composition of 3mer and 9mer amino acid fragments used in *de novo* folding. These fragments were generated based either on the primary sequence of the peptide alone (+CS and –CS). For all predictions, α-helices (H) are in green, β-strands (S) are in blue, and random coil (C) are in black.

De novo *folding of ghrelin in Rosetta*

During folding in the Topology Broker framework, 3- and 9-amino acid peptide fragments were inserted into an extended backbone of the peptide in a Monte Carlo fashion. The resulting conformations were scored with the RosettaMembrane (105, 130) potentials according to the Metropolis criterion (266). Ten thousand models were generated in the presence of the membrane and relaxed within the all-atom membrane potential.

Prediction of chemical shifts of de novo-*folded models*

Predicted CSs for the models generated from *de novo* folding were obtained by running PROSHIFT (253), SPARTA+ (261), SHIFTX (256), and SHIFTX2 (258). When running PROSHIFT, the temperature and pH were set to 303 K and 6.0, respectively. SPARTA+, SHIFTX, and SHIFTX2 were run using default settings. During CS analysis, CSs obtained for Gly^{1} were disregarded. (See *Protocol Capture* in Appendix C).

Selection of models that are of low energy and in contact with membrane

To maintain close contact between Ser^{3} and the membrane, all 10,000 models were filtered so that the Ser^{3} C_{α} of the filtered models were within the polar region of the RosettaMembrane implicit membrane environment. The remaining pool of 3,692 models was screened to filter out those models in which any peptide atoms found within 5 Å of any receptor atoms. All 3,692 of these models passed the filter and were further culled by keeping only those models whose Rosetta energies were within the top 10% of all 10,000 model energies, leaving a fully filtered pool of 355 models. This percentage was chosen after testing various ensemble sizes (Table 10).

Table 10: Ensemble average RMSDs (in ppm) resulting from filtering strategies

	Top 10% by Rosetta Energy	**Top 25% by Rosetta Energy**	**Top 50% by Rosetta Energy**	**Top 75% by Rosetta Energy**	**All**
PROSHIFT[b]	0.385 (22)	0.381 (18)	0.379 (26)	0.378 (26)	0.377 (26)
SHIFTX[c]	0.716 (18)	0.713 (28)	0.707 (23)	0.707 (17)	0.709 (13)
SHIFTX2[d]	0.722 (29)	0.719 (26)	0.718 (26)	0.702 (26)	0.717 (11)
SPARTA+[e]	0.718 (29)	0.717 (20)	0.711 (30)	0.711 (24)	0.718 (12)
# models in pool	355	856	1,790	2,683	3,692

[a] Ensemble size in parentheses
[b] References (19, 252-254)
[c] References (19, 254-257)
[d] References (19, 254, 257-259)
[e] References (259-261)

Generation of model ensembles in agreement with experimental chemical shifts

Ensembles of 10-30 models consistent with the experimental CSs were constructed from the resulting low-energy pool according to the algorithm summarized in Figure 13. PROSHIFT (253) was used to predict CSs for all models. The selection algorithm generates a random ensemble of 10 models. It then computes the average CS of each C_α, C_β, C_O, and H_α atom for which an experimental CS was determined (excluding those for Gly[1]). After all average CS values are determined, the root mean square deviation (RMSD) of the ensemble average-predicted CSs relative to the experimental CSs is calculated and reported. In order to avoid the average and RMSD being dominated by the larger magnitude of carbon CS values, carbon CSs were scaled down by a factor of 4. Next, the algorithm randomly chooses to add another model from the bigger pool to the ensemble (if not at the specified maximum ensemble size of 30), swap models between the ensemble and the pool, or remove a model from the ensemble (if not at the minimum ensemble size of 10). The process is repeated for 5,000,000 cycles.

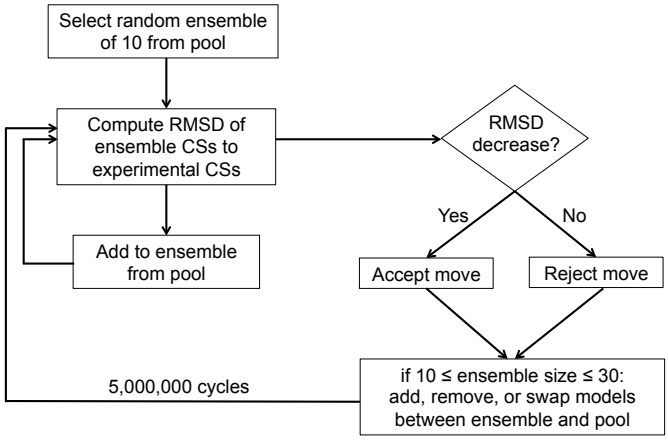

Figure 13: Outline of model ensemble selection algorithm
The above flowchart outlines the process by which the agreement with experimental data is determined for an ensemble of models selected from a large pool.

Structural analysis of final ensemble

The final ensemble of models was initially evaluated by the Protein Structure Validation Software suite (PSVS, http://psvs–1_5-dev.nesg.org/). The secondary structure information, including ϕ/ψ torsion angles, was obtained by running Define Secondary Structure of Proteins (DSSP (182), http://swift.cmbi.ru.nl/gv/dssp/). In addition, the DSSP analysis was modified to take into account polyproline II (PPII) helical structure using the same parameters presented by Adzhubei, Sternberg, and Makarav (267). Briefly, residues were only assigned PPII structure if they met the following conditions: 1) formerly assigned random coil (-) by DSSP, 2) ϕ = –75 ± 29 degrees, 3) ψ = 145 ± 29 degrees, 3) conditions 1) and 2) were met for two sequential residues.

Results

Ghrelin binds to negatively charged membranes

First, binding of ghrelin and desacyl-ghrelin to POPC/POPG (5/1, mol/mol) membranes was measured using an ultracentrifugation assay. Upon addition of sucrose loaded vesicles and ultracentrifugation, bound ghrelin co-precipitates with the liposomes and the percentage of bound peptide is measured with a fluorescamine assay, as shown in Figure 14. While about ~65% of the octanoylated ghrelin binds to the acidic liposomes with a K_D value of 100 ± 24 μM, only ~10% desacyl ghrelin is associated to the membranes at a lipid concentration of 5 mM without reaching saturation, indicating the importance of the octanoyl modification. The K_D-derived ΔG value for the binding of ghrelin to membrane surfaces is –28.6 kJ/mol. To confirm that the lipid membranes used

in this study were in a lamellar liquid crystalline phase state, static ^{31}P NMR spectra were recorded. All preparations showed the typical axially symmetric powder pattern, with a $\Delta\sigma$ = 45 ppm (data not shown).

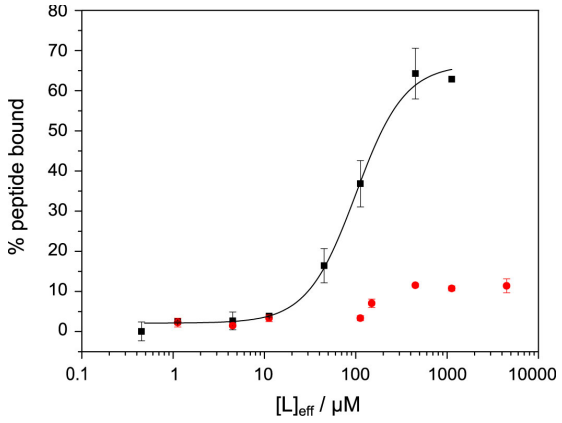

Figure 14: Binding isotherm of ghrelin and desacyl ghrelin to POPC/POPG membranes
The amount of bound ghrelin (black squares) and desacyl ghrelin (red circles) as a function of lipid concentration is given. The ghrelin binding curve was fitted according to Equation (1). No significant membrane binding is observed for desacyl ghrelin.

To understand the dynamics of the membrane lipids and the lipid modification of membrane-associated ghrelin, the properties of the lipid chains in four different samples were compared: 1) pure DMPC-d_{54}/DMPS, 2) DMPC-d_{54}/DMPS/ghrelin at a 30:1 molar lipid-to-peptide molar ratio, 3) DMPC-d_{54}/DMPS/desacyl-ghrelin, and 4) DMPC/DMPS/ghrelin-d_{15}, where ghrelin featured a perdeuterated octanoyl-d_{15} chain at Ser^3. This combination of samples allowed us to determine the effect of ghrelin on the bilayer properties of the host membrane. Typical ^{2}H NMR spectra of the DMPC-d_{54} and the ghrelin-d_{15} component of the mixtures are shown in Figure 15, panels A and B. The NMR spectrum of DMPC shows the typical superposition of Pake dubletts, which is typical for the lamellar liquid crystalline phase state of the membrane. A small isotropic

peak, as well as the bigger line width, indicate the presence of ghrelin. The ^{2}H NMR spectrum of ghrelin with a perdeuterated octanoyl chain also shows the features of a well-inserted peptide lipid chain, i.e., well dissolved Pake dublets. In addition, an isotropic peak that accounts for ~10% of the intensity is shown, indicating that about 10% of the octanoyl chain of ghrelin is isotropically mobile, or not inserted into the membrane.

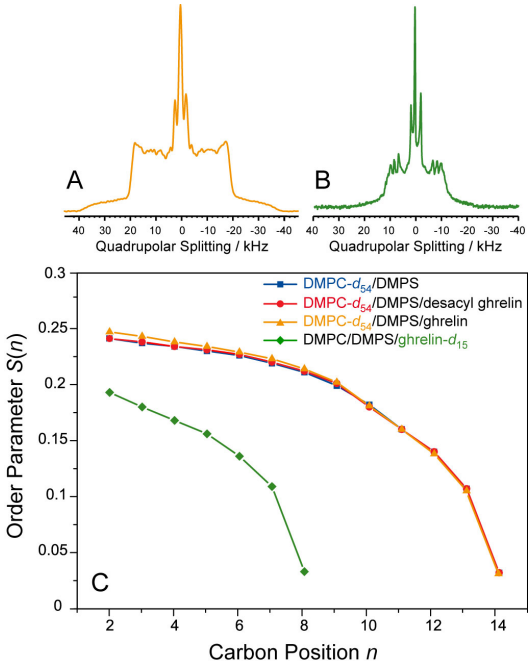

Figure 15: ^{2}H NMR spectra and order parameters of DMPC-d_{54}/DMPS membranes
^{2}H NMR spectra in DMPC-d_{54}/DMPS membranes (5/1, mol/mol) in the presence of ghrelin (A) and ghrelin-d_{15} in DMPC/DMPS membranes (B). C) ^{2}H NMR order parameters of DMPC-d_{54}/DMPS (5:1, mol/mol) membranes in the presence or absence of ghrelin (1:30 protein to lipid molar ratio) at a temperature of 30°C and a buffer content of 35 wt%.

From the ^{2}H NMR powder spectra of the four samples mentioned above, the segmental chain order parameters were determined. Smoothed chain order parameter profiles showing the dependence of the order parameter on the position of the carbon

segment in the acyl chain are presented in Figure 15C. The segments are numbered consecutively starting at the carbonyl group of the lipid or the C_β of ghrelin's Ser^3. Striking differences between the chain order parameters of DMPC-d_{54} and ghrelin-d_{15} are observed. The ghrelin octanoyl chain shows significantly lower order parameters than the host membrane for all carbon positions. In contrast, the order parameters of the host membrane are very similar in the absence and presence of both ghrelin and desacyl ghrelin. Virtually no differences are observed for DMPC-d_{54}/DMPS in the absence or presence of desacyl-ghrelin, confirming that there was no binding of the desacylated peptide to the membrane. Slightly higher order parameters are observed for the upper eight chain methylenes of the membrane in the presence of ghrelin. Using the mean torque model (268), the structural parameters of these lipid chains were calculated. The length of the DMPC chains in the mixture in the absence and presence of ghrelin was 11.1 Å and 11.3 Å, respectively. The length of the octanoyl chain of ghrelin was 4.8 Å.

^{13}C Chemical shifts were collected to study the structure of membrane-bound ghrelin

Next, the secondary structure of membrane-bound ghrelin was investigated. To this end, six peptides with varying labeling scheme were synthesized (Table 11). ^{13}C MAS NMR measurements were carried out in DMPC/DMPS (5:1, mol/mol) membranes. A comprehensive set of directly excited ^{13}C MAS NMR spectra, CP MAS spectra, and INEPT-based techniques were employed to find the most sensitive excitation scheme for membrane-bound ghrelin (269); the CP MAS technique with a contact time of 700 μs provided the most sensitivity. A typical ^{13}C CP MAS NMR spectrum of a ghrelin peptide in membranes is shown in Figure 16A. As membrane-bound peptides often aggregate at

high concentrations (248), the dependence of ghrelin CSs on peptide concentration was determined; ghrelin/lipid preparations of 1:30, 1:50, and 1:100 molar ratios were used. In all cases, there were no observable altered CSs, so a 1:30 ghrelin/lipid preparation was used for the remainder of this study.

Table 11: Overview of ghrelin peptide constructs and labeling schemes*

GHR1: H2N- **G**SS(**n-octanoyl**)F**L** **S**PEHQ RVQQR KESKK PPAKL QPR -OH	M = 3398,96 Da
GHR2: H2N- G**S**S(n-octanoyl)FL S**PE**HQ RVQQR KESKK PPAKL QPR -OH	M = 3384,96 Da
GHR3: H2N- GS**S**(n-octanoyl)FL SPEH**Q** RVQQR KESKK P**P**AKL QPR -OH	M = 3385,96 Da
GHR4: H2N- GSS(n-octanoyl)**F**L SPEHQ R**V**QQR KESKK PP**A**KL QPR -OH	M = 3388,96 Da
GHR5: H2N- GSS(n-octanoyl)FL SPEHQ RV**Q**QR KE**S**KK **P**PAKL QPR -OH	M = 3385,96 Da
GHR6: H2N- GSS(n-octanoyl)FL SPEHQ RVQ**Q**R KESKK PPAKL Q**P**R -OH	M = 3379,96 Da
∑: H2N- **GSS**(**n-octanoyl**)**FL SPE**H**Q** R**VQQ**R KE**S**KK **PPA**KL Q**P**R -OH	

* Several ghrelin peptides were synthesized having 17 of the 28 amino acids that were $^{13}C/^{15}N$ labeled. The peptides were allowed to bind to LUVs having a diameter of 100 nm and a composition of 80% DMPC-d_{67} and 20% DMPS-d_{54}. Experiments were performed with 35 wt% of 10 mM MES buffer containing 10 mM NaCl at pH 6.

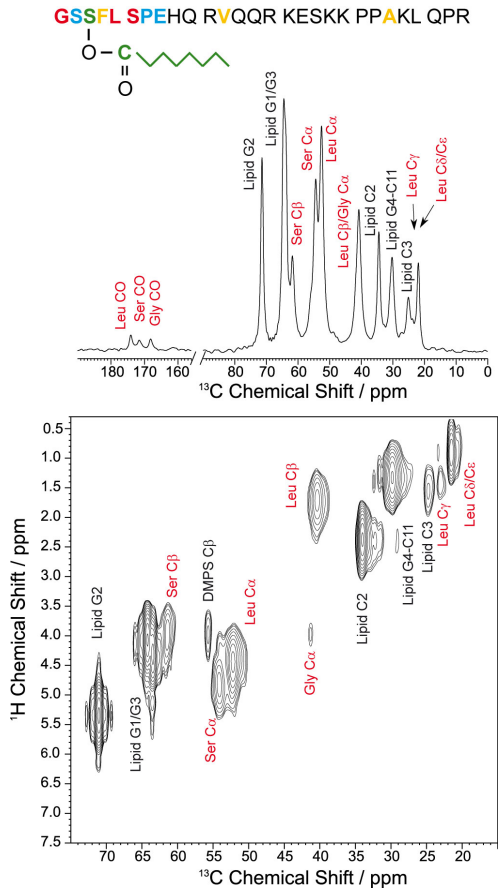

Figure 16: Ghrelin sequence showing the isotopic labeling scheme of the different molecules and ssNMR spectra of membrane-embedded ghrelin
Labeling scheme is shown in color (see Table 11). A) ^{13}C CP MAS NMR spectrum of ghrelin (with Gly[1], Leu[5], and Ser[6] labeled) in DMPC-d_{67}/DMPS-d_{54} (5:1, mol/mol) membranes at a ghrelin concentration of 3.3 mol%. B) ^{1}H-^{13}C MAS HetCor spectrum of the same preparation, all at 30°C and a MAS frequency of 7 kHz.

To achieve the full assignments of the ghrelin signals, ^{1}H–^{13}C HetCor and ^{13}C–^{13}C PDSP experiments were conducted. The basic connectivities within the labeled amino acid were determined in PDSP experiments using a mixing time of 50 ms. As membrane-associated ghrelin is relatively mobile (see below), the PDSP experiments were performed at –30°C. The high mobility of ghrelin helped in detecting ^{1}H CSs in ^{1}H–^{13}C HetCor experiments, which were well-resolved, even without application of homo-

nuclear decoupling. Typical peptide signals had a ^{1}H line width of 0.3–0.4 ppm. A characteristic ^{1}H-^{13}C HetCor NMR spectrum of membrane-associated ghrelin is shown in Figure 16B. A summary of the CS values determined for membrane-bound ghrelin is given in Table 12. The difference between $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ values for determination of secondary structure are reported in Figure 17.

Table 12: Chemical shifts measured for acylated ghrelin bound to DMPC/DMPS membranes (5/1,mol/mol) using MAS ssNMR

Residue	C_O	C_α	C_β	C_γ	C_δ	H_α	H_β	H_γ	H_δ
Gly[1]	167.0 ± 0.4	40.9 ± 0.2							
Ser[2]	172.1 ± 0.2	55.6 ± 0.5	62.5 ± 0.6						
Ser[3]		53.6 ± 0.1	63.3 ± 0.2			4.5± 0.3			
Phe[4]	172.1 ± 0.2	55.8 ± 1.2	37.0 ± 0.8						
Leu[5]	174.8 ± 0.4	51.9 ± 0.2	40.7 ± 0.5						
Ser[6]	169.3 ± 0.5	54.2 ± 0.5	61.2 ± 0.6						
Pro[7]	174.9 ± 1.2	61.2 ± 0.5	30.8 ± 1.7						
Glu[8]	174.1 ± 0.2	54.3 ± 0.9	25.8 ± 0.9						
Gln[10]	177.3± 0.3	55.4 ± 0.4	27.0 ± 0.0	34.4 ± 0.9		4.1± 0.3			
Val[12]	174.1 ± 0.2	60.3 ± 0.9	30.0± 0.3						
Gln[13]	173.4 ± 0.3	53.5 ± 0.2	27.0 ± 0.1			4.3± 0.3	2.0± 0.3		
Gln[14]	173.5± 0.3	53.4 ± 0.1	27.0 ± 0.1	33.7 ± 0.1		4.3± 0.3	2.1± 0.3	2.4± 0.3	
Ser[18]	171.8 ± 0.2	55.8 ± 0.1	61.3 ± 0.2			4.4± 0.3	3.9± 0.3		
Pro[21]	177.7± 0.3	59.0 ± 0.1	28.3 ± 0.0	24.8 ± 0.1	48.0± 0.3	4.7± 0.3		2.0± 0.3	3.8± 0.3
Pro[22]	173.7 ± 0.2	60.4 ± 0.2	29.4 ± 0.0	24.8± 0.3	47.9± 0.3	4.4± 0.3			
Ala[23]	175.5 ± 0.5	50.5 ± 0.6	17.0± 0.3						
Pro[27]	173.3 ± 0.0	60.8 ± 0.0	29.4 ± 0.1	24.8 ± 0.1	48.1± 0.3	4.4± 0.3	2.0± 0.3	2.0± 0.3	3.8± 0.3

* Gray cells indicate that these CSs were used in structure determination.

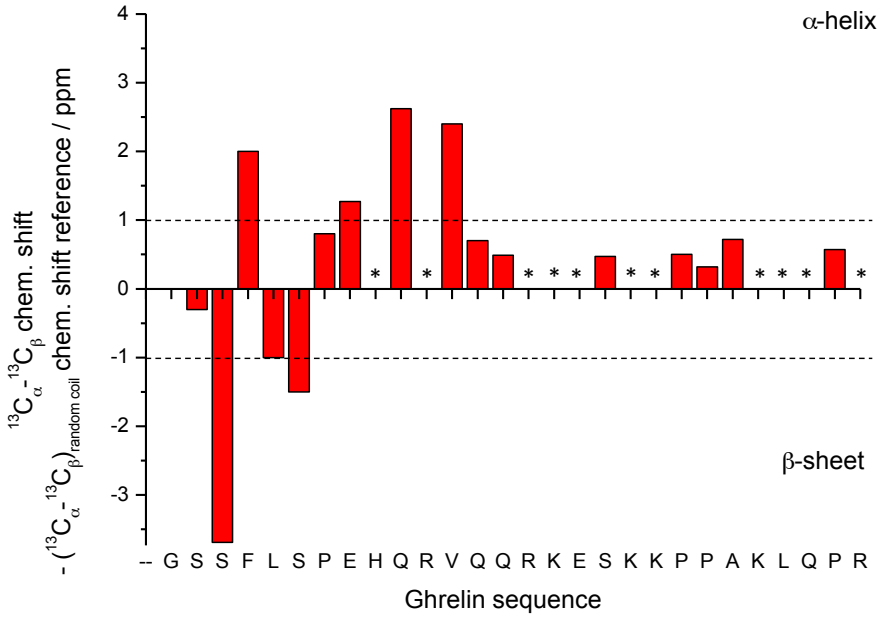

Figure 17: Chemical shift analysis of ghrelin based on MAS ssNMR data
The $^{13}C_\beta$–$^{13}C_\alpha$ values for reach residue are plotted. Positive values greater than 1 ppm indicate a tendency for α-helical structure, whereas values less than −1 ppm suggest some β-sheet character. Amino acids with a CS index close to 0 ppm are considered to have no secondary structure. Asterisks indicate that no CSs were available for that residue.

Dipolar couplings were measured to study the dynamics of membrane-bound ghrelin

Next, the dynamics of membrane-associated ghrelin were studied via dipolar coupling measurements (246). From the measurement of ^{13}C–^{1}H dipolar couplings, we determined the backbone and side-chain order parameters needed to characterize the amplitude of motion for the C-H-bond vectors. A fully rigid C-H-bond exhibits the maximal dipolar coupling strength of 22.8 kHz, corresponding to an order parameter of 1. An order parameter value of 0 corresponds to fully isotropic motion, which is expressed by a vanishing dipolar coupling. Molecular motions with a given amplitude lead to partial averaging of the dipolar coupling strength and can be characterized by a specific order parameter. The ^{1}H–^{13}C order parameters sample all motions with correlation times

shorter than ~10 μs (269). Overall, the order parameters for ghrelin in membranes are relatively low--around 0.2 for the backbone--with smaller values obtained for the side-chains. There are no significant differences in the order parameters for residues 1-12. However, the order parameter of Ala^{23} was significantly lower, indicating a large increase in the motional amplitude at the C-terminus (Figure 18).

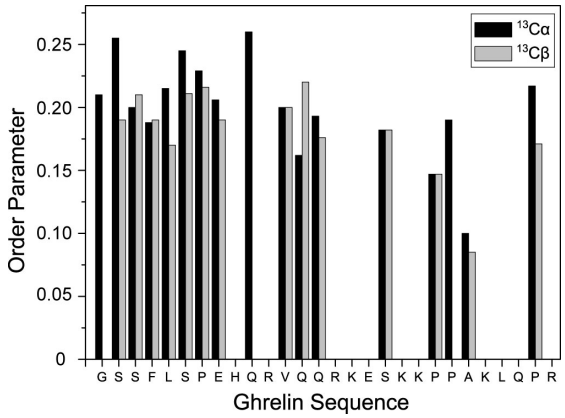

Figure 18: ^{1}H-^{13}C order parameters of ghrelin bound to DMPC/DMPS membranes
Order parameters were determined for 3.3 mol% ghrelin bound to DMPC/DMPS membranes (5:1, mol/mol) at a temperature of 30°C and a water content of 35 wt%.

Ghrelin interacts with membrane via Ser^{3} and Phe^{4}

Finally, the membrane topology of ghrelin was investigated by measuring spin diffusion from the lipid into the peptide (249). Ghrelin samples were prepared in DMPC-d_{67}/DMPS-d_{54} membranes in the presence of D_2O. Thus, spin diffusion originating from the glycerol backbone and the PS headgroup was detected in the ghrelin backbone. Typical spin diffusion curves for Ser^{3}, Phe^{4}, Val^{12}, and Ala^{23} are shown in Figure 19. At a mixing time of 0, all peptide magnetization was relaxed due to the T_2 filter of 6 ms. However, as the mixing time increases, the intensity of the ghrelin signals also increases. Qualitatively, magnetization buildup is strongest in Ser^{3} and Phe^{4}, while a significantly decreased magnetization buildup is detected for Val^{12} and Ala^{23}. This means that Ser^{3} and

Phe[4] are in close proximity to the membrane surface, while Val[12] and Ala[23] have no membrane contact because spin diffusion has to migrate longer to reach these sites.

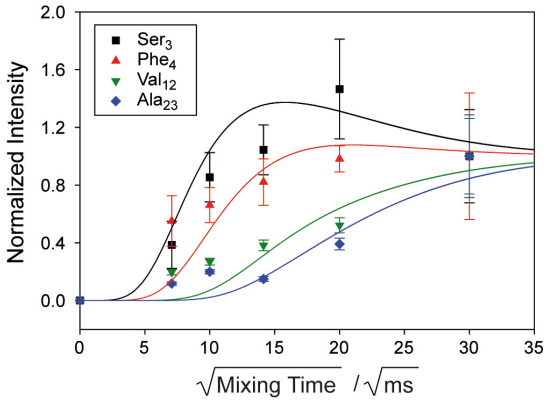

Figure 19: ^{1}H spin diffusion buildup curves of membrane-associated ghrelin
Spin diffusion spectra were determined for 3.3 mol% membrane-associated ghrelin in DMPC-d_{67}/DMPS-d_{54} (5:1, mol/mol) at a D_2O content of 35 wt%. Spin diffusion originates from the membrane's glycerol and the PS headgroups. Solid lines represent best-fit simulations using a lattice model with a spin diffusion coefficient of D = 0.001 nm^2/s and a distance between protons of 2 Å**.**

Magnetization buildup was also simulated using a simple lattice model for spin diffusion (250). As the mobilities of the lipids and ghrelin are comparable (see Figure 15 and Figure 18), a common spin diffusion coefficient of D = 0.001 nm^2/s was used for spin diffusion within the lipid, from lipid to peptide, and within ghrelin. With these simple assumptions, the magnetization buildup could be modeled relatively well using a 2-Å spacing between neighboring spins. In the lattice model, spin diffusion from the lipid reaches the peptide sites in close proximity to the membrane surface, Ser[3] and Phe[4], in 3 and 4 steps, respectively. On the other hand, 6 to 8 steps are necessary for the magnetization to diffuse to residues Val[12] and Ala[23].

PROSHIFT predicts CS of de novo *folded ghrelin with smallest deviation from experiment*

In order to construct an ensemble of ghrelin models in agreement with the experimental CS data (Table 12), an appropriate method for predicting CSs based on the *de novo* folded models was needed. We tested four CS prediction tools: PROSHIFT (253), SHIFTX (256), SHIFTX2 (258), and SPARTA+ (261). PROSHIFT employs an artificial neural network (ANN) trained on CS data from the Biological Magnetic Resonance Bank (BMRB). SHIFTX operates via a hybrid method, in which empirically derived CS hypersurfaces are combined with classical (i.e., Newtonian physics) or semi-classical equations for parameters, such as ring current, hydrogen bond, and solvent effects. SHIFTX2, like SHIFTX, employs structure-based concepts used by SHIFTX, but the algorithm also takes sequence homology information into account, as is done by SHIFTY (270). SPARTA+ uses an ANN, but, being a newer method, the ANN was trained on an approximately two-fold larger protein database than was used for training the PROSHIFT ANN. We hypothesized that the fragment-based assembly in Rosetta samples the conformational space likely occupied by the biologically active peptide and that, therefore, some models within the final ensemble represent conformations that give rise to the observed CSs. Accordingly, one can argue that the CS prediction algorithm most suitable for this particular application should give the lowest CS-RMSD between experimental and predicted CS. Because not all of the CS prediction methods predict values for side-chain atoms, including protons, only C_O, C_α, C_β, and H_α CSs were used in the determination of the CS-RMSD.

After using all four of the aforementioned methods to predict CSs for the 10,000 Rosetta-generated models, the CS-RMSD (in ppm) of each model to the experimental data was computed. The Rosetta score, or energy, was plotted against CS-RMSD, as determined by each CS prediction method (Figure 20). Surprisingly, it was found that PROSHIFT systematically created lower CS-RMSD values. Manual inspection of one selected model that agreed well with predicted CSs from all methods confirmed that more accurate CSs were predicted throughout the peptide and not located in one particular region (Table 13).

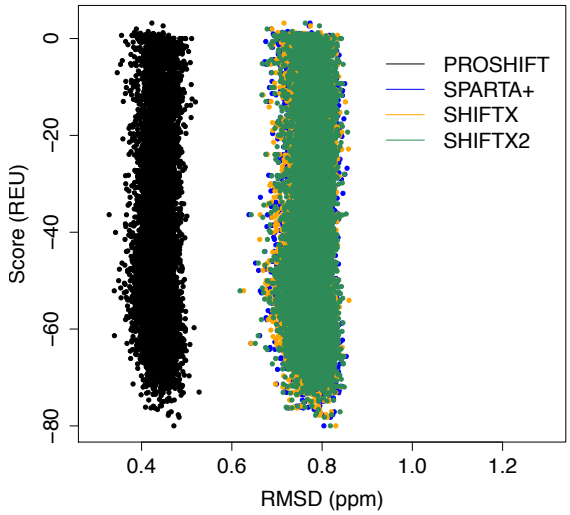

Figure 20: Assessment of four chemical shift prediction methods
Score vs. RMSD (in ppm) plot, where the RMSD of each model's predicted CSs to experimental values were computed. The RMSD was computed over each of the experimentally determined CSs, excluding the two CSs determined for Gly[1].

Table 13: Detailed analysis of low-RMSD model from set of filtered models in top 10% by score

Res	Atom	CS_{exp}[a,b]	$CS_{PROSHIFT}$	$\lvert CS_{PROSHIFT} - CS_{exp} \rvert$[c]	$CS_{SPARTA+}$	$\lvert CS_{SPARTA+} - CS_{exp} \rvert$	CS_{SHIFTX}	$\lvert CS_{SHIFTX} - CS_{exp} \rvert$	$CS_{SHIFTX2}$	$\lvert CS_{SHIFTX2} - CS_{exp} \rvert$
Ser[2]	C_O	172.1 ± 0.2	173.0	0.2	176.5	1.1	174.9	0.7	176.28	1.0
Ser[2]	C_α	55.6 ± 0.5	57.0	0.4	61.0	1.4	58.83	0.8	61.08	1.4
Ser[2]	C_β	62.5 ± 0.6	62.3	0.0	63.0	0.1	62.98	0.1	63.06	0.1
Ser[2]	C_α	53.6 ± 0.1	58.4	1.2	60.2	1.7	60.67	1.8	60.48	1.7
Ser[2]	C_β	63.3 ± 0.2	61.4	0.5	62.6	0.2	63.78	0.1	62.65	0.2
Ser[2]	H_α	4.5 ± 0.3	4.2	0.3	4.4	0.1	4.27	0.2	4.33	0.2
Phe[4]	C_O	172.1 ± 0.2	174.5	0.6	176.2	1.0	175.45	0.8	176.27	1.0
Phe[4]	C_α	55.8 ± 1.2	56.0	0.1	58.2	0.6	59.48	0.9	58.9	0.8
Phe[4]	C_β	37.0 ± 0.8	38.0	0.3	39.0	0.5	38.85	0.5	39.36	0.6
Leu[5]	C_O	174.8 ± 0.4	173.9	0.2	176.3	0.4	175.41	0.2	176.17	0.3
Leu[5]	C_α	51.9 ± 0.2	51.5	0.1	53.6	0.4	53.3	0.4	54.2	0.6
Leu[5]	C_β	40.7 ± 0.5	39.8	0.2	43.3	0.7	43.13	0.6	43.52	0.7
Ser[6]	C_O	169.3 ± 0.5	172.2	0.7	172.7	0.8	172.64	0.8	173.32	1.0
Ser[6]	C_α	54.2 ± 0.5	54.0	0.1	54.9	0.2	56.51	0.6	55.58	0.3
Ser[6]	C_β	61.2 ± 0.6	62.8	0.4	64.1	0.7	63.58	0.6	63.85	0.7
Pro[7]	C_O	174.9 ± 1.2	174.3	0.2	178.0	0.8	177.03	0.5	177.79	0.7
Pro[7]	C_α	61.2 ± 0.6	61.3	0.0	62.3	0.3	62.01	0.2	62.11	0.2
Pro[7]	C_β	30.8 ± 1.7	30.9	0.0	32.9	0.5	33.75	0.7	33.42	0.7
Glu[8]	C_O	174.1 ± 0.2	176.0	0.5	178.9	1.2	178.78	1.2	178.75	1.2
Glu[8]	C_α	54.3 ± 0.9	56.9	0.7	60.2	1.5	59.66	1.3	59.46	1.3
Glu[8]	C_β	25.8 ± 0.9	27.4	0.4	29.0	0.8	29.09	0.8	29.23	0.9
Gln[10]	C_O	177.3 ± 0.3	176.2	0.3	178.6	0.3	178.93	0.4	178.73	0.4
Gln[10]	C_α	55.4 ± 0.4	57.4	0.5	59.3	1.0	58.88	0.9	59.11	0.9
Gln[10]	C_β	27.0 ± 0.0	27.0	0.0	28.5	0.4	28.72	0.4	28.58	0.4
Gln[10]	H_α	4.1 ± 0.3	3.9	0.2	3.8	0.3	3.93	0.2	3.98	0.2
Val[12]	C_O	174.1 ± 0.2	175.5	0.3	177.0	0.7	177.75	0.9	178.15	1.0

Val[12]	C_α	60.3 ± 0.9	63.9	0.9	65.7	1.3	66.1	1.5	65.84	1.4
Val[12]	C_β	30.0 ± 0.3	28.8	0.3	31.5	0.4	31.63	0.4	31.75	0.4
Gln[13]	C_O	173.5 ± 0.3	175.2	0.4	178.7	1.3	177.83	1.1	176.87	0.8
Gln[13]	C_α	53.4 ± 0.2	56.3	0.7	57.1	0.9	58.16	1.2	57.74	1.1
Gln[13]	C_β	27.0 ± 0.1	27.0	0.0	28.7	0.4	28.64	0.4	28.66	0.4
Gln[13]	H_α	4.3 ± 0.3	4.1	0.2	4.1	0.2	4.11	0.2	4.14	0.2
Gln[14]	C_O	173.5	173.9	0.1	176.3	0.7	176.43	0.7	176.36	0.7
Gln[14]	C_α	53.4 ± 0.1	55.0	0.4	56.6	0.8	57.55	1.0	57.21	0.9
Gln[14]	C_β	27.0 ± 0.1	26.5	0.1	29.0	0.5	29.58	0.7	29.14	0.5
Gln[14]	H_α	4.3	4.3	0.0	4.1	0.2	4.07	0.2	4.19	0.1
Ser[18]	C_O	171.8 ± 0.2	173.2	0.4	174.8	0.8	174.27	0.6	173.9	0.5
Ser[18]	C_α	55.8 ± 0.1	56.4	0.2	57.9	0.5	57.89	0.5	58.23	0.6
Ser[18]	C_β	61.3 ± 0.2	60.3	0.3	64.7	0.8	64.13	0.7	64.47	0.8
Ser[18]	H_α	4.4 ± 0.3	4.4	0.0	4.5	0.0	4.42	0.0	4.41	0.0
Pro[21]	C_O	177.7 ± 0.3	173.4	1.1	175.1	0.6	174.84	0.7	175.33	0.6
Pro[21]	C_α	59.0 ± 0.1	59.1	0.0	61.8	0.7	61.79	0.7	62.32	0.8
Pro[21]	C_β	28.3 ± 0.0	29.8	0.4	31.5	0.8	31.89	0.9	30.97	0.7
Pro[21]	H_α	4.7 ± 0.3	4.6	0.1	4.4	0.3	4.42	0.3	4.6	0.1
Pro[22]	C_O	173.7 ± 0.2	174.3	0.2	176.1	0.6	176.41	0.7	177.46	0.9
Pro[22]	C_α	60.4 ± 0.2	60.7	0.1	62.6	0.5	62.68	0.6	62.72	0.6
Pro[22]	C_β	29.4 ± 0.0	30.9	0.4	32.3	0.7	31.89	0.6	32.19	0.7
Pro[22]	H_α	4.4 ± 0.3	4.5	0.1	4.1	0.4	4.21	0.2	4.42	0.0
Ala[23]	C_O	175.5 ± 0.5	175.0	0.1	177.1	0.4	177.18	0.4	177.2	0.4
Ala[23]	C_α	50.5 ± 0.6	50.2	0.1	51.4	0.2	51.51	0.3	51.62	0.3
Ala[23]	C_β	17.0 ± 0.3	16.8	0.0	20.3	0.8	20.15	0.8	19.27	0.6
Pro[27]	C_O	173.3 ± 0.0	175.0	0.4	176.0	0.7	176.67	0.8	177.01	0.9
Pro[27]	C_α	60.8 ± 0.0	61.7	0.2	63.1	0.6	62.73	0.5	63.13	0.6
Pro[27]	C_β	29.4 ± 0.1	30.2	0.2	32.4	0.7	32.32	0.7	31.98	0.6

Pro^{27}	H_α	4.4 ± 0.3	4.4	0.1	4.6	0.1	4.37	0.1	4.4	0.0
Average Deviation (± S.E.M.)				**0.3 ± 0.04**		**0.6 ± 0.05**		**0.6 ± 0.05**		**0.6 ± 0.05**
RMSD				**0.4**		**0.7**		**0.7**		**0.7**

[a] All values in ppm
[b] Experimental and predicted values not scaled. Difference values take scaling into account.
[c] All CS differences (in | |) are scaled. Scaling = CS_{carbon} * 0.25

Furthermore, the selection algorithm used to find the ensemble of models with the best overall agreement to the experimental CSs resulted in lower average RMSD values when the model CSs were predicted by PROSHIFT (Table 10 and Table 12). After running the selection algorithm over model pools of various sizes, each with PROSHIFT, SHIFTX, SHIFTX2, or SPARTA(+)-predicted CSs, it was determined that, for this system, the top 10% of models by total Rosetta score that had Ser^3 C_α atoms in proximity to the membrane plane struck the best compromise between favorable Rosetta energy and agreement with experimental CSs.

The final structural ensemble of ghrelin is highly flexible

The final ensemble of 22 ghrelin models had a CS-RMSD of 0.4 ppm relative to the experimental CSs according to the selection algorithm outlined in Figure 13. However, the ensemble is highly flexible and mobile. The backbone RMSD to mean structure is 4.0 ± 0.8 Å (Table 14**)**. There was no structural core by which the models could be aligned; therefore, the models' Ser^3 C_β atoms were superimposed for visualization (Figure 21).

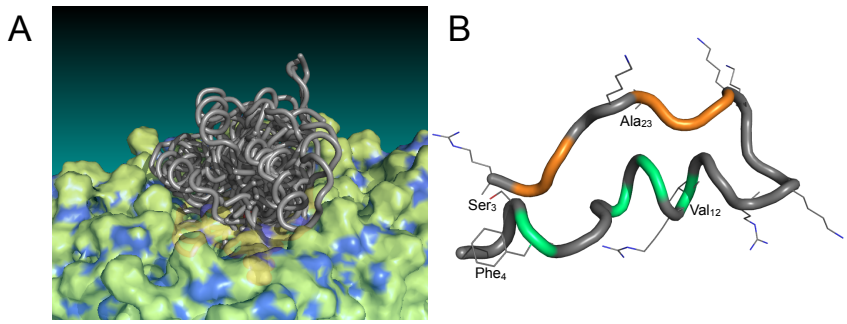

Figure 21: Structure of ghrelin based on MAS ssNMR chemical shift data
A) Final ensemble of ghrelin selected from the ensemble selection algorithm discussed in the main text. The Ser^3 C_α of each model was superimposed on the others. Ser^3 is shown as spheres. The ensemble was manually placed on the surface of a DMPC lipid bilayer, and the octanoic acid (spheres) was manually positioned in proximity to Ser^3. B) Model from the final ensemble. Residues predicted to be PPII helix (21-23 and 26-27) are colored in orange. Residues predicted to be helical according to Figure 12A (4, 8, 10, and 12) are colored in green. Positively charged residues (Arg and Lys), Ser^3, Phe^4, Val^{12}, and Ala^{23} are displayed as lines.

Table 14: Statistics for restraints, structural calculations, and structural quality for final ensemble of ghrelin models

NMR distance restraints used during folding and refinement	
Total restraints	55
Chemical shifts[a]	55
Structural statistics	
Number of models in ensemble	22
Deviations from idealized geometry	
Bond lengths (Å)	0.02
Bond angles (°)	0.7
Main chain RMSD to the mean structure (Å)	4.0 ± 0.8
Ensemble average RMSD to chemical shifts (ppm)	0.4
Ramachandran plot statistics (%)	
Most favored regions[b,c]	95.2, 99.5
Additionally allowed regions[b,c]	4.8, 0.5

[a] Chemical shifts were used during post-processing only; they were not used during fragment generation or *de novo* folding and refinement
[b] As determined by PROCHECK (http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/)
[c] As determined by MolProbity (http://molprobity.biochem.duke.edu)

Notice that the peptide exhibits an α-helical core but no β-strand character. After further inspection of the Ramachandran plot generated for all 10,000 models, as well as for the final ensemble, it appears that the final ensemble may exhibit some polyproline II helical character, which would be found in the $\phi = -75° / \psi = 150°$ area (Figure 22). Additionally, analysis of the ϕ/ψ angles using the PPII-DSSP method presented by Kabsch and Sander (182), residues 21–23 and 26–27 show significant PPII helical propensity (Figure 22B). The α-helical core agrees well with the secondary structure prediction of ghrelin based on PSIPRED (271), JUFO (104), SAM (272), and Figure 12. On the other hand, according to TALOS+ (143), which is based on the experimental CSs, the peptide, especially residues 21-23, is expected to be almost completely random coil. Ramachandran plots of residues 1-7, 8-12, 13-20, 21-28, and 21-28 indicate that the secondary structure of the final ensemble is not completely at odds with the secondary structure prediction or experimental CSs (Figure 22A).

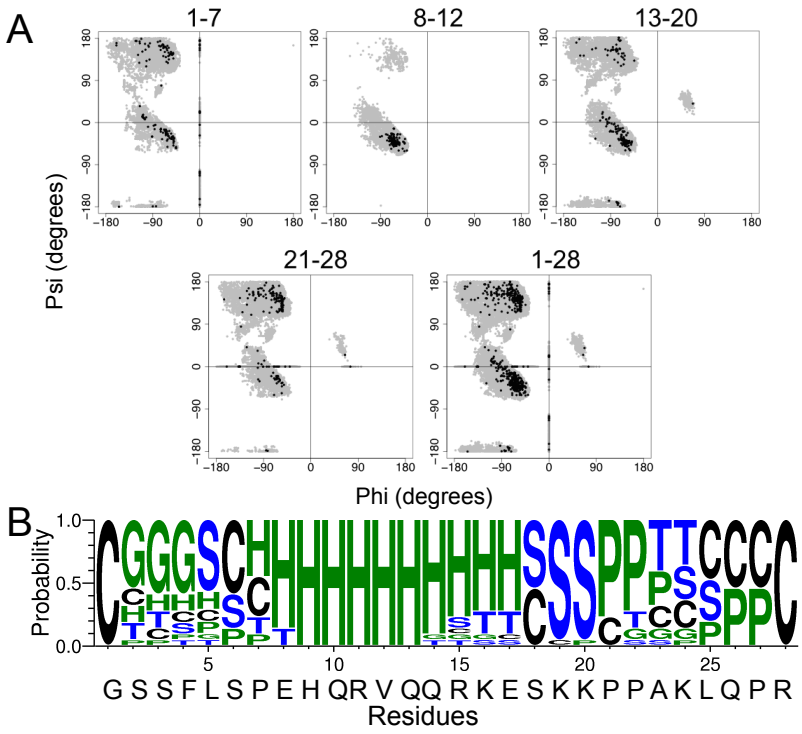

Figure 22: Secondary structure analysis of ghrelin
A) Ramachandran plots of various subsets of residues as labeled at the top of the plots. The torsion angles of all models generated in Rosetta (gray) and the final ensemble of models (black) are plotted. B) Weblogo (http://weblogo.threeplusone.com) of PPII-DSSP analysis of final ensemble of ghrelin models. Color key: black = random coil (C), blue = bend (S) or turn (T), and green = α-, 3_{10}-, or PPII helix (H, G, or P, respectively).

Discussion

Ghrelin interacts with the membrane via a small hydrophobic cluster

According to our spin diffusion studies, ghrelin interacts with the membrane via residues Ser[3] and Phe[4], whose side-chains and the octanoyl chain insert into the membrane (Figure 19); this is also in agreement with solution NMR data performed in detergent micelles (230). Due to the deuteration scheme of the membrane, ^{1}H spin diffusion can only originate from the glycerol backbone and the polar headgroup, suggesting localization of the Phe side-chain in this region. Generally speaking, the interface region of the membrane represents the preferred localization for membrane-

bound lipidated peptides (273, 274). Spin diffusion into residues Val^{12} and Ala^{23} is significantly slower, implying that these residues have no membrane contact. Due to the highly dynamic ghrelin structure at the membrane surface and the fact that the octanoyl chain is in equilibrium between an inserted state (~90% of the time) and a desorbed state (~10% of the time), spin diffusion from the membrane into the peptide is significantly slower than what is observed for membrane proteins with a transmembrane segment (249).

The small hydrophobic cluster of amino acids of octanoylated Ser^3, Phe^4, and Leu^5 account for about –13.4 kJ/mol (275, 276) of the energy corresponding to the ghrelin-membrane interaction. At the lipid concentrations used in our experiments, this is insufficient for a permanent association with the membrane. Using a simple membrane partition model (232), this would only account for binding of ~8% of ghrelin. Clearly, a second mechanism is required for anchoring ghrelin to the membrane. This second mechanism is electrostatic attraction of the positively charged C-terminal two-thirds of the ghrelin sequence to the lipid headgroups. Indeed, ghrelin holds an electrostatic charge of +5.8 at pH 6, which was used for our studies to prevent the hydrolysis of the octanoyl chain. Numerous calculations based on the Gouy Chapman theory have been carried out to determine the electrostatic contribution to membrane binding of lipidated peptides (277). For instance, pentalysine binds to a slightly negatively charged membrane, as in our case, with a Gibbs free energy of approximately –12 kJ/mol (278). Together with the hydrophobic contribution from the N-terminus of ghrelin, we estimate a total membrane binding energy of $\Delta\Delta G_0$ about –25 kJ/mol, which corresponds to approximately 90% of

bound ghrelin. This corresponds relatively well with the value of –28.6 kJ/mol determined from the binding measurement.

The octanoyl chain might play a role in a fine-tuned membrane association mechanism

Given the above observations, why is ghrelin not modified with a longer lipid chain, which would provide the peptide much better membrane partitioning properties? Clearly, the short octanoyl chain is not optimal for membrane binding. Further, chemical biology studies have shown that longer lipid chains and even more bulky groups are accepted by the GHSR (226). However, a replacement of the octanoyl chain to Ser^{6} or Ser^{18} is not tolerated. Further, the lack of the octanoyl chain, as in descyl-ghrelin, poorly activates the receptor. This could, however, also be explained by the fact that desacyl-ghrelin does not bind negatively charged membranes, as shown here and by others (62). However, a computer-generated model of the GHSR-ghrelin complex revealed hydrophobic contacts between the receptor and Phe^{4}, as well as the octanoyl side-chain (279). Apparently, the short ghrelin octanoyl chain is responsible for a fine-tuned membrane association mechanism, which catalyzes receptor binding and activation (228). Although there is some disagreement about the exact hydrophopic contribution of ghrelin to membrane binding, most studies agree that desacyl-ghrelin does not significantly bind membranes (62, 230). It is obvious that the ghrelin octanoyl chain has not been optimized for the purpose of membrane binding; longer acyl chains or prenyl groups provide much more favorable membrane anchors (232). The octanoyl chain is therefore primarily needed for receptor activation.

Previous studies of membrane-associated ghrelin and related peptides primarily indicate α-helical structure

Earlier ^{1}H studies of acylated and desacylated ghrelin in aqueous solution at low pH indicate that both forms of the peptide are highly unstructured in water. Indeed, the poor dispersion of CSs, as well as the lack of nuclear Overhauser effects (NOEs) typically seen of α-helices and β-sheets support the CD data (61). It is also possible that ghrelin experiences structural inter-conversion on a faster timescale than the NMR measurements, resulting in no detection of transient secondary structure. A 10-ns MD simulation performed in water at constant temperature and neutral pH (preceded by 2 ns of simulated annealing MD, or SAMD) provided evidence that ghrelin may sample a helix from residues 7 to 13 in both environments. MD studies in DMPC bilayers for 15 ns, initiated with the energy-minimized final peptide from the previous 10-ns MD simulation in water, did not show any significant differences in secondary structure from the peptide in aqueous conditions. However, the presence of the membrane appeared to reduce ghrelin's flexibility. Interestingly, the octanoyl side-chain, while initially pointed to the lipid bilayer, did not anchor the peptide to the membrane. Instead, during the simulation, residues 15–18 served as contact points with the lipid headgroups (63, 280).

CD spectroscopy of ghrelin and desacyl-ghrelin performed in aqueous solution (20 mM Tris buffer) and in 100% TFE at pH 7.4 provide experimental support for the MD studies, in that the acylated peptide exhibits 12% helical character in aqueous solution and in TFE. Desacyl-ghrelin, on the other hand, showed a significant increase in helical character when going from an aqueous environment (23%) to TFE (48%) (281).

In contrast to the MD simulations performed by Beevers and Kukol (63), Staes, *et al*. showed via a variety of biochemical assays that, while ghrelin and desacyl-ghrelin both electrostatically interact with the membrane, only acylated ghrelin penetrates into negatively charged membranes. However, the interaction of ghrelin with membranes was not such that it otherwise significantly disturbed the membrane surface. The same authors also investigated the secondary structure of ghrelin and desacyl-ghrelin via *in silico* modeling and CD spectroscopy. Similar to previous MD studies (63), an α-helix spanning residues Pro^{7} to Ser^{18}, which was flanked by two loops, for both acylated and desacylated ghrelin. The authors' models were supported by CD data collected in water, dodecylphosphocholine (DPC) micelles, SDS micelles, and TFE. For both forms of ghrelin, the helicity increased significantly in SDS micelles and TFE (62). A similar trend was observed for the prolactin releasing peptide (PrRP), another peptide that plays a role in food intake and body weight homeostasis (37). In the case of PrRP, it was demonstrated that the peptide likely exists in a conformational equilibrium between α- and 3_{10}-helix, and the helical propensity of the peptide is essential for its ability to activate the PrRP receptor, another GPCR. Another peptide that is involved in the regulation of appetite, galanin-like peptide (GALP), also shows nascent helical character, which may increase upon binding to galanin receptors (204). More recently, the neuropeptide, substance P (SP), was also found to have α-helical character in negatively charged SDS micelles and DMPG liposomes. However, in aqueous solution and in sub-micellar concentrations of SDS and DMPC liposomes, CD spectra indicate the presence of polyproline II (PPII) helix (282).

Ghrelin exhibits a highly flexible structure containing some polyproline II-, α-, and 3_{10}-helix

Based on previous structural studies, structural characterization of related peptides, and secondary structure prediction based on the primary sequence (Figure 12), it was expected that our current ssNMR studies would point to a dynamic peptide having transient α- and/or 3_{10}-helical character in conformational equilibrium. Interestingly, $^{13}C_{\alpha}-^{13}C_{\beta}$ CS values indicate helical propensity for residue 4, 8, 10, and 12, but according to CS index analysis with TALOS+ (143), only Arg^{11} exhibits a small amount of helical propensity (Figure 12). This is in agreement with the high mobility inferred from the low order parameters that have been measured for the peptide (Figure 18).

The final ensemble of Rosetta-generated models in best agreement with the experimental CSs provides a set of three-dimensional (3D) structures that allow for the visualization of the information obtained by NMR. As expected from the order parameters, we modeled a very loose conformational ensemble (Figure 22A and Table 14). Interestingly, while Rosetta sampled ϕ/ψ torsion angles expected for all common secondary structures (i.e., α-helix and β-sheet), the final ensemble exhibits a strongly helical core with what initially appeared to be "random coil" in the N- and C-terminal region, in agreement with previous studies (41, 62, 63, 213, 214). However, upon closer inspection of the Ramachandran plots and a modified DSSP analysis of these 22 models, it is probable that the final ensemble exhibits a small amount of 3_{10}-helical character, as well as a significant amount of PPII helix, especially for Pro^{21}-Ala^{23} and Gln^{26}-Pro^{27} (Figure 21 and Figure 22). The helical character of ghrelin does appear to allow it to

frequently adopt amphipathic conformations, thus allowing the basic residues to interact with the membrane's polar headgroups (Figure 21).

Polyproline II helical conformation in ghrelin may play a biologically significant role

Pure PPII helix is left-handed, is often characterized as a triangular prism, and has a helical pitch of 9.3 Å/turn; it contains ϕ and ψ angles of –75° and 145°, respectively. However, other amino acids and combinations of amino acids can form PPII helices.(283) Stapley and Creamer state that, in addition to Pro, Gln and positively charged residues having an increased probability of existing in PPII helices, Gly and aromatic residues show decreased probability (284). Other analyses of proteins of known structure agree that Gly and aromatic residues have low propensities to form PPII helices and that Pro appears most often, the increased observation of Gln and positively charged residues in PPII helices is disputed (285). In addition to being sampled during protein folding and unfolding, PPII helical structure has been implicated in amyloid formation (286, 287), nucleic acid binding (288), and muscle tissue elasticity (289). Statistical analysis of a database of 274 non-homologous protein structures shows that, while only 2% of residues are found in PPII helices, more than half of all polypeptide chains contain PPII helix of at least three residues in length (284).

To our knowledge, ghrelin is the first membrane-associated peptide to have PPII helical character for some residues in the presence of lipid bilayers. We point out that this character is likely transient and involves only short stretches of 2-3 residues. However, it is possible that, like the aforementioned peptides, ghrelin's α-helical content increases and extends into the ten C-terminal residues when it binds to GHSR. However, the PPII

helical character could allow for increased solvent accessibility while simultaneously providing for structural flexibility in areas such as flanking α-helices, linker regions, etc. Further, PPII helices have been found to be structural motifs involved in protein-protein interactions, which may result from their tendency to form amphipathic helices and to bind in a rapid and reversible fashion (267, 290).

Reliance of peptide fragment selection on chemical shifts and secondary structure prediction

While CSs can be used to guide fragment selection for *de novo* folding in Rosetta, we ultimately chose to utilize the original fragment selection protocol and filter by CS agreement after modeling was completed. Given that CS data for ghrelin is sparse, i.e. only for a subset of all residues secondary structure can be determined from the CSs, this protocol was chosen to prevent biases from residues with determined CSs on other regions of the peptide. In the case of ghrelin, Rosetta selects fragments based on agreement of CS for a subset of residues with little or no secondary structure information for other residues. When generating fragments for ghrelin, this led to a bias of β-hairpin fragments, which was not in agreement with other experimental data that pointed to a highly flexible and mobile peptide. We therefore opted to select fragments based on predicted secondary structure for all residues and filtered the models based on agreement of experimental CS later. Indeed, upon analysis of the secondary structure of fragments selected with and without experimental CSs, we see that the fragment selection scheme depends heavily on CS data when available, as is described in the literature.(106) On the

other hand, when CS data are not included in fragment selection, the secondary structure prediction of all residues is critical (Figure 12).

PROSHIFT gives systematically best agreement between experimental and predicted CS values

In order to compare the Rosetta-generated models with the experimentally determined CSs, we tested four CS prediction methods: PROSHIFT, SPARTA+, SHIFTX, and SHIFTX2. While SPARTA+, SHIFTX, and SHIFTX2 performed similarly, PROSHIFT appears to be the best method for prediction of CSs for ghrelin (Figure 20). To rule out systematic error and artifacts, one low-energy model that had minimal deviations between predicted and experimental CSs was chosen for in-depth analysis (Table 13 and Figure 23). This was also carried out on a few randomly selected models (data not shown).

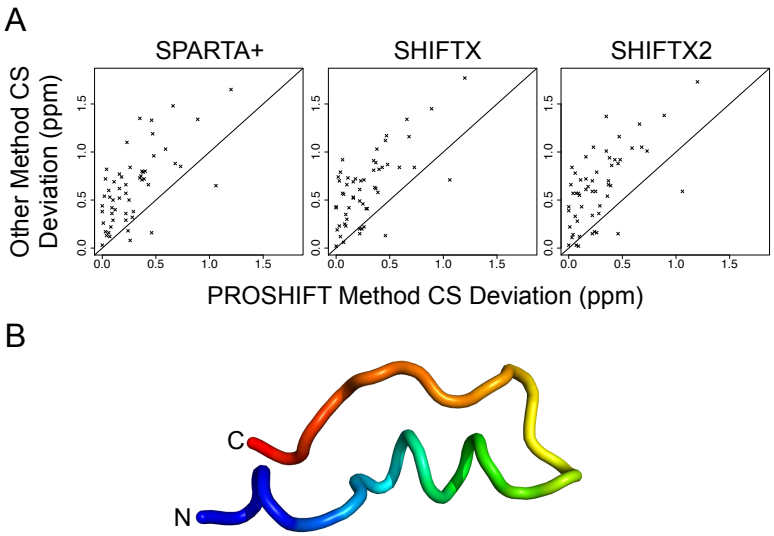

Figure 23: In-depth analysis of chemical shiftx for one model
A) The $CS_{SPARTA+}$, CS_{SHIFTX}, and $CS_{SHIFTX2}$ deviations from $CS_{experimental}$ values (Table 13) are plotted against $CS_{PROSHIFT}$ deviations from $CS_{experimental}$ values. All values are in ppm. B) Structure of the model chosen for in-depth analysis. While it was not in the final ensemble of models reported in this work, it was within the top 10% by Rosetta energy and the best or second-best model with respect to CS-RMSD relative to experimental CSs.

This result was somewhat surprising, given that PROSHIFT is an older method than the three to which it was compared. The reason for PROSHIFT's superior performance is not obvious, especially considering that the same 55 (C_O, C_α, C_β, and H_α) CSs were used for all analysis. Our explanation for this phenomenon is that PROSHIFT might be less biased than other methods in predicting CSs for well-structured proteins with large amounts of secondary structure, thereby making it more suitable for prediction of CSs for peptides or intrinsically disordered proteins.

A combined ssNMR-Rosetta protocol for studying structure and dynamics of flexible peptides and proteins

Due to the lack of regular inter-residue hydrogen bonding characteristic of α-helices and β-strands, it is likely that PPII helices are often categorized as "random coil" by secondary structure analysis software, such as DSSP. Furthermore, while PPII helices are difficult to detect directly by NMR (285) there have been attempts using CS data (291, 292). More generally, determining the structural ensemble that best represents sparse NMR CSs is especially challenging for biomolecules expected to be highly flexible and potentially unstructured. In addition to presenting a 3D structural ensemble of the biologically active form of ghrelin, we provide a novel, thorough method for predicting membrane-associated peptides, as well as for selecting a set of models based on ssNMR CSs. As NMR is often used to characterize protein unfolding and intrinsically unstructured proteins (IUPs) (289, 292-294), we believe our approach of combing NMR with Rosetta and a Monte Carlo ensemble selection algorithm may be useful for future studies of other structurally flexible and mobile systems.

Conclusion

To date, the results on the structure and dynamics of ghrelin have been controversial and inconclusive. In order to elucidate the mechanism by which ghrelin interacts with the membrane, as well as its 3D structure and its dynamics in the membrane environment, CSs and order parameter data were collected via MAS ssNMR. The primary sequence of ghrelin was then used to *de novo* the peptide in Rosetta using the RosettaMembrane energy functions. A final ensemble of models was then selected

based on the CS data. Unlike other peptides that activate GPCRs and in contrast to previous studies of ghrelin, our model of ghrelin is extremely flexible (4-Å RMSD) while strongly sampling both α- and PPII helical character. This unique secondary structure may allow the peptide to adopt an amphipathic structure, which would allow it to bind electrostatically to the membrane. Finally, the protocol employed to fold ghrelin and select the final ensemble of models can be used to structurally characterize other flexible proteins and peptides for which only sparse CS data are available, including those that act in a lipid environment.

Availability

The protocol capture for comparative modeling, CS prediction, and ensemble selection, can be found in Appendix C. The coordinates for the final ensemble will be available on a hard drive upon final submission of the dissertation.

Acknowledgements

The authors would like to thank Greg Sliwoski for providing template PDB files and advice for comparative modeling, Sam DeLuca for assistance in writing the code for the ensemble selection algorithm. We acknowledge Vanessa Grote and Regina Reppich-Sacher for support in peptide synthesis and analysis. The study was supported by the Europäischer Sozialfonds (ESF 22117016)) and U.S. National Institutes of Health (NIH) F31-GM100742. Work in the Meiler laboratory is supported through NIH (R01 GM080403) and NSF (CHE 1305874).

CHAPTER IV

ROSETTAEPR: AN INTEGRATED TOOL FOR PROTEIN STRUCTURE DETERMINATION FROM SPARSE EPR DATA

This work is based on publication (Hirst, Alexander, Mchaourab, and Meiler, 2011).

Summary

Site-directed spin labeling electron paramagnetic resonance (SDSL-EPR) is often used for the structural characterization of proteins that elude other techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR). However, high-resolution structures are difficult to obtain due to uncertainty in the spin label location and sparseness of experimental data. Here, we introduce RosettaEPR, which has been designed to improve *de novo* high-resolution protein structure prediction using sparse SDSL-EPR distance data. The "motion-on-a-cone" spin label model is converted into a knowledge-based potential, which was implemented as a scoring term in Rosetta. RosettaEPR increased the fractions of correctly folded models ($RMSD_{C\alpha}$ < 7.5Å) and models accurate at medium resolution ($RMSD_{C\alpha}$ < 3.5Å) by 25%. The correlation of score and model quality increased from 0.42 when using no restraints to 0.51 when using bounded restraints and again to 0.62 when using RosettaEPR. This allowed for the selection of accurate models by score. After full-atom refinement, RosettaEPR yielded a 1.7Å model of T4-lysozyme, thus indicating that atomic detail models can be achieved by combining sparse EPR data with Rosetta. While these results indicate RosettaEPR's

potential utility in high-resolution protein structure prediction, they are based on a single example. In order to affirm the method's general performance, it must be tested on a larger and more versatile dataset of proteins.

Introduction

Protein modeling with Rosetta can serve as an alternative means of structure elucidation

The vast majority of proteins in the Protein Data Bank (PDB) have been determined by X-ray crystallography or nuclear magnetic resonance (NMR) (1). However, a large number of biomedically relevant proteins continue to evade structural elucidation by these techniques due to membrane environment (295), high flexibility (296), and size (297). Alternative techniques, such as computational structure prediction methods, can be employed in order to define the structure of such proteins. The usual experimental bottlenecks, such as obtaining highly pure, concentrated samples of protein, are thereby avoided. Rosetta routinely folds soluble proteins of less than 150 amino acids correctly (298). It is generally among the top performers in the Critical Assessment of protein Structure Prediction (CASP) experiments (109, 111, 299-301). In addition, Rosetta's ability to obtain the correct fold of membrane proteins of various sizes and topologies has been demonstrated (105, 117, 130). More recently, Das, *et al.* introduced RosettaFold-and-Dock, which allows for the *de novo* structure prediction of homomeric proteins (302).

Rosetta's sampling and scoring capabilities for protein folding have been reviewed extensively elsewhere (101, 102, 110, 303). Briefly, the Rosetta *de novo* protein structure prediction algorithm is divided into two steps: low-resolution protein folding to

obtain the overall topology and high-resolution refinement of the backbone and side-chains. Metropolis Monte Carlo peptide fragment insertion is driven by a variety of knowledge-based potentials to rapidly predict protein folds. In high-resolution refinement, the protein backbone ϕ and ψ angles are perturbed while the overall fold is maintained. Side-chain conformations are predicted via a Metropolis Monte Carlo search of rotamer space, and all torsional degrees of freedom are subjected to gradient-based minimization.

Sparse NMR restraints can be combined with Rosetta to obtain atomic detail structures

While the algorithm described above performs well in the *de novo* prediction of relatively small, soluble proteins, effectively sampling protein conformational space remains the limiting factor in the accurate prediction of more complex proteins. To this end, distance and orientational restraints, such as those obtained by NMR, have been incorporated into the Rosetta protein folding protocol (106). Chemical shifts are converted into backbone torsional angle restraints, which are used in the generation of the peptide fragment libraries. Distance restraints from nuclear Overhauser effects (NOEs) are also employed in this process. Additionally, distance and orientaitonal restraints (NOEs and residual dipolar couplings, or RDCs, respectively) have been incorporated into the scoring function and are evaluated during protein folding. Bowers, *et al.* demonstrated that Rosetta, combined with a sparse set of NOEs (approximately one restraint per residue) and backbone chemical shifts, can produce models with atomic detail accuracy (168). Similarly, a combination of sparse RDCs and chemical shifts was used to produce correctly folded models (303). Shen, *et al.* have made significant

progress in improving the robustness and accuracy of CS-Rosetta with incomplete chemical shift datasets, obtaining atomic detail models based on much data that would otherwise be considered unsuitable for high-resolution structure determination (139, 263, 264).

SDSL-EPR offers an advantage over traditional structure determination techniques

Despite such advances, some proteins remain un-amenable to structure determination by these methods. Site-directed spin labeling electron paramagnetic resonance spectroscopy (SDSL-EPR) allows for structural studies of membrane proteins and large macromolecular assemblies in native or native-like environments (74, 146-148, 304-306). SDSL involves mutating residues of interest to cysteines, which can be reacted with a paramagnetic spin label, such as methanethiosulfonate (MTS). A sensitive structural probe at a known sequence position is created, forgoing the need to "assign" signals in the spectrum as is necessary in NMR spectroscopy. Additionally, resolution of SDSL-EPR is not limited by the size of the system. Similar to fluorescence and NMR spectroscopy, however, SDSL-EPR generates information concerning both the local environment of the spin label and the overall global fold of the protein. SDSL-EPR has been used to characterize conformational changes, such as those seen in MsbA (74, 147), rhodopsin (307-309), and KcsA (77, 146, 310). More recently, it has been demonstrated that the fold of a protein can be determined by structural restraints derived from SDSL-EPR data alone (134).

Atomic detail protein structure determination by SDSL-EPR is difficult and computationally demanding

Challenges in using SDSL-EPR structural data arise from the possible perturbation of the system by introduction of the spin label, sparseness of datasets resulting from the need to construct a dedicated mutant for every data point collected, and uncertainty in the position and dynamics of the spin label relative to the protein backbone. In the past, proteins have displayed a surprising robustness with respect to the introduction of spin labels (151, 157, 161, 311, 312). Molecular dynamics simulations (162) and crystallography (161, 313) have been employed to explicitly model the spin label in order to help interpret SDSL-EPR structural data. However, these calculations are relatively slow and computationally demanding. In addition, most studies of this nature are designed to examine a specific protein and are not easily expanded to other systems. For the purpose of protein structure determination, a faster, broadly applicable approach to relate the spin label position to the protein backbone is needed. As an exhaustive experimental mapping of intra-protein distances is infeasible given time and the labor intensiveness of the SDSL-EPR method, a limited dataset that unambiguously defines the fold of the protein needs to be defined (314).

RosettaEPR is designed specifically to work with sparse SDSL-EPR data

In 2008, Alexander *et al* introduced the implicit "motion-on-a-cone" model, or cone model (Figure 26B), which is based on the structure of the MTS spin label (Figure 26A) (134). This model was used to convert an observed spin label distance, d_{SL}, into an "allowed" range for the distance of the C_β atoms, $d_{C\beta} \in [d_{SL}-12.5Å, d_{SL}+2.5Å]$ (Figure

26C). The authors demonstrate that these distance restraints are sufficient to determine the structure of T4-lysozyme to atomic detail accuracy from 25 SDSL-EPR restraints The present study introduces RosettaEPR, which replaces the soft interpretation of the distance constraints used in the previous study with a knowledge-based restraint potential optimized for SDSL-EPR distance data. Alexander, *et al.* utilized RosettaNMR, with the consequence that all $d_{C\beta}$ distances falling within the allowed range were considered equally favorable during *de novo* folding. All other distances were disfavored using a quadratic penalty function (Figure 24). However, while the distance difference, $d_{SL}-d_{C\beta}$, falls within a wide range, values between 0Å and 5Å are more likely than values outside this range. We used the cone model, in combination with the PDB, to derive a probability function for $d_{SL}-d_{C\beta}$, which was then converted into a scoring function using the Boltzmann relation. We demonstrate that treatment of SDSL-EPR distance restraints with this scoring function is superior. Following the benchmarking presented in this paper, RosettaEPR will be made available to the scientific community.

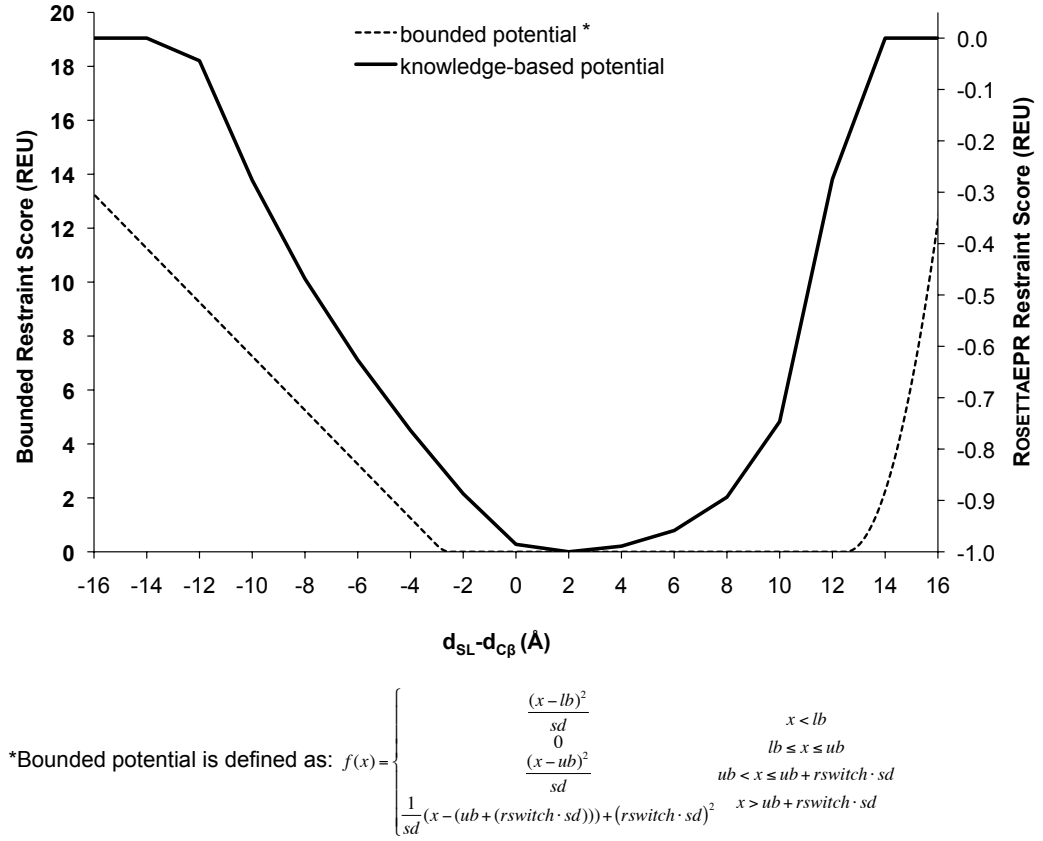

Figure 24: Comparison of the RosettaEPR knowledge-based potential with the bounded potential

The bounded potential against which restraint violations are scored is defined according to the equation reported in the figure, where *ub* = upper bound, *lb* = lower bound, *sd* = standard deviation of 1.0, and *rswitch* = 0.5

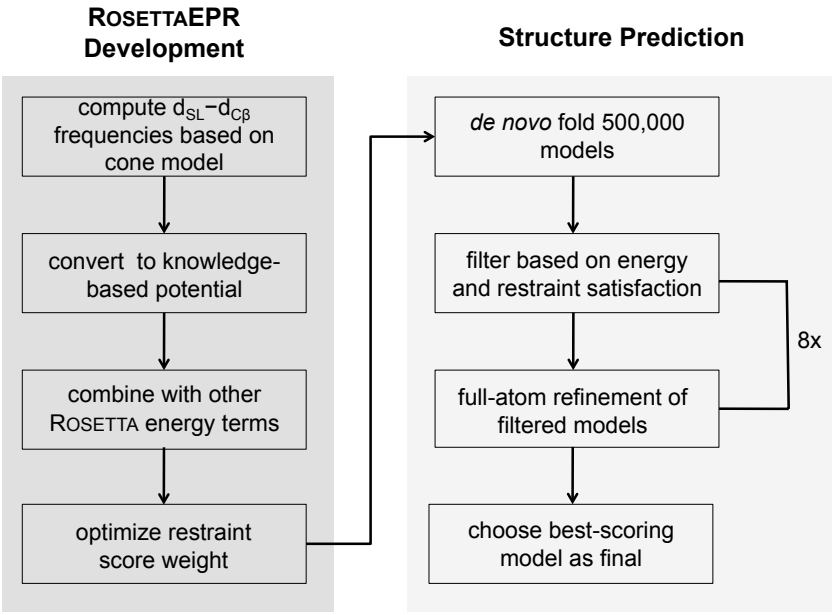

Figure 25: Flowchart outlining the currently described protocol

Materials and methods

The protocol described in the present work is outlined in Figure 25. It is divided into two subsections corresponding to the implementation and development of RosettaEPR and the prediction of the T4-lysozyme structure to atomic detail.

Conversion of the motion-on-a-cone model into a knowledge-based potential

The d_{SL}–$d_{C\beta}$ histogram (Figure 26D) was generated by placing a cone model-based simulated spin label at every exposed amino acid position in 3,584 proteins from a non-redundant protein database (315). That is, the simulated spin label was placed at residue positions that had a neighbor count (316) of less than ten, resulting in over 140 million measured distances. For every pairwise distance within each protein, the protein's $d_{C\beta}$ was subtracted from the simulated d_{SL} and stored in 0.5Å-wide bins. Because the

highest frequency of $d_{SL}-d_{C\beta}$ values was on the order of 10^6, a pseudocount of 10^6 was added to the total counts computed so that less commonly observed values are also considered.

The potential (Figure 26E) was calculated by taking the negative logarithm (−ln) of the propensity of each $d_{SL}-d_{C\beta}$ value, where the propensity is defined as:

$$propensity = \frac{\left(\frac{frequency + PseudoCounts}{TotalCounts}\right)}{\left(\frac{1}{\#\,bins}\right)}$$

PseudoCount equals 10^6, and # *bins* equals 64. The resulting values were normalized and shifted such that they were all negative. This relationship is based on the Boltzmann relationship, which is used to correlate a population of a species to an associated energy. The potential was re-scaled to give a maximum bonus of −1.0 for $d_{SL}-d_{C\beta}$ values between −12.0 and 12.0 (observed by the cone model) and a 0.0 penalty for values outside this range.

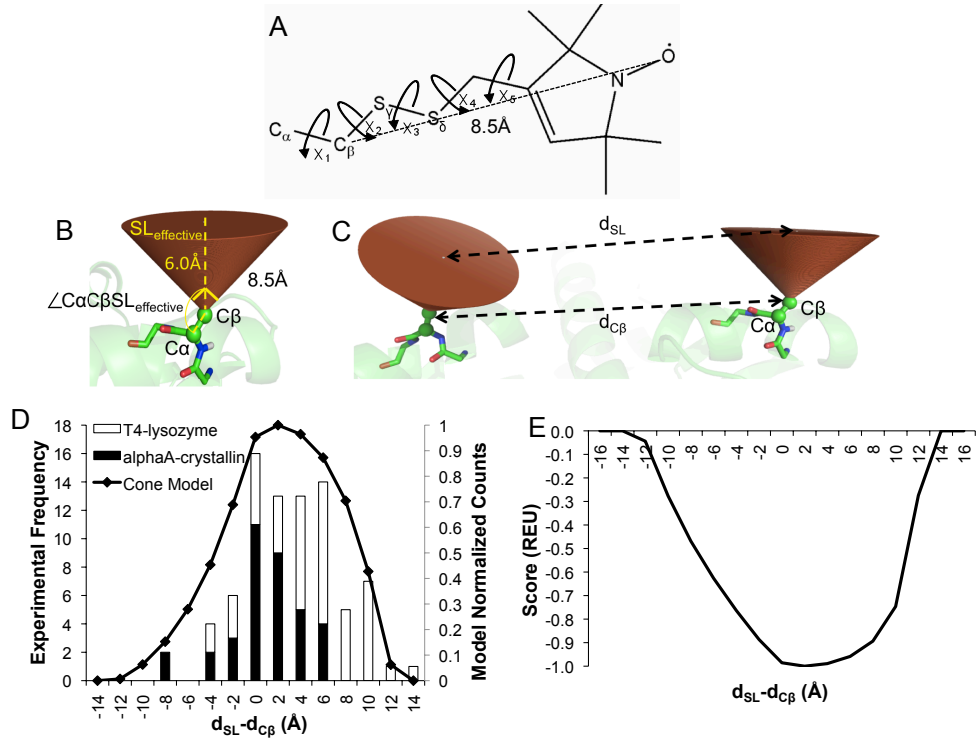

Figure 26: The "motion-on-a-cone" model
A) Methanethiosulfonate (MTS) spin label. The C_β-SL distance is approximately 8.5Å. B) In the cone model, the C_β-SL distance ($SL_{effective}$) is assumed to be 6Å, and the cone has an opening angle of 90°. The C_α-C_β-$SL_{effective}$ angle is restrained to angles $135° \leq (\angle C_\alpha C_\beta SL_{effective}) \leq 180°$. C) The cone model is used to calculate d_{SL}–$d_{C\beta}$ values. D) The normalized frequency of d_{SL}–$d_{C\beta}$ values for a database of proteins (black line, right y-axis) compared to experimentally observed values for T4-lysozyme and αA-crystallin (open and filled bars, respectively, left y-axis. E) The propensity of d_{SL}–$d_{C\beta}$ values can be converted into a knowledge-based potential according to the Boltzmann relation. The resulting energies were normalized such that the most favored d_{SL}–$d_{C\beta}$ value correlates with an energy of −1.0 Rosetta Energy Unit (REU), and the least favored d_{SL}–$d_{C\beta}$ value correlates with a Rosetta energy of 0.0 REU.

Model quality was assessed according to $RMSD_{C\alpha}$ relative to the 2LZM crystal structure

In order to best assess the ability of RosettaEPR to recover native-like folds, only the α-helical core domain of T4-lysozyme (residues 58-164) was modeled, as experimental restraints for other regions of this protein were not available. The experimentally determined distances used as restraints are reported in Table 15 and are mapped onto the T4-lysozyme crystal structure in Figure 27. Models of the protein were

generated a) without restraints, b) with restraints using RosettaEPR's knowledge-based potential, and c) with restraints defined by the same boundaries as those used by Alexander, *et al*. Model quality was assessed by computing the $RMSD_{C\alpha}$ relative to the X-ray crystal structure of T4-lysozyme (PDBID: 2LZM (317)). Only core residues 70-155, excluding loops, were considered in computing the $RMSD_{C\alpha}$ (see Table 16).

Table 15: T4-lysozyme EPR distance restraints in comparison with the crystal structure

AA1-AA2[a]	$d_{C\beta}$ (Å)[b]	d_{SL} (Å)[c]	σ_{SL} (Å)[d]	Reference
061-135	37.7	47.2	2.2	Borbat, *et al.*, 2002
065-135	34.3	46.3	2.2	Borbat, *et al.*, 2002
061-086	34.5	37.5	2.0	Borbat, *et al.*, 2002
065-086	28.9	37.4	2.7	Borbat, *et al.*, 2002
080-135	26.7	36.8	1.0	Borbat, *et al.*, 2002
061-080	28.7	34.0	2.2	Borbat, *et al.*, 2002
065-080	22.6	26.5	3.8	Borbat, *et al.*, 2002
119-131	13.2	25.0	5.0	Alexander, *et al.*, 2008
123-131	14.6	23.0	5.0	Alexander, *et al.*, 2008
065-076	16.8	21.4	2.8	Borbat, *et al.*, 2002
116-131	11.1	19.0	10.0	Alexander, *et al.*, 2008
119-128	10.4	19.0	4.0	Alexander, *et al.*, 2008
140-151	15.5	18.0	9.0	Alexander, *et al.*, 2008
089-093	9.8	16.0	3.0	Alexander, *et al.*, 2008
086-119	10.0	15.0	3.0	Alexander, *et al.*, 2008
120-131	10.5	14.0	3.0	Alexander, *et al.*, 2008
127-151	9.6	14.0	2.4	Alexander, *et al.*, 2008
140-147	10.1	13.0	7.0	Alexander, *et al.*, 2008
131-150	8.7	5.7	0.4	Alexander, *et al.*, 2008
127-154	5.9	7.0	3.0	Alexander, *et al.*, 2008
131-154	9.5	6.5	4.0	Alexander, *et al.*, 2008
134-151	10.7	7.0	0.8	Alexander, *et al.*, 2008
131-151	10.4	9.0	8.0	Alexander, *et al.*, 2008
088-100	8.9	<6.0	3.0	Alexander, *et al.*, 2008
089-096	8.4	<6.0	3.0	Alexander, *et al.*, 2008

a Indices of spin labeled amino acids with respect to the crystal structure
b Cβ distance as reported in the crystal structure
c Spin label distance as observed by EPR
d Standard deviation as observed by EPR

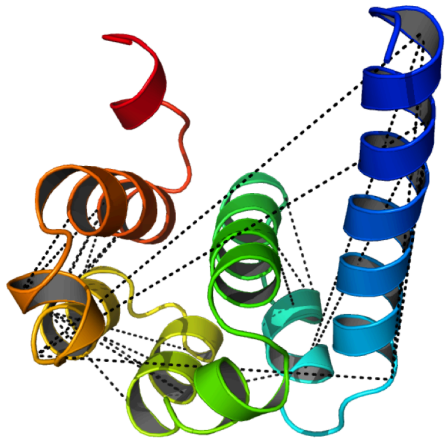

Figure 27: Map of EPR distance restraints on the T4-lysozyme crystal structure
The 107 C-terminal residues of the T4-lysozyme crystal structure are shown in rainbow with inter-residue distances used as restraints in RosettaEPR depicted as black dotted lines. A full list of experimentally determined EPR distances used in the benchmarking of RosettaEPR for this protein is reported in Table 15.

Table 16: Residues over which RMSDs and rotamer recovery were computed

RMSD	Rotamer Recovery
70-80, 82-90, 93-106, 108-113, 115-123, 126-134, 137-141, 143-155	74-75, 78, 84, 87-88, 91, 94-104, 106, 110-111, 113-114, 116-118, 120-121, 125-126, 128-130, 132-134, 136, 138-139, 145-153, 156

Weight optimization for the knowledge-based SDSL-EPR restraint potential

To optimize the factor by which the RosettaEPR scoring function should be applied, 10,000 models of the α-helical region of T4-lysozyme were constructed for a wide variety of weights (Table 17). The fraction of models with $RMSD_{C\alpha}$ values below 7.5Å was taken as measure for the correct fold. The fraction of models with $RMSD_{C\alpha}$ values below 3.5Å was employed to identify candidate models for successful atomic detail refinement; models generated with this level of accuracy are considered to be "native-like." The knowledge-based potential was implemented as a spline approximation in the Rosetta AtomPairConstraint score. The bounded restraint uses the

AtomPairConstraint score as computed according to a bounded quadratic equation (Figure 24).

Rosetta was used to de novo *fold and refine T4-lysozyme*

Secondary structure prediction of the 107 C-terminal residues of T4-lysozyme was performed using Jufo (104), Psipred (271), and Sam (318). Peptide fragments to be used in *de novo* structure prediction were generated as previously described, and fragments based on homologous proteins were excluded during folding. Rosetta's low-resolution *de novo* protein folding algorithm was used to generate 10,000 models of T4-lysozyme guided by experimental restraints (Table 15) (134) weighted to various extents, resulting in models containing structural information of the protein backbone only. During *de novo* folding, residues are represented as superatoms, or "centroids" (102). After determining that the RosettaEPR knowledge-based potential optimally predicts the fold of T4-lysoyzme when multiplied by a factor of 4.0, this weight was used in the generation of 500,000 models of the protein.

The 500,000 models were filtered according to their overall Rosetta energy and the extent to which they satisfied the experimental restraints. Only the top 1% of models by total score that had a restraint score of at least 85% of the optimum value was included in the filtered ensemble. These 1,388 models were then refined to atomic detail, in which the centroids were replaced with side-chain rotamers based on a backbone-dependent rotamer library (319). During refinement, Rosetta's full-atom scoring potentials are used to guide refinement through an iterative cycle of side-chain repacking and gradient-based minimization (110, 320). Each round of refinement yielded ten times the initial number

of models. That is, one round of refinement resulted in 13,880 new, refined models. All *de novo* folding and full-atom refinement computations were performed using Rosetta trunk revision 34586.

Structure determination with RosettaEPR is computationally feasible

All models were generated by independent simulations using Vanderbilt University's Center for Structural Biology computing cluster and the university's Advanced Computing Center for Research and Education (ACCRE). Computations were performed on a combination of AMD Opteron and Intel Nehalem processor nodes. The average time needed to fold one model of the 107 C-terminal residues of T4-lysozyme was approximately 240 seconds. The same time is required for a single round of high-resolution refinement for one model.

Results

Knowledge-based potential reflects likelihood of model in light of observed SDSL-EPR distance

Cone model-based statistics were collected over a database of non-redundant proteins (see *Materials and methods*) and compared to $d_{SL}-d_{C\beta}$ values determined experimentally for T4-lysozyme and αA-crystallin (Figure 26D). The set of cone model statistics recovers several features of the experimental data, including the range of $d_{SL}-d_{C\beta}$ values and a shift towards $d_{SL}-d_{C\beta}$ values greater than 0Å. The shift towards positive $d_{SL}-d_{C\beta}$ values indicates that spin labels are more likely to point away from each

other. This is expected for soluble proteins, where mutations of surface residues are not expected to destabilize the protein.

For conversion into a knowledge-based potential, the negative logarithm (–ln) of the propensity of each $d_{SL}-d_{C\beta}$ value was computed such that less frequently seen $d_{SL}-d_{C\beta}$ values are considered less favorable than one that is more often observed (Figure 26E). In result, a restraint that is fulfilled in the most likely area of the distribution improves the total score by one point, and a restraint that is violated is not counted towards the total score. This knowledge-based potential was then incorporated into Rosetta's low-resolution scoring function where it is affiliated with a dedicated weight (see *Knowledge-based potential* section below). The current model is an improvement upon the original implementation of the cone model, in that a) protein structures, not ellipsoids, were used to generate the statistics, and b) the knowledge-based potential considers the likelihood of $d_{SL}-d_{C\beta}$ values instead of a simple binary classification.

Knowledge-based potential achieves up to 55% correctly folded T4-lysozyme models

Ten thousand T4-lysozyme models were folded *de novo* in the presence of the same restraints used previously (Table 15 and Figure 27) (134). Restraints were incorporated with various weights, and the results were compared to the bounded potential used by Alexander, *et al.* (Table 17). The usage of restraint scoring functions results in more native-like folds than when folding with no restraints at all (Figure 28 and Table 18). This reaffirms that experimental data increases sampling of more native-like structures. RosettaEPR recovers the native topology of the T4-lysozyme α-helical region in up to 55% of the models. This compares to 7% if no restraints are used and 42% when

using bounded restraints. Furthermore, folding with bounded restraints consistently resulted in approximately 1.0-1.5% of all built models having native-like conformations, compared to 2.1% when using the EPR knowledge-based potential with an optimal weight of 4.0. This improvement is significant, as additional starting structures for high-resolution refinement increase the chance of successfully obtaining atomic detail models (see *Ten-fold enrichment of low-RMSD models*). Further, conversion to a knowledge-based potential enabled fine-tuning of the weight of the SDSL-EPR potential for optimal performance, while the bounded potential provided constant suboptimal performance over wide ranges of the weight.

Table 17: Benchmarking results of T4-lysozyme using no restraints, 25 restraints scored according to the RosettaEPR knowledge-based potential, and 25 bounded restraints

Weight	% Models with $RMSD_{C\alpha}$ < 3.5Å	% Models with $RMSD_{C\alpha}$ < 7.5Å	% Models with $RMSD_{C\alpha}$ < 3.5Å	% Models with $RMSD_{C\alpha}$ < 7.5Å
0	0.03	7.17		
	RosettaEPR		**Bounded**	
1	0.73	21.98	0.89	37.56
2	1.41	31.07	1.18	40.95
3	2.01	37.20	1.58	41.84
4	2.05	42.08	1.62	41.09
5	1.83	45.65	1.43	40.44
6	1.60	47.29	1.40	39.50
7	1.35	49.60	1.40	38.42
8	1.31	51.21	1.62	38.01
9	0.87	50.89	1.59	37.42
10	1.02	52.70	1.57	37.22
20	0.51	54.89	1.44	34.02
30	0.46	53.28	1.22	32.77
40	0.25	49.74	1.27	32.16
50	0.17	47.43	1.12	32.27
60	0.07	43.86	1.01	31.07
70	0.03	43.95	1.29	31.67
80	0.02	43.07	1.34	31.05
90	0.01	40.92	1.39	31.22
100	0.01	41.11	1.12	30.62

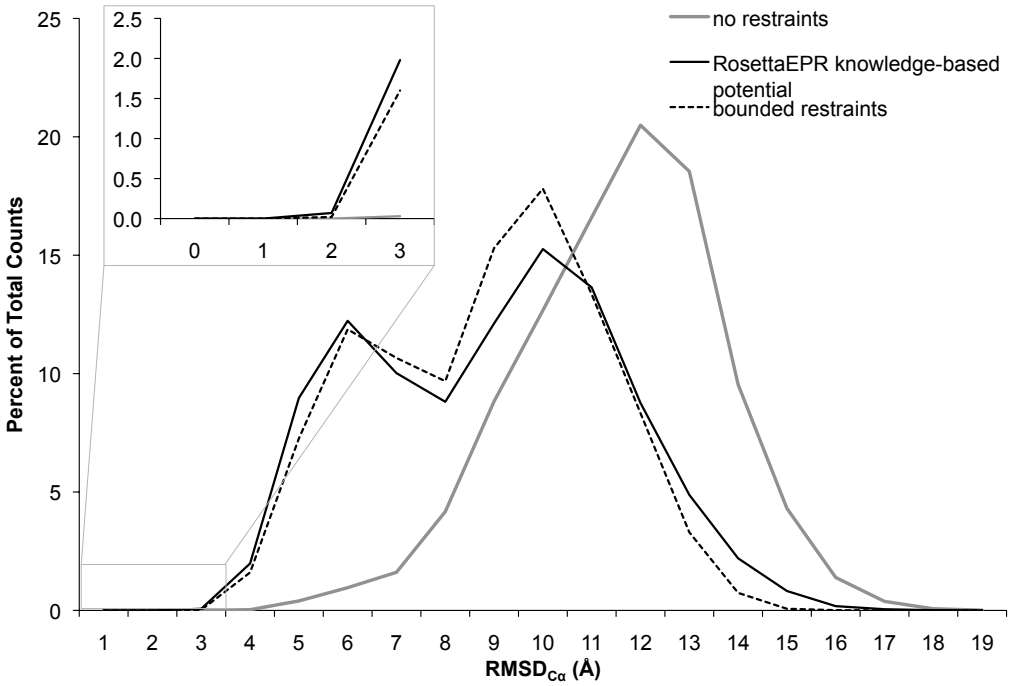

Figure 28: Comparison of the RosettaEPR knowledge-based potential to the bounded potential

T4-lysozyme was folded *de novo* in Rosetta guided by 25 experimental restraints. Restraint violations were scored according to either a bounded potential or the EPR knowledge-based potential. The $RMSD_{C\alpha}$ distributions of the resulting models when folded with optimally weighted restraint energies are compared to folding without restraints.

Table 18: Summary of benchmarking results of T4-lysozyme using no restraints, 25 restraints scored according to the optimally weighted RosettaEPR knowledge-based potential, and 25 bounded restraints with a weight of 4.0[a]

Restraint Type	% Models with $RMSD_{C\alpha}$ < 3.5Å	% Models with $RMSD_{C\alpha}$ < 7.5Å	Enrichment[b]
none	0.03	7.17	--[c]
knowledge-based potential (weight = 4.0)	2.05	42.08	7.0
bounded restraints (weight = 4.0)	1.62	41.09	5.3

[a] Results for all tested weights reported in Table 17

[b] *Enrichment = (fraction of low-RMSD models in filtered ensemble) ÷ (fraction of low-RMSD models of all models generated)*; *filtered ensemble* = within the top 1% of models by total score, the top 35% of models according to restraint score

[c] *Enrichment* could not be computed as with the other data sets due to lack of restraint score

Knowledge-based function improves correlation of score and model quality

The correlation of the scoring function with model quality is key to selection of native-like models when the structure is not known. The correlation coefficient improves from 0.42 in the absence of restraints to 0.51 when using the bounded function and further to 0.62 when using RosettaEPR (Figure 29). To quantify the value of the score for filtering native-like models, the enrichment for each optimized scenario was also computed (see Table 18). For the knowledge-based potential weighted by a factor of 4.0, the benchmark resulted in an enrichment of 7.0. The same analysis was performed on the models folded with the equally weighted bounded restraint potential, resulting in an enrichment of 5.3. The ensemble of models generated with no restraints contained only three native-like models, all of which were among the 10% best-scoring models, but this method was unable to produce enough native-like models to justify any high-resolution refinement.

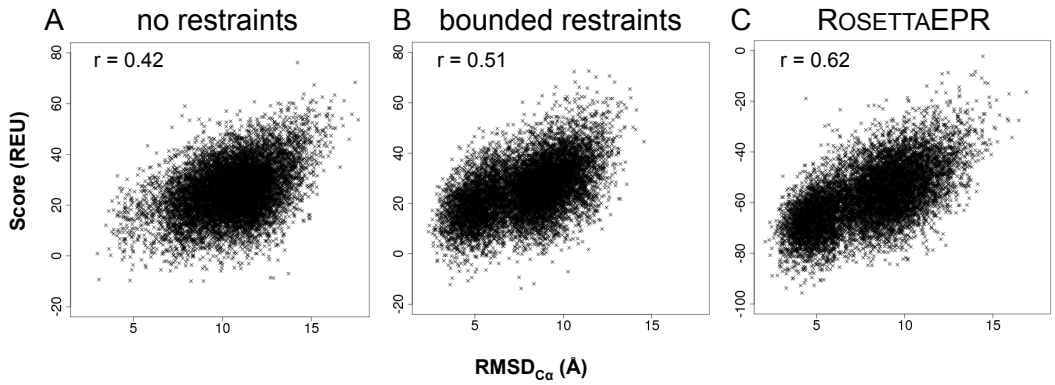

Figure 29: Correlation between total Rosetta energy and $RMSD_{C\alpha}$ of *de novo* folded models
Score vs. $RMSD_{C\alpha}$ for 10,000 models *de novo* folded A) with no restraints, B) with 25 bounded restraints, and C) with 25 restraints guided by the RosettaEPR knowledge-based potential

Ten-fold enrichment of low-RMSD models through knowledge-based SDSL-EPR score for high-resolution refinement

500,000 models of T4-lysozyme were *de novo* folded in Rosetta guided by 25 EPR distance restraints (weight equals 4.0). From the 1% best-scoring models, models achieving at least 85% of the optimal knowledge-based restraint score were selected for high-resolution refinement. The enrichment of native-like models in the filtered pool was 10.6, while the enrichment of correctly folded models was 2.3, where enrichment was defined as the fraction of native-like or correctly folded models in the filtered pool divided by the fraction of native-like or correctly folded models in the entire ensemble. Filtering decreases the number of models considered for high-resolution refinement to a more manageable ensemble and enriches the fraction of low-RMSD models such that more native-like folds are refined to full-atom detail.

High-resolution refinement of T4-lysozyme yields structural model that is accurate at atomic detail

The resulting 1,388 models of T4-lysozyme were refined to high-resolution using Rosetta's full-atom potentials, which include knowledge-based van der Waals attraction, repulsion, hydrogen bonding, solvation, and electrostatic terms (110). Each input model was refined ten times without experimental restraints, resulting in 13,880 models. Ideally, low-RMSD models would be considered energetically favored according to Rosetta's scoring function. Therefore, the models were then filtered such that only the top 10% by total score were carried on to the next round of refinement. This process was repeated through eight iterations, at which point the score of the refined models converged. The

total score of each model was plotted against its $RMSD_{C\alpha}$ (Figure 30A). The correlation between energetically favorable and low-RMSD models improves after each round of refinement until it converges after the eighth iteration. The lowest energy model produced with this strategy had an $RMSD_{C\alpha}$ of 1.76Å relative to the native (Figure 31), and the lowest $RMSD_{C\alpha}$ observed was 1.73Å. The previously reported model was determined to have an $RMSD_{C\alpha}$ of 1.66Å.

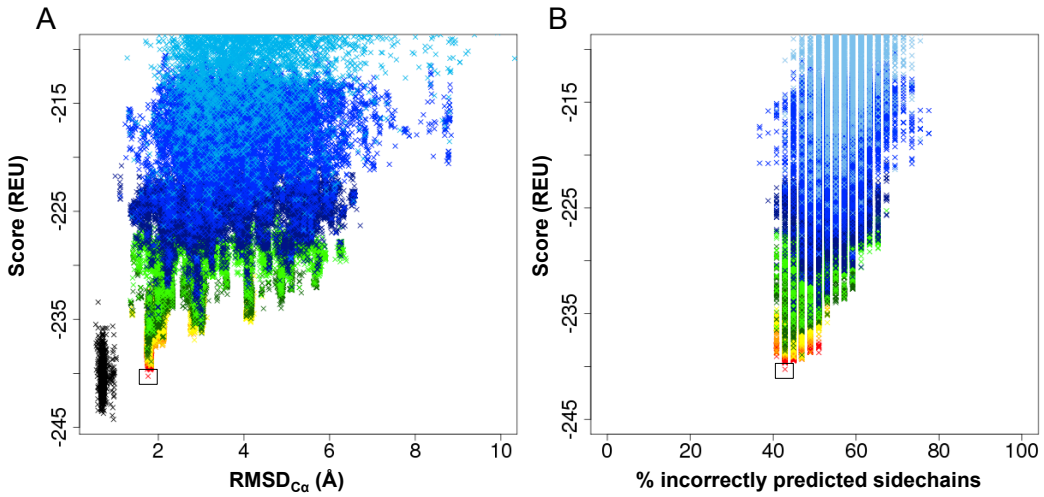

Figure 30: Correlation between Rosetta energy and $RMSD_{C\alpha}$ of refined models
A) Score vs. $RMSD_{C\alpha}$ plot of T4-lysozyme models for eight cycles of full-atom refinement. Each cycle of refinement resulted in ten times the number of input models. After each cycle, the refined models were filtered by total Rosetta energy, and the top 10% were refined again. Color key: refined crystal structure – black; round 1 = sky blue; round 2 = bright blue; round 3 = dark blue; round 4 = light green; round 5 = dark green; round 6 = yellow; round 7 = orange; round 8 = red. B) Percent of incorrectly predicted side-chains of core residues (see Table 16) as a function of total Rosetta score. The same coloring scheme in Panel A was used.

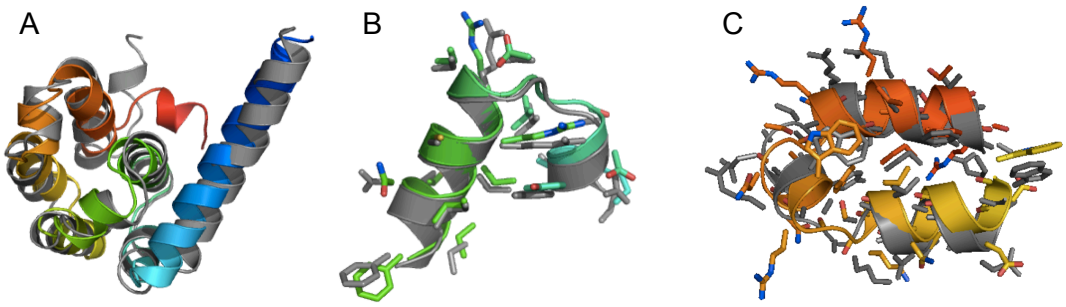

Figure 31: Atomic detail model of T4-lysozyme *de novo* folded with RosettaEPR
A) Superimposition of the lowest-scoring model of T4-lysozyme (rainbow) with the 2LZM crystal structure (gray). The $RMSD_{C\alpha}$ for the lowest-scoring model to the native is 1.76Å. Side-chains are displayed as sticks. B) Residues 86-104. C) Residues 126-154

The ability of Rosetta to recover native-like side-chain conformations was tested by comparing side-chain rotamer agreement of refined models of T4-lysozyme with the X-ray crystal structure. A rotamer of a given amino acid residue is defined by its χ_{1-4} angles. Side-chain conformations are classified by assigning them to the closest rotamer in terms of χ_{1-4} angle deviation (319, 321). The total Rosetta energy is plotted as a function of the percentage of incorrectly predicted side-chain rotamers (Figure 30). In general, the Rosetta energy correlates well with rotamer agreement, with the percent of correct rotamers predicted increasing after each round of refinement.

Discussion

The RosettaEPR knowledge-based potential proves to be superior to the bounded potential during de novo *folding*

We have demonstrated the advantages of using a knowledge-based potential to convert EPR distance data into structural restraints. The potential is derived from the cone model (134) and has been shown to perform better than a simple bounded potential.

From a conceptual standpoint alone, the energetic bonus correlates with the likelihood of observing $d_{SL}-d_{C\beta}$ values. As a result, the knowledge-based potential inherently uses the structural information from SDSL-EPR data more completely compared to the bounded scoring function used by Alexander, *et al*. Furthermore, the knowledge-based potential, in combination with Rosetta's low-resolution scoring function and *de novo* folding algorithm, proves more robust in obtaining low-RMSD models of T4-lysozyme, from which atomic detail structures can be generated through full-atom refinement.

The correlation between score and RMSD improves through multiple rounds of refinements

The Rosetta full-atom scoring function allows the most native-like model to be identified unambiguously by its overall score, if model accuracy is better than 2.0Å. This model should have the lowest overall Rosetta energy and therefore exhibit not only the correct topology, but also native-like side-chain and backbone conformations. Similarly, less favorable conformations should have higher computed energies; these models will also have higher computed RMSDs relative to the native structure. One therefore expects to observe an energy "funnel" after several rounds of full-atom refinement, where both the score and RMSD of the models converge to the native structure. The overall scores of the predicted models of T4-lysozyme are plotted against their $RMSD_{C\alpha}$ relative to the crystal structure in Figure 30. The correlation improves after each round of filtering and refinement, resulting in several atomic detail models with Rosetta energies comparable to the 2LZM crystal structure, which was refined using the same potentials as the predicted models.

RosettaEPR will be developed continuously as more data become available

Although a larger benchmarking set would be ideal, there are a limited number of systems for which both experimentally determined three-dimensional structures and EPR data can be obtained. However, the resulting atomic detail models of T4-lysozyme generally satisfy the experimental EPR data, and benchmarking will be expanded to more diverse systems as more data become available. In the mean time, a larger benchmark on a variety of proteins of known structure using simulated data will be performed to assess the general performance of the method. The current work serves as a proof of principle. It will be interesting to test whether the similar results will be obtained for other proteins. It has already been shown that NMR restraints greatly aid Rosetta's ability to recover native-like models (106, 136, 168, 303, 322), a method which is widely applicable to other biological systems, including the fumarate sensor DcuS (137) and a chordin-like cysteine-rich (CR) repeat from procollagen IIA (323). It is believed that the same will be true with RosettaEPR after further testing and refinement.

Sparse SDSL-EPR distance data alone are not able to yield atomic detail models

SDSL-EPR affords several advantages over other structure determination techniques, such as X-ray crystallography and NMR. No crystallization is required, there are few size constraints, proteins, and membrane proteins in particular, can be studied in a native-like environment, and there is no need to assign resonance signals. Thereby, SDSL-EPR overcomes some experimental limitations in the high-resolution structure determination of proteins that are large, highly flexible, or natively reside in lipid bilayers.

However, while quantitative in nature, the structural information obtained by SDSL-EPR is limited due to the flexibility of the spin label, which adds large uncertainties to the distances determined. Introduction of spin labels into proteins requires removal of native cysteine residues without affecting the protein structure and assumes that the spin label does not perturb the structure. Datasets obtained by SDSL-EPR remain sparse due to the requirement to create a dedicated double-mutant for each distance to be measured. Therefore, SDSL-EPR a) will be applied to systems where crystallography and NMR spectroscopy are not applicable and b) will be combined with crystallography and other techniques to study structural dynamics of proteins.

The current work and the results presented by Alexander, *et al.* (134) provide the first indication that sparse (approximately 0.25 restraints per residue) SDSL-EPR distance data can be combined with Rosetta for *de novo* protein structure elucidation with atomic detail accuracy. While RosettaEPR can be applied to soluble proteins, it is expected that the need and applicability of RosettaEPR will be highest for the structure determination of membrane proteins, the majority of which continue to evade more traditional techniques. A benchmark of RosettaEPR involving more proteins and membrane proteins in particular will be executed as suitable datasets become available

RosettaEPR will be accessible to the scientific community

Other researchers will have access to RosettaEPR via software licenses granted by the RosettaCommons (www.rosettacommons.org). These licenses are free for academic and non-profit institutions. To encourage usage of RosettaEPR, web tutorials will be made available.

Conclusion

RosettaEPR is the first tool designed to generate high-resolution protein structures from sparse EPR data. It can also be used in combination with an optimized restraint-selecting algorithm (314) to assist experimentalists in determining protein structures to high-resolution. In the future, RosettaEPR will be modified such that it can be used to effectively determine the structures of membrane proteins, an EPR accessibility knowledge-based potential will be implemented, and high-resolution modeling of the MTS spin label will be included. The ultimate goal of this research is to optimize the structural information that can be achieved through EPR spectroscopy. RosettaEPR will enable the high-resolution structure elucidation of a plethora of proteins for which structures have, until now, not yet been determined.

Acknowledgements

We would like to thank members of the Rosetta community for sharing their knowledge of various aspects of the software. We are specifically grateful to Kristian Kaufmann, Samuel DeLuca, and Kelli Kazmier for their insight and assistance throughout the development of RosettaEPR. This work was funded in part by the NIH R01 GM080403 to Jens Meiler and GM077659 to Hassane Mchaourab. Nathan Alexander is funded by NIH NIMH Award Number F31MH086222.

CHAPTER V

ROSETTATMH: MEMBRANE PROTEIN STRUCTURE ELUCIDATION BY COMBINING EPR DISTANCE RESTRAINTS WITH ASSEMBLY OF TRANSMEMBRANE HELICES

This work is based on the manuscript submitted to *PLoS ONE* of the same title by Stephanie DeLuca, Samuel DeLuca, Andrew Leaver-Fay, and Jens Meiler

Summary

Membrane proteins make up approximately one third of all proteins, and they play key roles in a plethora of physiological processes. Even though significant advances have been made in structure determination methods, such as X-ray crystallography, nuclear magnetic resonance spectroscopy, and cryo-electron microscopy, integral membrane proteins make up less than 2% of experimentally determined structures. Furthermore, few computational methods for *de novo* folding of integral membrane proteins have been presented. One potential alternative means of structure elucidation is to combine computational methods with experimental EPR data. In 2011, Hirst and others introduced RosettaEPR; the authors showed that this approach could be successfully applied to soluble proteins. In this work, we present RosettaTMH, a novel algorithm for structure prediction of helical membrane proteins. A benchmark set of 34 proteins, in which the proteins ranged in size from 91 to 565 residues, was used to compare RosettaTMH to Rosetta's two existing membrane protein folding protocols: the published RosettaMembrane folding protocol ("MembraneAbinitio") and folding from an extended chain ("ExtendedChain"). In the absence of EPR restraints, RosettaTMH folds

more models having the correct topology than MembraneAbinitio in 8 cases, whereas it performs better in 9 cases in comparison with ExtendedChain. When EPR distance restraints are used, RosettaTMH+EPR outperforms MembraneAbinitio for 30 proteins and ExtendedChain+EPR for 14 proteins. RosettaTMH+EPR is capable of achieving native-like topologies for the majority of proteins tested, including receptors and transporters. For example, a model of rhodopsin of 4.9Å $RMSD_{100}$SSE accuracy to the crystal structure was achieved, and a model of 6.7Å accuracy was obtained for the 565-residue Na^+/galactose transporter, vSGLT. The addition of RosettaTMH and RosettaTMH+EPR to the Rosetta family of *de novo* folding methods broadens the scope of helical membrane proteins that can be accurately modeled with this software suite.

Introduction

Approximately one-third of all proteins are integral membrane proteins (MPs) (324), and, due to their prevalence in a wide variety of biological functions, MPs comprise more than half of all drug targets (25, 26, 325). However, of the > 100,000 proteins with experimentally determined three-dimensional (3D) structures in the Protein Data Bank (PDB) (1), only about 2,000 are MPs (295). Further, according to Stephen White's database of MPs of known structure (http://blanco.biomol.uci.edu/mpstruc/), fewer than 500 unique MP structures have been determined. This disparity between the importance of MPs and the available 3D structures reflects the technical difficulties associated with MP structure determination by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. To study MPs in their biologically relevant native conformation(s), a membrane mimic must be present during the experiment. While

X-ray crystallographers have developed techniques to obtain diffracting crystals, such as the use of femto-second crystallography (326, 327), robotics (328), and antibodies (66, 329, 330), MP crystallization remains a bottleneck. Line broadening due to slow tumbling times of large MPs embedded in membrane mimics and decreased sensitivity due to the presence of additional nuclei are often limiting factors for solution NMR spectroscopy. Cryo-probes, increasingly powerful NMR magnets, selective labeling, and the development of solid-state NMR techniques (331) are continuously pushing the MP NMR field forward, but challenges remain here as well (72, 332-334).

EPR spectroscopy can serve as an alternative means of membrane protein structural characterization

Site-directed spin labeling electron paramagnetic resonance (SDSL-EPR) spectroscopy may serve as another means of MP structure determination because it has a number of advantages compared to more traditional methods. For example, proteins can be studied in their native environment, such as in lipid bicelles or vesicles, and no crystallization is required. Further, it does not require large amounts of protein, which is important in the case of MPs that are often difficult to express and purify. EPR is also an extremely sensitive technique because it measures the resonance of two unpaired electrons, so the signal-to-noise ratio is largely uninterrupted, unlike in NMR spectroscopy (148, 154, 311, 335, 336).

However, EPR is not without its disadvantages. Like NMR spectroscopy, structure determination is indirect in that the spectroscopic data are first converted to structural restraints (134, 135). Also, for distance measurements, SDSL requires the

removal of all endogenous cysteines in the protein and the mutation of the residues of interest into cysteines. As a result, in contrast to NMR spectroscopy, only one inter-residue distance can be measured per experiment. This results in low throughput and sparse datasets. In addition, the spin label itself introduces uncertainty, as the distance between the paramagnetic spin labels, which are at the tips of long and flexible side-chains, is measured. This distance then needs to be converted into a structural restraint based on MP backbone coordinates (134, 135).

There is a need for novel de novo *membrane protein structure prediction tools*

In order to aid in MP structure determination, several computational methods have been developed. These methods can be divided into two categories: template-based, or comparative modeling, and *de novo* folding. Template-based methods, such has Modeller (95, 337), Rosetta (96), SWISS-MODEL (338), and I-TASSER (339), are commonly used when the structure of a homologous protein exists. Template-based modeling methods are so named because they require a structural template onto which a target sequence can be threaded. For the sequence in question, a template structure, whether it is a sequence homolog or a structure exhibiting the same expected topology, must first be identified. Next, often after performing one or more sequence alignments, the target sequence is threaded onto the 3D coordinates of the template structure, thus replacing the sequence of the template with that of the target (252).

In the case of MPs, it is often difficult to identify a suitable template structure. As mentioned previously, there are a limited number unique MP structures available in the PDB. Additionally, even though templates having a similar fold may exist, it is possible

that the sequence homology between the target and the template is too low to be confidently detected. For example, of the more than 20 experimentally determined structures of G-protein coupled receptors (GPCRs), the majority are of class A, or class 1 (http://gpcr.scripps.edu/index.html) (23), even though there are probably 5 or 6 GPCR classes (13). Similarly, while there are some structures of transporters, such as LeuT (340), vSGLT (341), BetP (342), and GadC (343), MPs having the LeuT fold perform a large variety of functions and diverge in sequence significantly and can belong to a number of different protein superfamilies (344). While comparative modeling based on an evolutionarily distant template can be useful for hypothesis generation, especially when combined with experimental methods in an iterative fashion (254), *de novo* structure prediction of MPs is needed when no structural template is available. Additionally, *de novo* folding methods allow for an unbiased exploration of the conformational space, which is one disadvantage of using template-based methods.

Even though significant advances are being made in the structure determination of GPCRs, progress appears slower for other MP folds. The MP structural biology community is still striving for the structure determination of biomedically significant proteins, such as hERG, hSERT, the NPY receptor, etc. (http://blanco.biomol.uci.edu/mpstruc) It is possible, or perhaps even likely, that these structures will be determined in the future, but in the mean time, the need for advances in computational methods for MP structure prediction persists.

Compared to template-based MP modeling methods, there are only a handful of tools for *de novo* folding of MPs. RosettaMembrane was introduced in 2006 (105) and was later expanded to include full-atom scoring potentials (130), but its capabilities were

limited to MPs of fewer than 150 amino acids. The addition of limited restraints derived from sequence conservation allowed for accurate modeling of larger proteins (117), but this method could only account for one restraint at a time. Furthermore, the utility of RosettaMembrane in its current state is limited. For technical reasons that originate in the RosettaMembrane code base, it is not possible to *de novo* fold MPs with multiple restraints, such as those obtained from NMR and EPR.

Other methods to predict membrane protein structure, such as FILM3, exhibited mild success for predicting large MPs, but they rely on correlated mutational information to score MP models. Of 71 MP sequences, FILM3 was able to correctly predict 100% of inter-helix contacts for 17 proteins. Upon, comparison with two-dimensional slices of the experimental structures, 9 predicted structures had the correct topology (119). EVfold_membrane is also a promising method for MP structure determination but again relies on information from evolutionary covariation (118). On the other hand, BCL::MP-Fold does not depend on mutational information. It reduces the conformational search space by assembling secondary structure elements (SSEs) combined with knowledge-based potentials (KBPs) to assess model quality (345). The disadvantage of BCL-generated models is the lack of inter-helix loop regions and, because they are comprised of idealized α-helices, it under-predicts secondary structural features often present in MPs, such as helical kinks.

RosettaTMH allows for folding of membrane proteins, both with and without experimental restraints

To address the limitations of previously reported MP *de novo* folding methods, we have developed RosettaTMH, which, like BCL::MP-Fold, assembles MP topologies via rigid body perturbations of transmembrane helices (TMHs). Further, 3- and 9-amino acid fragment insertions, as used in the traditional Rosetta *de novo* folding algorithm (100), are used to more thoroughly sample helical orientations and introduce bends and kinks. Throughout the *de novo* folding process, RosettaMembrane's MP-specific scoring functions are used (105, 130). However, in contrast to previously published RosettaMembrane folding protocols, RosettaTMH can be combined with multiple experimental restraints, such as inter-residue distance information from EPR. This additional feature allows for improved sampling of native-like topologies that are in agreement with empirical information.

In this work, RosettaTMH was benchmarked on 34 MPs of known structure. It was compared to the original RosettaMembrane folding algorithm, "MembraneAbinitio" (105) and the traditional fragment assembly-only method used for folding soluble proteins in Rosetta, "ExtendedChain" (100) (but using the RosettaMembrane scoring function). In order to assess the performance of combining RosettaTMH with experimentally obtained structural data, EPR distance restraints were simulated for all MPs in the benchmark set. The purpose of the benchmark was to determine if these restraints increase the sampling of native-like MP folds. The simulated distance restraints were generated using the BioChemical Library (BCL, http://bclcommons.vueinnovations.com/bclcommons) and the restraint-picking algorithm

introduced by Kazmier, *et al*. (314). We show that, by implementing the ability to fold MPs with structural restraints, native-like folds can be obtained for 33 MPs in the benchmark set.

Materials and methods

Setup of RosettaTMH parameterization and benchmarking datasets

Thirty-four α-helical MPs and MP subunits of known structure were chosen to test the RosettaTMH folding algorithm (Table 19). Nine of these proteins (in italics in (Table 19) were used for the initial testing and parameter optimization of the RosettaTMH protocol. The benchmarking set exhibits a wide range of sizes and topological complexity. The number of EPR distance restraints simulated was computed as:

$$\#restraints = 0.2 * \#aa_{\text{TMH}},$$

where *#restraints* refers to the number of simulated EPR restraints generated, and $\#aa_{TMH}$ refers to the amino acids in TMHs defined in the experimental structures. This number of restraints was chosen because it is on the order of the maximal number of distance restraints that have been obtained for several MPs (145-147, 346). Further, it is a good compromise between prediction accuracy and plausibility. The input files used (i.e., fragments, secondary structure prediction, span, lipophilicity, and native PDB files) were the same or based on those employed for benchmarking of BCL::MP-Fold (345).

Table 19: Proteins used for benchmarking

PDB	Protein Name	Domain	# Res	# TMH	Absolute Contact Order	# Restraints
Small MPs						
3SYO	subunit of G protein-gated inward rectifier K^+ channel GIRK2 (Kir3.2)	76–197	122	2	14.4	12
2BG9	subdomain of nicotinic acetylcholine receptor	A: 211–301	91	3	6.9	16
1J4N	subdomain of aquaporin water channel, AQP1	4–119	116	3	15.2	17
2KSF	subdomain of histidine kinase receptor, KdpD	396–502	107	4	11.9	13
1PY6[a]	subdomain of bacteriorhodopsin	77–199	123	4	13.3	20
2PNO	human leukotriene C4 synthase	A: 2–131	130	4	13.6	22
2BL2	subdomain of V-type Na-ATPase	12–156	145	4	20.7	25

Medium MPs						
2K73	disulfide bond formation protein, DsbB	1–164	164	4	15.5	19
2ZW3	subdomain of connexin 26 gap junction channel	A: 2–217	216	4	25.7	24
1IWG	subdomain of multidrug efflux transporter, AcrB	336–498	163	5	17.4	26
1RHZ	subdomain of protein-conducting channel, SecYE	A: 23–188	166	5	19.8	21
2YVX	subdomain of magnesium transporter, MgtE	A: 284–471	188	5	20.6	26
1OCC	subdomain of cytochrome C oxidase, aa3	C: 71–261	191	5	24.1	29
4A2N	isoprenylcysteine carboxyl methyltransferase	1–192	192	5	22.4	24

1KPL	subdomain of H^+/Cl^- exchange transporter, CIC	31–233	203	5	23.4	31
2BS2	subdomain of quinol:fumarate reductase	C: 21–237	217	5	17.5	29
3P5N	S component of the ECF-type riboflavin transporter	10–188	179	6	17.9	22
2IC8	rhomboid peptidase, GlpG (*E. coli*)	91–272	182	6	17.9	23
1PV6	subdomain of lactose permease transporter	1–190	189	6	28.3	33
2NR9	rhomboid peptidase, GlpG (*H. influenzae)*	4–195	192	6	17.6	24
Large MPs						
1OKC[b]	mitochondrial ADP/ATP carrier	2–293	292	6	25.8	34
3B60	subdomain of lipid flippase, MsbA	A: 10–328	319	6	25.7	52

2KSY	sensory rhodopsin II	1–223	223	7	20.1	37
1PY6	bacteriorhodopsin (full length)	5–231	227	7	25.2	36
3KCU	formate channel, FocA	29–280	252	7	29.7	33
1FX8	glycerol facilitator channel, GlpF	6–259	254	7	28.6	38
1U19	rhodopsin	33–310	278	7	25.0	41
3KJ6	methylated β_2 adrenergic receptor	A: 35–346	311	7	39.5	31
Very large MPs						
3HD6	human Rh C glycoprotein, RhCG	6–448	403	12	43.6	59
3GIA	amino acid, polyamine, and organocation transporter, ApcT	3–435	433	12	62.5	64

3O0R	nitric oxide reductase subunit B	B: 10–458	449	12	30.6	69
3HFX	carnitine transporter, CalT	12–504	493	12	68.0	63
2XUT	peptide transporter, PepT1 and PepT2	A: 13–500	488	14	42.8	71
2XQ2	K294A mutant of Na^+/galactose transporter, VSGLT	A: 9–573	565	15	71.8	79

[a] Referred to as 1PY7 in this chapter; [b] Italicized PDB IDs indicate that this protein was used in RosettaTMH parameter optimization.

Modification of BCL::MP-Fold benchmark models for comparison with Rosetta

In order to compare the performance of BCL::MP-Fold with Rosetta, Rosetta loop definition files based on the models resulting from the BCL::MP-Fold benchmark for the 34 proteins in Table 19 were generated using the BCL (http://www.meilerlab.org/index.php/bclcommons/show/b_apps_id/1). Next, the model PDB files were converted to be compatible with the Rosetta cyclic coordinate descent (CCD) loop modeling application, according to the protocol outlined by Combs, *et al*. (252). The resulting PDB files were then used as input for fragment-based loop building in Rosetta. Only one output model was generated per input. That is, for 1,000 models that

were input for a given protein, only 1,000 models with loops were constructed. This procedure allowed for the calculation of $RMSD_{100}SSE$ (C_α $RMSD_{100}$ (347) in predicted native SSEs) over the same residues as that computed over Rosetta-built models.

Loop building on RosettaTMH-generated models

Because RosettaTMH makes cuts in the protein fold tree in order to perform rigid body sampling (see *The RosettaTMH* de novo *folding algorithm*), the TMHs needed to be reconnected with loops. Therefore, the models built using RosettaTMH with and without restraints were subjected to Rosetta fragment-based loop building, as described in the previous section.

Modification of Rosetta radius of gyration score for folding membrane proteins

In addition to implementing the ability to fold MPs with multiple experimental restraints, a modified version of the Rosetta radius of gyration (RG) scoring term was introduced to help keep the TMHs from drifting too far away in 3D space, as well as to prevent the TMHs from collapsing into the membrane. Generally, the RG of a protein is directly proportional to the extent to which it is "spread out" in Cartesian space (348). The existing RG scoring term in Rosetta is computed over all residues in the protein (100). For MPs, the new RG scoring term takes only the TMH centers of mass (CoMs) into account and is computed over only those residues' coordinates in the membrane, or X-Y, plane. In result, the scoring term, which is an energetic penalty, will disfavor conformational changes that cause the TMH CoMs to move, either laterally or along the membrane normal, far away from one another.

Weight optimization of Rosetta default radius of gyration score

Preliminary testing data indicated that the default weight for the Rosetta RG score was sub-optimal for folding MPs. Therefore, multiple weighting factors, ranging from 0.0 to 10.0 in increments of 0.25 (as well as 0.01), were tested. For each simulation, 1,000 models of the 9 MPs italicized in Table 19 were folded using the RosettaTMH and ExtendedChain protocols, as both protocols have not yet been optimized.

Determination of sampling efficiency for de novo *folding*

In order to measure sampling efficiency, or how many models need to be constructed for reliable benchmarking, 5,000 models based on the 1FX8, 1U19, and 3O0R primary sequences were folded with the MembraneAbinitio folding algorithm, with RosettaTMH with and without simulated EPR distance restraints ($weight_{KBP}$ = 20.0, $weight_{quadratic}$ = 1.0), and from an ExtendedChain with and without simulated EPR distance restraints ($weight_{KBP}$ = 50.0, $weight_{quadratic}$ = 20.0). After this was completed, the average $RMSD_{100}$SSE and standard deviation of a randomly selected subset of the 5,000 models were computed.

Simulation of EPR distance restraints using the BCL

For the benchmark in this chapter, 10 sets of EPR distance restraints were generated for each protein. This was done to avoid bias resulting from using any single restraint set. The restraint selection algorithm developed by Kazmier, *et al*. (314) was used employed. The algorithm optimizes the information content of the restraint set by maximizing the sequence separation between spin labeling sites. At the same time, the

algorithm finds restraint sets that link all pairs of SSEs in the protein. In order to convert the resulting restraint sets to EPR-like distance restraints for testing during *de novo* folding, the Euclidian distances between the specified residues were determined from the MP experimental structures. Next, a spin label uncertainty was added to each distance, based on the cone model-based spin label statistics generated for the RosettaEPR KBP (135). These statistics were generated by placing a pseudo-spin label in the form of a right-angle cone (based on methanethiosulfonate, or MTS) on exposed residue pairs in a database of over 3,500 proteins. The frequency of observed values for the calculated difference between spin label distance and C_β distance ($d_{SL}-d_{C\beta}$) were collected in a histogram, which was shown to match relatively well to experimentally determined $d_{SL}-d_{C\beta}$ values for T4-lysozyme and αA-crystallin. This histogram of spin label statistics quantifies the expected uncertainty associated with EPR distances measured on proteins spin labeled with MTS.

Optimization of EPR distance restraint scoring term weighting

The EPR distances for the residue pairs were simulated as described in the previous section. Preliminary benchmarking indicated that the EPR score used for the folding of T4-lysozyme (135) was insufficient to improve MP model quality of large MPs, such as rhodopsin. Instead, it was determined that a two-component scoring term was needed.

The modified EPR restraint potential for folding MPs consists of an energetic bonus derived from the aforementioned cone model statistics. Indeed, this energetic bonus is the same KBP used in the *de novo* folding of T4-lysozyme by Hirst, *et al*. (135).

However, in addition to the KBP energetic bonus, the EPR restraint score contains an energetic penalty characterized by the equation:

$$f(x)=\begin{cases} \left(\frac{x\text{-}lb}{sd}\right)^2 & \text{for } x < lb \\ 0 & \text{for } lb \le x \le ub \\ \left(\frac{x-ub}{sd}\right)^2 & \text{for } ub < x \le ub + rswitch * sd \\ \frac{1}{sd}\left(x-(ub+rswitch*sd)\right)+\left(\frac{rswitch*sd}{sd}\right)^2 & \text{for } x > ub + rswitch * sd \end{cases}$$

where x is the currently measured distance within the model, *lb* is the restraint lower bound, *ub* is the restraint upper bound, *sd* is the restraint standard deviation, and *rswitch* is set to 0.5. This quadratic penalty is similar to that used for nuclear Overhauser effect (NOE)-derived distance restraints in NMR structure calculations. The EPR scoring potential is designed such that the quadratic penalty is enforced if, during folding, the simulated model's d_{SL}–$d_{C\beta}$ value for a given residue pair is greater than –12.0Å and less than 12.0Å.

The weight of each EPR scoring term component was optimized separately. One thousand models of each protein were folded using RosettaTMH for each EPR restraint weighting scheme. For each protein under each of 49 weighting schemes, the percentage of models having $RMSD_{100}SSE$ < 8Å was computed, and the average of these values across the 9 proteins used for optimization are reported in Table 20. In addition, the enrichment was computed based on the models obtained from each weight scheme (Table 21). Enrichment was computed as:

$$enrichment = \frac{TP}{TP+FP} * \frac{P+N}{P},$$

where the *(P+N) / P* ratio = 10, limiting the maximum obtainable enrichment to 10.0. The models were sorted according to Rosetta score. Models that fell within the top 10% by score were counted as "positive," (*P*), and all models whose scores fell into the bottom 90% by score were counted as "negative" (*N*). The positives were then sorted by $RMSD_{100}SSE$ relative to the native structures, and those models that fell within the top 10% by $RMSD_{100}SSE$ were labeled "true positives" (*TP*s). All other low-scoring models were considered "false positives" (*FP*s).

During EPR restraint weight optimization, the default Rosetta RG score weighted at 4.25 was used (see *Weight optimization of Rosetta default radius of gyration score*). Further, each restraint is scored independently, and the sum of individual restraint scores constitutes the total raw restraint score. The total restraint score was multiplied by a normalization factor that is equal to:

$$weight_{cst} = \frac{\log(\#\,cst)}{\#\,cst} * \#\,aa\,,$$

where $weight_{cst}$ is the weight by which the entire restraint score is multiplied before it is added to the total Rosetta score, or energy, *#cst* is equal to the number of simulated EPR restraints used, and *#aa* is the number of residues in the protein. Because the total restraint score is the sum of individual restraint scores, the weighted restraint score can be represented by:

$$cst\_score_{weighted} = average(cst\_score_{raw}) * \log(\#\,cst) * \#\,aa\,.$$

The RosettaTMH de novo *folding algorithm*

The RosettaTMH MP folding algorithm differs significantly from both the Rosetta folding algorithm for soluble proteins, "ExtendedChain" (100), as well as the

published RosettaMembrane folding protocols (105, 117). It allows for enhanced sampling of MP topologies by treating TMHs as rigid bodies. Each helix can be rotated or translated, or transformed, as an independent entity. In order to implement this new algorithm in Rosetta folding, the model's fold tree was modified. The fold tree of a model is a directed acyclic graph--a data structure that represents the connectivity of the model in internal coordinate space. This connectivity is distinct from chemical connectivity and thus enables Rosetta to rapidly move large sections of the protein without disturbing other parts (138). In the case of a helical MP, a radial, or star, fold tree is used, in which the CoM of each helix is connected to a central node (Figure 32).

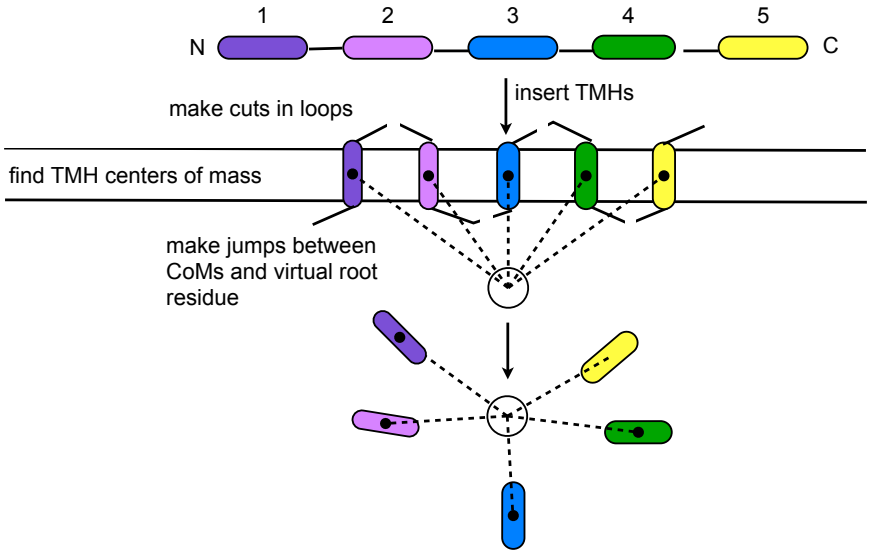

Figure 32: Generation of membrane protein fold tree in RosettaTMH
This schematic outlines how RosettaTMH generates a radial fold tree for a 5-helix membrane protein. In preparation for generating the fold tree, the primary sequence of the protein is read in and used to create an idealized α-helix. RosettaTMH utilizes user-defined TMH definitions to divide the idealized helix and insert each individual TMH into the implicit membrane. It then calculates each helix's center of mass (CoM). The CoMs connect the helices to a central root residue (open circle) in internal coordinate space.

Before *de novo* folding begins, each helix is inserted into the implicit RosettaMembrane environment (105). The CoM of each helix is set at the membrane center, and the helices are aligned along the membrane normal such that each helix is antiparallel to its sequential neighbors. The helices are arranged in a hexagonal grid and are initially separated from each other by 15Å. This grid point separation value was chosen after briefly testing distances of 5Å, 10Å, and 20Å and was selected based on the qualitative observation that, when TMHs were placed 5-10Å apart, they were more likely to "clash" into each other in sterically hindered conformations. On the other hand, a distance of 20Å caused the TMHs to never "see" each other during the first stage of *de novo* folding, making scoring by the Rosetta energy function difficult. The starting topology of the model is randomized; that is, the arrangement of helices in the hexagonal grid is different for each starting model (Figure 33).

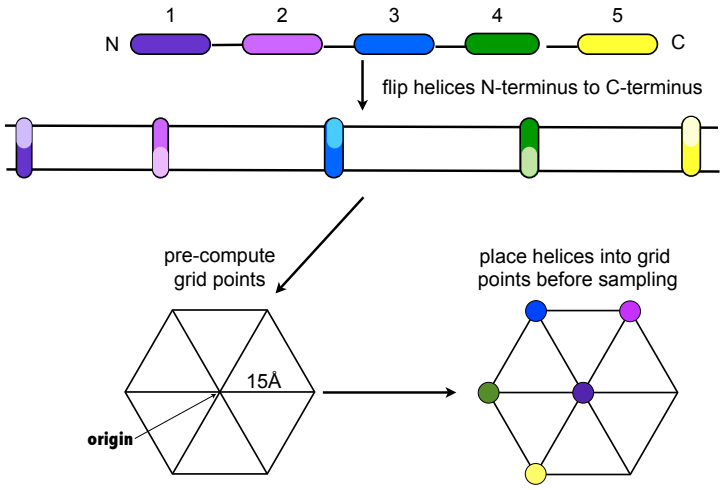

Figure 33: Initial placement of transmembrane helices before *de novo* folding
Schematic outlining initialization of protein model conformation before sampling. In this case, a 5-helix MP is inserted in the RosettaMembrane implicit membrane such that the TMHs run antiparallel to one another and aligned along the membrane normal. A hexagonal grid is computed, such that the vertices are aligned along the membrane center plane and are 15Å away from one another and from the origin. Then, for each grid point, a TMH is chosen randomly, and the helix is transformed to that grid point such that its CoM is aligned with the origin. The hexagonal grid can be expanded as needed, depending on the number of TMHs in the protein.

Stages of de novo *folding with RosettaTMH*

The pre-processing and *de novo* folding stages of RosettaTMH are summarized in Figure 34. Folding begins after the initialization of the model. The first stage of *de novo* folding consists entirely of rigid body transformations performed in a Monte Carlo Metropolis (MCM) fashion (266). For each MCM move, the helix is allowed to either rotate by up to 0.1° about any axis or translate up to 0.5Å in any direction from its current position. These values were selected based on preliminary testing and qualitative observation of the resulting models. The conformation resulting from each transformation is scored according to the RosettaMembrane centroid-based scoring function and, if specified, the MP-specific RG score (see *Modification of Rosetta radius of gyration score for folding MPs*). Stage 1 of folding consists of 2,000 MCM moves, and the RG and RosettaMembrane-specific "density" term and are turned on (105). These scoring terms aid in improving the compactness of the model. After the first stage, the model undergoes 9- and 3-amino acid fragment insertions using a protocol analogous to the one used for soluble proteins (100). Briefly, in Stage 2, 2,000 MCM cycles are performed, during which 9mer fragments are inserted onto the helical protein backbone. The density scoring term is turned off, and residue pairing, membrane environment, and membrane-specific penalties are added (105). The density term is re-introduced in Stage 3, which consists of 10 inner cycles; during these inner cycles, the scoring function can be alternated if desired. However, for MPs, the scoring function is the same for each of two inner cycle sub-stages. Each sub-stage consists of 2,000 MCM cycles for inserting 9mer fragments, resulting in a total of 20,000 fragment insertions. Finally, the density term is up-weighted

in Stage 4, and 4,000 MCM cycles of 3mer fragment insertions are performed (see Appendix E).

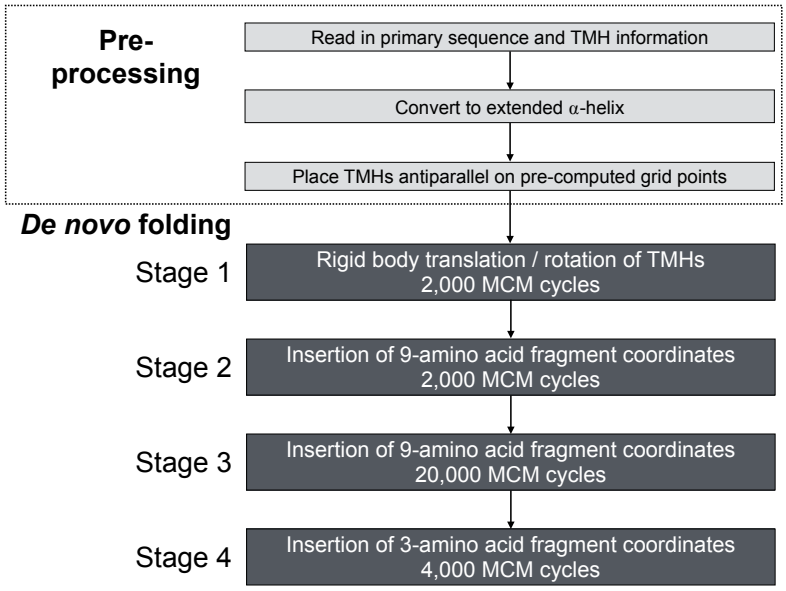

Figure 34: Outline of stages for RosettaTMH *de novo* folding

Benchmarking of RosettaTMH in the absence and presence of simulated EPR restraints

The generation of input files for this benchmark, except for the simulated restraints, is described in Weiner, *et al.* 's work on BCL::MP-Fold (345). Briefly, the primary sequence of each protein listed in Table 19 was used to generate 3- and 9-amino acid fragment files required for *de novo* folding in Rosetta. The Rosetta spanfiles containing the TMH definitions were obtained by using predictions from OCTOPUS (349). Rosetta lipophilicity files were also generated for each protein using the LIPS algorithm (350). One thousand models were folded from the primary sequence, using TMH information and the RosettaMembrane centroid-based scoring function (105). When multiple EPR restraint sets were used, the number of total models generated per restraint set was equal to the total number of models generated divided by the number of different restraint sets (i.e., 10 sets of 100 models for each protein). All computations

were performed on the Vanderbilt University Advanced Computing Cluster for Research and Education (ACCRE) using Rosetta revision numbers d592380 and d7b5a70 for RosettaTMH parameter optimization and benchmarking, respectively. The source code is available in the Rosetta3 master branch, which is available to developers in the RosettaCommons via https://github.com/RosettaCommons. The complete protocol capture for this work is described in Appendix E.

Results

Modified radius of gyration score does not significantly affect folding with RosettaTMH

Development of MP *de novo* folding methods often has a distinct advantage in that the membrane environment imposes a spatial constraint on the orientation of the protein. One way to leverage this constraint is to modify MP-specific scoring terms. In the case of RosettaTMH, an MP-specific RG scoring term ("new RG") was tested. This RG scoring term computes the RG over X- and Y- Cartesian coordinate values, disregarding the Z-coordinates that indicate vertical position of the MP in the membrane bilayer. The objective of this scoring term is to compress the model in the X/Y plane but not, or less drastically, along the Z-axis. This is different than the Rosetta default RG scoring term, which computes the RG over all three Cartesian dimensions. However, after testing this new RG score on the 9 MPs italicized in Table 19, it was found that the modified RG score did not affect overall performance for *de novo* folding with RosettaTMH (Figure 35). Therefore, the default Rosetta RG score was used for all further simulations in this chapter.

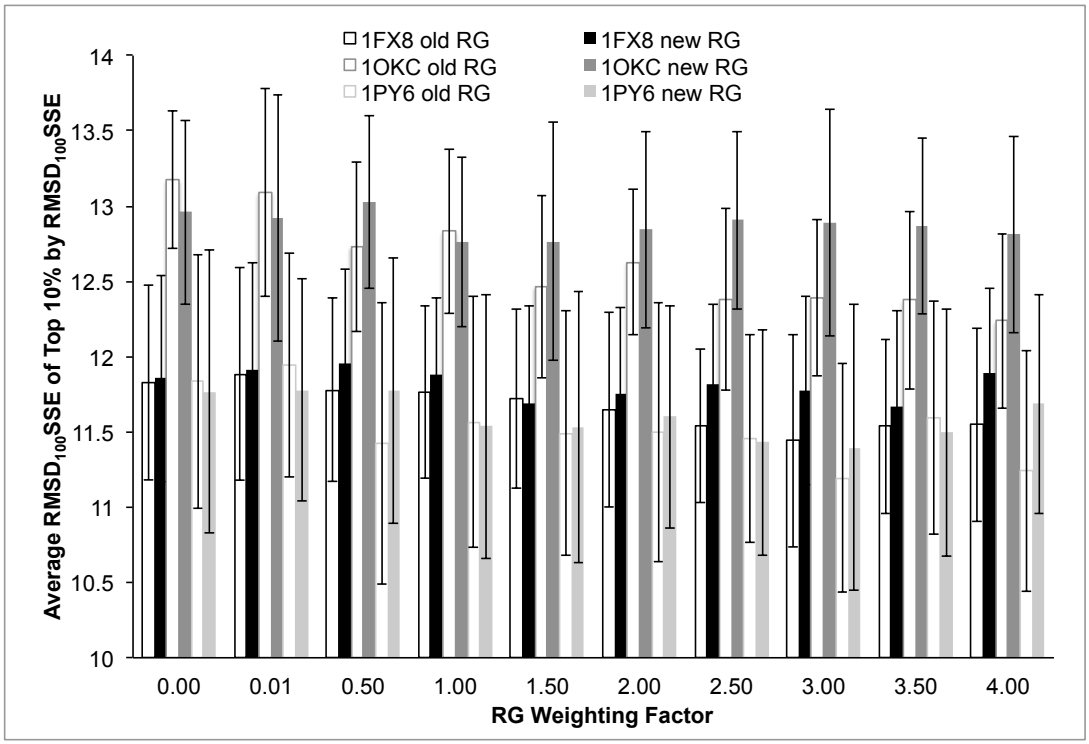

Figure 35: Comparison of RosettaTMH *de novo* folding with the original and modified radius of gyration scores
The mean $RMSD_{100}SSE$ (± S.E.M) of the top 10% of total models built for 9 MPs is plotted as a function of the factor by which the RG score is weighted (beyond the default Rosetta weight of 0.3).

Weight optimization for default radius of gyration score

Preliminary testing of *de novo* folding of MPs with RosettaTMH indicated that the weight of the Rosetta default RG score needed optimization. For each of 42 RG weighting factors ranging from 0.0 to 10.0, 1,000 models of the 9 proteins italicized in Table 1 were folded using RosettaTMH. The average enrichment and standard error of the mean (S.E.M.) were computed for each folding simulation (Figure 36). Based on these results, a weighting factor of 4.25 was used for the benchmarking of RosettaTMH.

It should be noted that, by default, the RG score is already multiplied by 0.3, thus resulting in an overall RG score weight of 1.275.

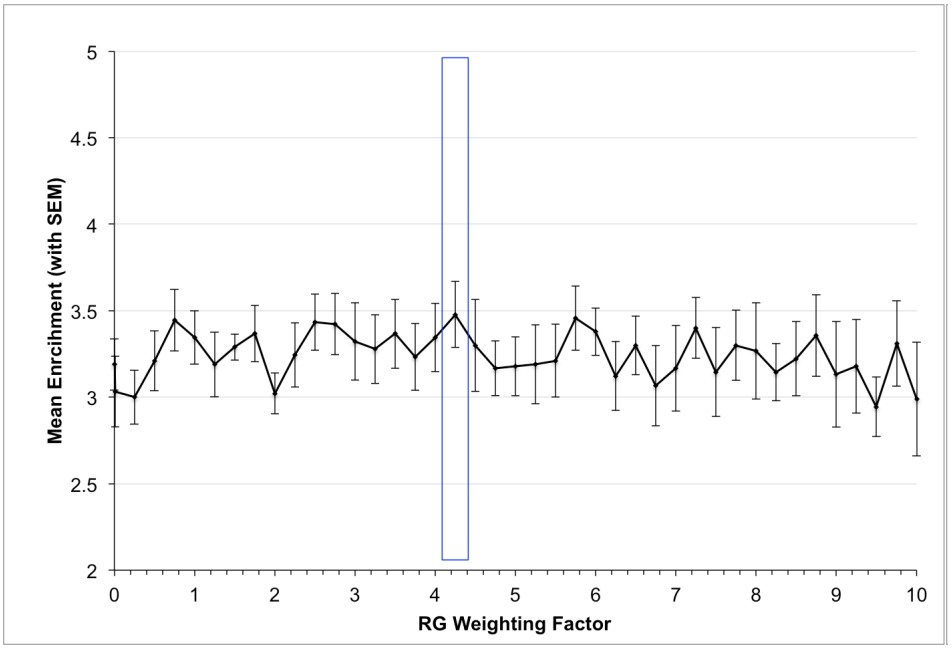

Figure 36: Optimization of the default Rosetta radius of gyration score weighting factor for *de novo* folding with RosettaTMH
The average enrichment (± S.E.M) obtained for folding 9 MPs with RosettaTMH is plotted as a function of RG weighting factor. The blue box indicates the optimum value for that folding method. Further benchmarking was performed with an RG weight of 4.25 for all folding protocols.

The sampling efficiency of Rosetta de novo *folding methods is similar*

To test how quickly *de novo* folding results of Rosetta for MPs converge, i.e., sampling efficiency, 5,000 models of 3 large MPs were folded using MembraneAbinitio, ExtendedChain, and RosettaTMH protocols. The latter two methods were also tested in the presence of simulated EPR restraints ("ExtendedChain+EPR" and "RosettaTMH+EPR," respectively). Then, randomized subsets of varying sizes were

selected, and the average $RMSD_{100}SSE$ and standard deviation was computed for that fraction of models. The standard deviation was then plotted as a function of fraction of total models built (Figure 37). Interestingly, all five folding methods' results appear to converge at around 3,750 – 5,000 models.

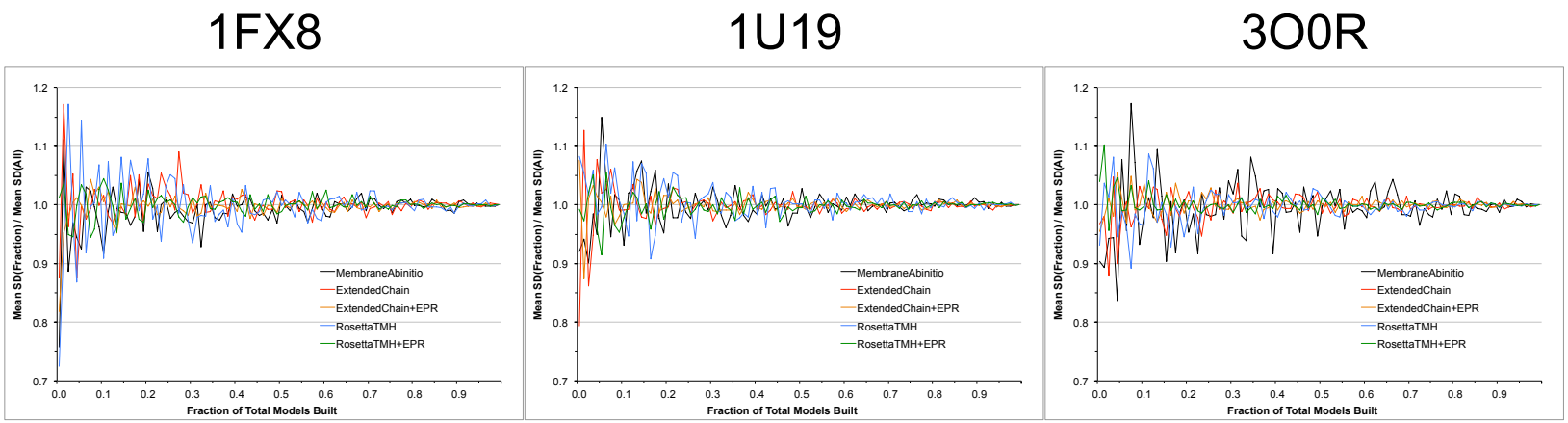

Figure 37: Sampling efficiency of various Rosetta *de novo* folding methods for three membrane proteins
Ratio between the mean standard deviation of $RMSD_{100}SSE$ (SD_{RMSD}) of a subset of 5,000 models and the mean SD_{RMSD} of all 5,000 models is plotted as a function of the fraction of total models built.

Optimal EPR restraint potential weighs both knowledge-based potential and quadratic penalty equally

De novo folding of soluble proteins with EPR restraints in Rosetta had been optimized previously (134, 135). However, it was found that, for MPs, a quadratic penalty was needed in addition to the EPR KBP energetic bonus to sufficiently improve conformational sampling of native-like folds. After rigorous weight optimization of this hybrid scoring term, it was determined that, for folding with RosettaTMH, the optimal weight scheme for the EPR distance restraint score was to multiply the EPR KBP by a factor of 1.0 and multiply the quadratic penalty by a factor of 1.0. This was based on the observation that, of the weights tested, the mean percentage of models with $RMSD_{100}$SSE < 8Å across 9 proteins was highest with this weighting scheme (Table 20). The enrichment for folding with this set of weights was 2.93 (Table 21). Interestingly, the enrichment for *de novo* folding with EPR restraints was generally lower than folding with no restraints. By definition, this is because the number of false positives, or low-scoring, high-RMSD models, was higher when folding with simulated restraints. This is perhaps due to the higher promiscuity of the EPR restraints, which are broader than distance restraints resulting from NMR NOEs. Therefore, models that fulfill the simulated restraints and are lower-scoring do not necessarily have native-like topologies.

Table 20: Percentage of correctly folded models obtained for folding nine membrane proteins with RosettaTMH using a variety of restraint score weighting schemes

		Percent of Models $RMSD_{100}SSE < 8Å$						
		Quadratic Penalty						
		0	1	10	20	30	40	50
EPR KBP	0	0.02	1.00	0.91	0.89	0.81	0.82	0.78
	1	1.56	4.03	3.74	3.26	3.48	3.41	3.51
	10	1.58	5.24	4.21	4.03	4.28	4.01	4.03
	20	1.54	4.53	4.03	3.84	3.96	3.67	3.64
	30	1.63	4.34	4.02	3.94	3.99	3.98	3.72
	40	1.30	4.28	3.99	3.81	3.82	3.60	3.57
	50	1.33	4.07	3.73	3.50	3.79	3.72	3.66

Table 21: Enrichment obtained for folding nine membrane proteins with RosettaTMH using a variety of restraint score weighting schemes

		Enrichment						
		Quadratic Penalty						
		0	1	10	20	30	40	50
EPR KBP	0	3.16	2.41	2.54	2.51	2.47	2.28	2.44
	1	3.40	2.93	2.58	2.66	2.71	2.64	2.54
	10	3.06	1.99	2.09	2.01	1.99	2.23	2.12
	20	2.87	1.76	2.00	1.70	1.93	1.78	1.89
	30	2.54	1.54	1.84	1.90	1.60	1.70	1.83
	40	3.12	1.67	1.69	1.78	1.82	1.90	1.73
	50	2.73	1.78	1.64	1.71	1.57	1.66	1.86

The data in Table 22 and Table 23 compares the enrichment and quality of models generated by MembraneAbinitio (105), ExtendedChain (100), ExtendedChain+EPR, RosettaTMH, and RosettaTMH+EPR. We also compare the various Rosetta MP folding methods to BCL::MP-Fold, another MP *de novo* folding method that forms 3D protein conformations via assembly of SSEs (345). There are no significant or apparent trends in the enrichment data, other than that enrichments for datasets generated with RosettaTMH are often higher for small- to medium-sized MPs. The addition of EPR restraints results

in a general decrease in enrichment for all but the largest MPs, which was unexpected (Table 22).

Table 22: Enrichment obtained for folding thirty-four membrane proteins with and without simulated EPR distance restraints*

PDB	Folding Method					
	MembraneAbinitio	ExtendedChain	+EPR	RosettaTMH	+EPR	BCL
3SYO	1.2	1.0	0.8 ± 0.9	3.2	2.1 ± 1.3	1.5
2BG9	1.5	0.7	2.5 ± 1.4	4.1	2.9 ± 1.0	2.2
1J4N	1.6	1.1	2.1 ± 1.5	0.0	3.0 ± 1.1	0.4
2KSF	1.0	0.9	1.0 ± 1.5	3.2	2.4 ± 2.0	1.5
1PY7	2.9	2.7	4.1 ± 2.1	4.0	3.1 ± 1.4	1.8
2PNO	1.0	2.3	1.7 ± 0.9	3.7	2.1 ± 1.2	0.8
2BL2	3.5	3.2	3.9 ± 1.7	4.3	2.6 ± 1.7	1.1
2K73	1.5	1.6	1.4 ± 1.2	3.1	2.3 ± 1.7	1.5
2ZW3	2.4	1.3	1.5 ± 0.7	3.1	1.8 ± 0.9	0.5
1IWG	1.9	1.0	2.4 ± 0.7	2.9	1.8 ± 1.5	0.6
1RHZ	1.9	1.7	2.2 ± 0.9	2.9	2.3 ± 0.9	0.9
2YVX	2.3	0.4	1.3 ± 0.8	2.5	2.0 ± 1.1	0.7
1OCC	2.5	2.5	1.8 ± 1.4	3.8	2.1 ± 1.1	1.1
4A2N	3.1	1.1	1.2 ± 1.1	3.5	2.7 ± 1.5	0.7
1KPL	1.7	2.3	1.3 ± 0.9	1.8	1.3 ± 1.2	0.6
2BS2	1.9	2.9	2.3 ± 1.1	3.6	2.0 ± 1.4	1.5
3P5N	2.8	2.6	3.0 ± 1.9	3.4	2.3 ± 1.2	1.5
2IC8	1.8	1.6	1.6 ± 1.5	2.3	2.0 ± 1.2	1.3
1PV6	1.4	2.5	2.8 ± 0.8	3.0	2.3 ± 1.6	1.2
2NR9	2.6	0.6	1.5 ± 1.0	2.6	2.0 ± 1.1	1.5
1OKC	3.4	4.0	2.2 ± 0.9	2.6	2.9 ± 1.3	0.7
3B60	2.0	3.5	4.1 ± 1.0	2.3	2.2 ± 1.1	0.0
2KSY	1.0	0.9	1.0 ± 1.5	3.2	2.4 ± 2.0	1.5
1PY6	1.3	3.1	3.5 ± 1.4	3.3	2.9 ± 1.0	1.4
3KCU	1.4	1.1	1.1 ± 0.9	2.2	1.4 ± 1.2	0.5
1FX8	2.5	4.9	2.4 ± 1.3	2.9	1.7 ± 0.9	0.9
1U19	5.2	3.3	1.9 ± 1.7	2.7	1.9 ± 1.1	1.4
3KJ6	1.2	0.6	0.8 ± 0.6	2.3	1.3 ± 0.9	1.0
3HD6	2.0	5.7	4.4 ± 0.5	1.9	1.8 ± 1.1	0.4
3GIA	2.7	4.2	2.3 ± 1.2	1.6	1.5 ± 1.1	0.0
3O0R	2.2	4.9	3.2 ± 1.2	1.7	2.3 ± 1.1	0.1
3HFX	3.5	1.9	2.1 ± 0.9	1.2	1.8 ± 1.4	0.4
2XUT	2.4	1.1	1.3 ± 1.2	1.6	2.1 ± 1.3	0.0
2XQ2	2.7	1.1	1.3 ± 1.4	0.8	1.1 ± 0.9	0.1
Mean	2.2	2.2	2.1 ± 1.2	2.7	2.1 ± 1.3	0.9
std. dev.	0.9	1.3	1.0 ± 0.4	1.0	0.5 ± 0.3	0.6

* Enrichment values with standard deviations were obtained from *de novo* folding 1,000 models with 10 different EPR distance restraint sets.

Addition of EPR restraints significantly improves sampling for RosettaTMH and ExtendedChain

In order to assess the overall sampling capability of each folding protocol, the average $RMSD_{100}SSE$ of the top 10% of models by $RMSD_{100}SSE$ ($\mu_{10\%}RMSD$) was computed relative to the experimental, or native, structure. Additionally, we computed the percentage of models having an $RMSD_{100}SSE < 8Å$, which serves as a cutoff for determining if models have the correct topology. We also report the best $RMSD_{100}SSE$ ($Best_{RMSD}$) obtained for each method and the mean $RMSD_{100}SSE$ of the five lowest-scoring models ($\mu_{5model}score$). As was observed with T4-lysozyme (135), the addition of EPR restraints increases the likelihood of obtaining the correct MP fold for both RosettaTMH and ExtendedChain. When looking at the percentage of models with $RMSD_{100}SSE < 8Å$, for 12 of 34 proteins, RosettaTMH performs worse than ExtendedChain, while RosettaTMH+EPR performs better than ExtendedChain+EPR. Several of these proteins consist of over 200 residues, indicating RosettaTMH's ability to fold large MPs in the presence of restraints. Further, when compared to other Rosetta MP folding methods, RosettaTMH+EPR obtains the highest percentage of correctly folded models for 4 of the 13 medium-sized proteins, 5 of the 8 large proteins, and 5 of the 6 very large proteins. However, BCL::MP-Fold out-performs all Rosetta methods for the vast majority of benchmark cases – a fact that requires further research outside the scope of the present work (Table 23).

Table 23: Overall performance of *de novo* folding membrane proteins with Rosetta and BCL::MP-Fold

PDB	Metric	Folding Method					
		MembraneAbinitio	ExtendedChain	+EPR	RosettaTMH	+EPR	BCL
3SYO	μ_{5model}score[a]	11.6	8.0	7.7	6.3	5.2	4.0
	$Best_{RMSD}$[b]	6.9	2.5	2.0	2.9	2.3	1.6
	$\mu_{10\%}$RMSD[c]	9.5	4.9	3.5	5.1	3.4	3.0
	% < 8Å[d]	0.4	43.5	74.6	30.4	71.3	100.0
2BG9	μ_{5model}score	8.7	10.4	7.2	8.3	7.3	6.4
	$Best_{RMSD}$	4.2	4.1	3.9	5.5	2.9	2.7
	$\mu_{10\%}$RMSD	5.8	5.3	5.6	8.4	5.1	3.6
	% < 8Å	28.1	22.1	40.0	3.0	32.7	50.7
1J4N	μ_{5model}score	10.0	8.2	7.6	11.2	10.0	9.2
	$Best_{RMSD}$	4.9	4.8	3.1	6.6	5.7	4.6
	$\mu_{10\%}$RMSD	7.0	6.1	4.5	8.9	7.9	6.0
	% < 8Å	12.3	28.8	44.7	1.4	4.4	32.3
2KSF	μ_{5model}score	8.5	9.2	10.0	8.8	9.6	6.6
	$Best_{RMSD}$	5.6	4.8	4.0	5.5	3.9	3.2
	$\mu_{10\%}$RMSD	6.9	6.4	5.8	9.1	5.9	4.2
	% < 8Å	18.8	27.0	28.8	1.3	20.6	41.3
1PY6*	μ_{5model}score	4.5	7.3	2.3	9.6	6.6	8.0
	$Best_{RMSD}$	2.5	1.9	1.9	6.0	2.9	3.8
	$\mu_{10\%}$RMSD	3.9	3.5	2.6	9.4	5.0	4.8
	% < 8Å	63.7	70.7	57.9	0.9	31.0	57.4
2PNO	μ_{5model}score	7.9	8.1	7.9	10.4	7.8	59.6
	$Best_{RMSD}$	4.1	3.3	2.8	7.2	4.4	4.5
	$\mu_{10\%}$RMSD	6.0	5.7	4.1	10.1	6.0	7.2
	% < 8Å	29.0	26.0	56.6	0.5	23.3	13.6
2BL2	μ_{5model}score	6.7	3.9	5.4	9.9	7.3	38.7
	$Best_{RMSD}$	2.3	2.3	2.5	6.7	3.4	3.0

	$\mu_{10\%}$RMSD	3.6	3.4	3.7	9.8	5.0	4.0
	% < 8Å	70.1	54.8	70.9	1.0	53.8	79.9
2K73	μ_{5model}score	10.1	6.9	5.7	10.6	5.9	31.4
	$Best_{RMSD}$	6.4	3.1	2.8	6.3	3.1	2.8
	$\mu_{10\%}$RMSD	8.7	4.5	4.2	9.0	4.9	3.8
	% < 8Å	1.1	43.5	55.2	1.6	48.3	72.3
2ZW3	μ_{5model}score	11.8	12.1	9.7	10.5	7.7	48.3
	$Best_{RMSD}$	10.1	5.2	5.2	5.7	3.7	3.1
	$\mu_{10\%}$RMSD	11.9	8.3	6.8	8.8	5.4	4.5
	% < 8Å	0.0	2.7	16.4	1.8	30.2	73.2
1IWG	μ_{5model}score	8.1	10.4	8.4	11.5	7.2	8.6
	$Best_{RMSD}$	5.8	5.0	4.8	7.6	4.2	4.2
	$\mu_{10\%}$RMSD	7.3	11.2	5.8	10.0	5.9	5.9
	% < 8Å	12.2	9.0	42.6	0.3	27.0	41.9
1RHZ	μ_{5model}score	9.8	9.4	8.2	11.9	7.7	9.4
	$Best_{RMSD}$	7.1	5.2	3.9	7.5	5.4	4.9
	$\mu_{10\%}$RMSD	8.8	7.1	5.2	10.1	7.0	7.2
	% < 8Å	0.7	11.9	44.4	0.2	12.9	12.6
2YVX	μ_{5model}score	8.9	14.3	8.1	11.3	6.6	9.4
	$Best_{RMSD}$	6.7	6.0	4.1	7.5	3.8	5.8
	$\mu_{10\%}$RMSD	7.9	8.2	5.6	10.4	6.6	7.3
	% < 8Å	4.8	3.5	35.8	0.1	17.6	13.3
1OCC	μ_{5model}score	9.8	10.1	9.1	9.1	6.9	7.4
	$Best_{RMSD}$	5.9	7.0	5.9	7.8	4.1	4.5
	$\mu_{10\%}$RMSD	9.1	8.8	7.9	10.0	5.4	6.2
	% < 8Å	0.8	1.2	5.0	0.3	45.9	49.0
4A2N	μ_{5model}score	9.9	14.0	8.7	9.6	8.0	8.7
	$Best_{RMSD}$	6.4	5.6	3.8	6.4	4.5	3.8
	$\mu_{10\%}$RMSD	8.2	7.8	5.5	9.3	6.7	5.6

	% < 8Å	3.4	5.0	32.2	1.3	10.9	29.2
1KPL	μ_{5model}score	13.3	13.8	13.8	13.9	11.5	145.4
	$Best_{RMSD}$	10.3	9.9	6.9	11.2	7.3	9.9
	$\mu_{10\%}$RMSD	13.0	12.5	9.0	13.0	9.8	11.3
	% < 8Å	0.0	0.0	1.0	0.0	0.1	0.0
2BS2	μ_{5model}score	9.9	9.0	8.1	10.6	9.1	8.1
	$Best_{RMSD}$	6.0	6.3	5.1	6.4	4.1	4.9
	$\mu_{10\%}$RMSD	8.8	9.0	6.7	10.2	6.1	6.4
	% < 8Å	1.5	1.0	15.1	0.1	26.9	31.4
3P5N	μ_{5model}score	9.0	9.4	7.3	12.0	9.1	114.9
	$Best_{RMSD}$	5.5	5.1	3.8	7.5	4.6	4.5
	$\mu_{10\%}$RMSD	8.3	7.4	5.4	10.0	7.0	6.5
	% < 8Å	2.0	7.1	36.6	0.1	16.3	23.6
2IC8	μ_{5model}score	10.1	9.3	9.7	10.1	8.3	9.6
	$Best_{RMSD}$	5.2	4.9	3.5	7.8	5.0	5.1
	$\mu_{10\%}$RMSD	7.9	7.3	5.6	10.2	6.8	6.8
	% < 8Å	3.6	8.0	30.0	0.1	13.0	16.7
1PV6	μ_{5model}score	10.5	9.9	8.0	11.7	7.8	9.3
	$Best_{RMSD}$	5.7	6.9	4.2	7.5	4.2	5.7
	$\mu_{10\%}$RMSD	8.2	8.6	6.3	10.7	5.9	7.3
	% < 8Å	3.5	1.5	20.6	0.1	23.8	10.5
2NR9	μ_{5model}score	9.6	9.6	9.0	11.3	9.1	8.6
	$Best_{RMSD}$	6.1	5.6	3.7	8.3	5.3	5.1
	$\mu_{10\%}$RMSD	8.3	7.8	5.6	10.2	7.1	6.8
	% < 8Å	2.5	4.5	27.8	0.0	10.9	16.7
1OKC	μ_{5model}score	13.1	10.5	10.6	12.0	9.2	78.9
	$Best_{RMSD}$	9.0	8.4	6.9	8.0	5.1	5.7
	$\mu_{10\%}$RMSD	11.6	12.1	8.6	10.7	7.3	7.8
	% < 8Å	0.0	0.0	2.1	0.1	9.5	5.9

3B60	μ_{5model}score	9.4	9.4	4.3	11.0	8.6	74.4
	$Best_{RMSD}$	5.5	6.0	3.2	7.8	4.8	9.2
	$\mu_{10\%}$RMSD	8.3	8.3	4.7	10.2	6.3	11.1
	% < 8Å	2.2	3.2	48.0	0.1	30.7	0.0
2KSY	μ_{5model}score	7.7	10.0	8.4	11.7	6.6	8.2
	$Best_{RMSD}$	4.3	3.6	3.9	8.2	4.2	5.1
	$\mu_{10\%}$RMSD	6.0	6.2	5.3	10.9	5.8	6.7
	% < 8Å	27.9	17.8	33.4	0.0	28.6	17.7
1PY6	μ_{5model}score	8.4	7.1	5.8	12.3	7.7	8.4
	$Best_{RMSD}$	4.6	4.0	4.2	8.3	4.1	4.7
	$\mu_{10\%}$RMSD	6.3	6.6	5.9	10.6	5.8	6.2
	% < 8Å	29.3	16.5	23.9	0.0	32.7	24.4
3KCU	μ_{5model}score	10.0	11.7	10.9	12.2	11.1	100.7
	$Best_{RMSD}$	6.8	6.8	4.7	8.8	5.9	6.5
	$\mu_{10\%}$RMSD	8.9	9.2	7.6	11.1	7.6	8.5
	% < 8Å	0.4	0.7	6.1	0.0	6.6	1.0
1FX8	μ_{5model}score	11.3	11.5	11.6	11.5	9.3	210.4
	$Best_{RMSD}$	7.7	7.0	6.2	9.6	6.7	7.2
	$\mu_{10\%}$RMSD	10.1	11.2	8.8	10.9	8.8	8.3
	% < 8Å	0.1	0.1	1.2	0.0	0.7	1.9
1U19	μ_{5model}score	11.9	15.8	10.8	12.1	8.3	8.5
	$Best_{RMSD}$	9.7	12.7	7.3	8.3	4.9	5.3
	$\mu_{10\%}$RMSD	12.5	15.0	9.2	10.7	6.6	7.0
	% < 8Å	0.0	0.0	0.5	0.0	16.8	14.2
3KJ6	μ_{5model}score	15.1	13.0	10.1	10.8	8.6	152.6
	$Best_{RMSD}$	12.4	7.5	5.1	6.2	3.6	4.4
	$\mu_{10\%}$RMSD	13.7	10.3	7.2	9.6	6.4	5.9
	% < 8Å	0.0	0.1	10.8	0.6	20.0	43.9
3HD6	μ_{5model}score	10.6	16.5	11.1	19.2	10.8	10.2

	$Best_{RMSD}$	6.8	13.1	8.3	9.7	7.2	7.0
	$\mu_{10\%}$RMSD	9.6	19.4	10.9	12.2	9.4	8.5
	% < 8Å	0.1	0.0	0.1	0.0	0.6	1.6
3GIA	μ_{5model}score	14.0	25.9	13.5	15.0	14.0	122.3
	$Best_{RMSD}$	11.6	20.4	8.0	11.2	8.8	9.1
	$\mu_{10\%}$RMSD	14.0	24.4	10.2	13.2	10.5	10.8
	% < 8Å	0.0	0.0	0.2	0.0	0.0	0.0
3O0R	μ_{5model}score	10.2	24.1	9.9	14.5	9.5	51.3
	$Best_{RMSD}$	6.5	20.2	6.0	10.3	6.2	6.9
	$\mu_{10\%}$RMSD	8.6	24.6	9.2	12.3	7.6	8.7
	% < 8Å	2.2	0.0	1.3	0.0	6.8	1.6
3HFX	μ_{5model}score	13.2	24.7	12.2	15.1	9.9	77.8
	$Best_{RMSD}$	10.0	15.2	7.0	8.4	5.7	6.0
	$\mu_{10\%}$RMSD	12.8	20.5	9.1	10.1	7.9	8.0
	% < 8Å	0.0	0.0	0.8	0.0	4.4	4.2
2XUT	μ_{5model}score	14.5	48.2	12.8	16.1	9.5	75.1
	$Best_{RMSD}$	12.5	22.6	8.3	12.1	6.5	7.4
	$\mu_{10\%}$RMSD	13.6	29.4	11.4	13.6	8.5	9.4
	% < 8Å	0.0	0.0	0.1	0.0	2.7	0.3
2XQ2	μ_{5model}score	17.2	52.3	13.1	17.5	11.5	103.7
	$Best_{RMSD}$	13.8	31.7	9.8	12.1	6.7	8.2
	$\mu_{10\%}$RMSD	15.6	39.5	11.5	13.6	8.6	10.3
	% < 8Å	0.0	0.0	0.1	0.0	1.8	0.0
Mean ± std. dev.	μ_{5model}score	10.4 ± 2.6	13.9 ± 10.6	9.0 ± 2.6	11.8 ± 2.6	8.6 ± 1.8	48.1 ± 46.0
	$Best_{RMSD}$	7.0 ± 2.8	8.2 ± 6.8	4.9 ± 2.0	7.9 ± 2.0	4.9 ± 1.4	5.3 ± 2.0
	$\mu_{10\%}$RMSD	9.1 ± 2.9	11.2 ± 8.2	6.7 ± 2.4	10.4 ± 1.7	6.8 ± 1.5	6.9 ± 2.2
	% < 8Å	9.4 ± 17.6	12.1 ± 18.0	25.4 ± 22.5	1.3 ± 5.3	20.1 ± 16.6	26.0 ± 26.8

[a] μ_{5model}score = mean $RMSD_{100}$SSE to native structure of the top five models by score; [b] $Best_{RMSD}$ = $RMSD_{100}$SSE of the best model by $RMSD_{100}$SSE compared to the native structure; [c] $\mu_{10\%}$RMSD = mean of the top 10% of models by $RMSD_{100}$SSE compared to the native structure; [d] % < 8Å = percentage of total models folded having an $RMSD_{100}$SSE < 8Å

De novo *folding with RosettaTMH improves sampling over other methods for large proteins*

For ease of visualization, a representative set of 7 proteins was chosen from the 34-protein benchmark set for further $RMSD_{100}SSE$ analysis. For each protein and for each folding method, the $RMSD_{100}SSE$ values were sorted from lowest to highest, and the top 5% of models by $RMSD_{100}SSE$ were selected. Next, $RMSD_{100}SSE$ vs. $RMSD_{100}SSE$ plots comparing RosettaTMH and RosettaTMH+EPR with the other Rosetta and BCL MP folding methods were generated. This analysis clarifies a few key conclusions concerning RosettaTMH. First, RosettaTMH+EPR samples lower-RMSD conformations for larger MPs when compared to RosettaTMH, MembraneAbinito, and ExtendedChain. Also, when comparing RosettaTMH with MembraneAbinito and ExtendedChain, the latter two methods are more suitable for structure prediction when the proteins are small- to medium-sized. Finally, RosettaTMH performance is comparable to MembraneAbinitio and ExtendedChain for 2K73 and 1FX8, respectively (Figure 38).

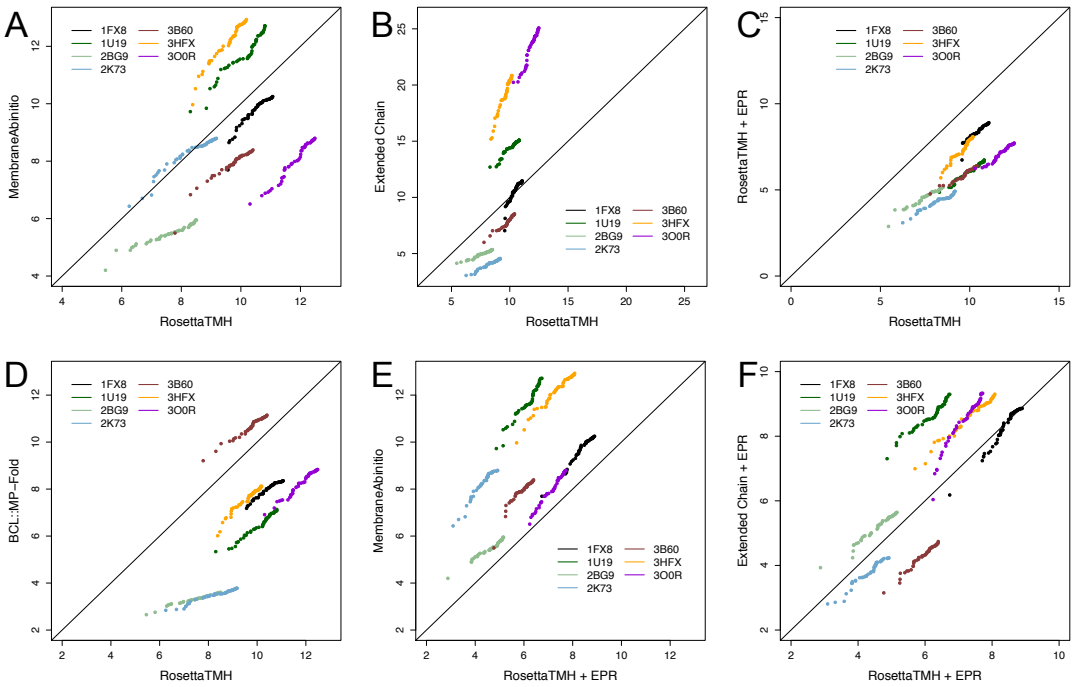

Figure 38: Sampling performance for *de novo* folding with RosettaTMH compared to other folding methods
For each panel, the $RMSD_{100}$SSE of the top 5% of models by score were selected for 7 proteins. A) MembraneAvinitio vs. RosettaTMH. B) Folding from an extended chain vs. RosettaTMH. C) RosettaTMH with EPR restraints vs. RosettaTMH (without EPR restraints). D) BCL::MP-Fold vs. RosettaTMH. E) MembraneAbinitio vs. RosettaTMH with EPR restraints. F) Folding from an extended chain with EPR restraints vs. RosettaTMH with EPR restraints.

Addition of EPR restraints primarily responsible for improvement seen in RosettaTMH folding

The MembraneAbinitio folding algorithm was first benchmarked on a dataset of relatively small proteins and performed best with small helical bundles (105). However, it was found that MPs having more complex topologies posed a much more difficult challenge, which RosettaTMH could possibly address. Indeed, 6 of the 34 proteins tested in this benchmark set show improved quality when using RosettaTMH over folding with MembraneAbinitio or ExtendedChain. Further, the addition of the EPR distance restraint

potential improves sampling of native-like folds significantly. This appears to be primarily due to the influence of the EPR restraints, as folding with ExtendedChain+EPR also increases sampling efficiency to include the correct fold. In order to test this hypothesis, rhodopsin (PDB ID: 1U19 (260)) was selected for an in-depth analysis of the relationship between the number of EPR restraints and overall conformational sampling ability.

Rhodopsin was folded using all of the Rosetta methods listed in Table 22. However, for RosettaTMH+EPR and ExtendedChain+EPR, multiple sets of models were generated based on whether 0, 10, 20, 40, 80, or 160 simulated EPR distance restraints were used. Unlike in the 34-protein benchmark, only one EPR restraint set for each scenario was generated, and 1,000 models were folded for each case. Box-and-whisker plots of the resulting $RMSD_{100}$SSE distributions are displayed in Figure 39. When no restraints are used, MembraneAbinitio and folding from an extended chain perform similarly, while RosettaTMH generally appears to generate lower-RMSD models. When using 10 restraints, RosettaTMH+EPR and ExtendedChain+EPR exhibit similar median $RMSD_{100}$SSE values, but RosettaTMH+EPR samples a wider range of conformations. However, when 20 or more restraints are used, RosettaTMH+EPR is consistently better in sampling the correct fold. As expected, the number of outliers correlates inversely with the number of restraints (Figure 39).

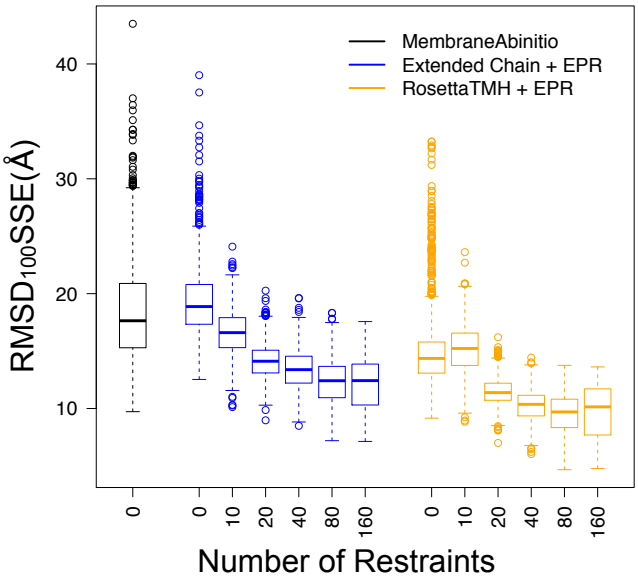

Figure 39: Sampling performance of RosettaTMH with various EPR restraint set sizes for folding rhodopsin
Box-and-whisker plot indicating the breadth of model accuracy obtained for folding rhodopsin with RosettaTMH with 10, 20, 40, 80 and 160 simulated EPR distance restraints. The thick line indicates the median $RMSD_{100}SSE$ obtained, while the boxes indicate the interquartile range. The highest and lowest $RMSD_{100}SSE$ values, excluding outliers, are indicated by the "whiskers," and outliers are shown as open circles.

Detailed analysis of individual de novo *folding stages indicate rigid body sampling not necessary*

In addition to studying the overall performance of RosettaTMH with and without EPR restraints, the ability of the protocol to sample MP topologies during each stage of folding (see Figure 34) was also analyzed. As with the above experiment, rhodopsin was chosen as an example protein, and, when indicated that EPR restraints were employed, only one set of 41 optimally weighted restraints was used. For each folding method, 1,000 individual trajectories were run, and the conformations before folding began and after each stage of folding were output. Then, similar to in Figure 39, the $RMSD_{100}SSE$ distributions for each folding stage were plotted. The single, high-scoring conformations

observed at folding initiation, or Stage 0, are all an extended chain, which is how both the default Rosetta folding algorithm, and MembraneAbinitio, begin. Accordingly, for MembraneAbinitio and ExtendedChain model quality significantly improves from initiation to Stage 1 and then from Stage 1 to Stage 2. In contrast, RosettaTMH-generated model accuracy decreases during Stage 1 of folding. That is, the rigid body sampling causes the quality of rhodopsin models to decrease. The $RMSD_{100}SSE$ values do not improve significantly for Stages 2-4 when no restraints are used. When EPR restraints are used, the models' accuracy improves from Stage 1 to Stage 2 but does not change significantly thereafter. This was also observed for ExtendedChain+EPR (Figure 40).

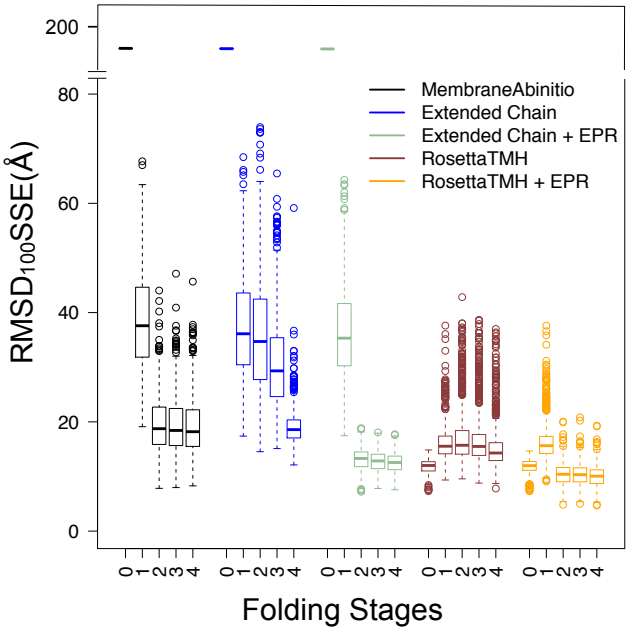

Figure 40: Sampling performance of various Rosetta methods during each stage of *de novo* folding using rhodopsin as an example
Box-and-whisker plot indicating the breadth of model accuracy obtained during each stage of folding with MembraneAbinitio, folding from an extended chain with and without EPR restraints, and folding with RosettaTMH with and without restraints. The thick line indicates the median $RMSD_{100}SSE$ obtained, while the boxes indicate the interquartile range. The highest and lowest $RMSD_{100}SSE$ values, excluding outliers, are indicated by the "whiskers," and outliers are shown as open circles.

RosettaTMH-generated models exhibit large inter-helical distances

RosettaTMH assembles MP folds by breaking up the proteins into individual TMHs and allowing these helices to move as rigid bodies. Therefore, the resulting arrangements could feature distances between subsequent SSEs that cannot be connected by a loop. In order to determine the extent to which this is true, a representative set of 7 proteins chosen from the 34-protein benchmark was selected, and the Euclidian distance between subsequent SSEs was measured for 1,000 models generated with RosettaTMH in the presence and absence of restraints. The $\log_{10}$ of the Euclidian distance was plotted as a function of $\log_{10}$ of the number of amino acids in the loop (Figure 41). The distance-loop length relationship for the 7 native proteins, as well as the maximum Euclidian distance possible ($3.8*(loop\_length-1)$), were also determined and plotted. According to the information in Figure 41, it appears that, especially for shorter loops, RosettaTMH often places TMHs too far away in 3D space. Similarly, RosettaTMH fails to reflect the dependence of Euclidean distance from loop length accurately. Indeed, for all proteins, excluding 3HFX, the inter-helix distances of the vast majority of models generated can theoretically be spanned by a loop, but only a small percentage—if any--exhibits native-like inter-helix distances (Table 24).

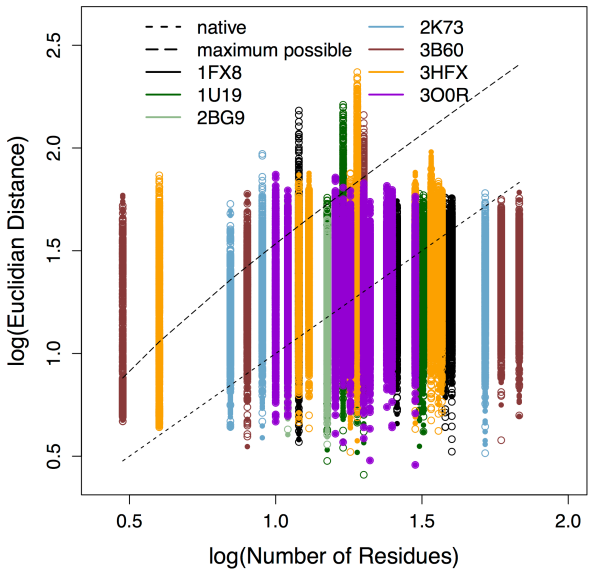

Figure 41: Analysis of inter-SSE distances for RosettaTMH-folded models
The $\log_{10}$ of the inter-SSE Euclidian distance (i.e., loop distance) as a function of the $\log_{10}$ of the number of residues in the loop is plotted for a representative set of membrane proteins. The long-dashed line indicates the maximum Euclidian distance possible for *n* residues, and the short-dashed line indicates the Euclidian distances for *n* residues found in the proteins' native structures.

Table 24: Percentage of models having loops that can or are likely to be closeable

PDB	% Models with Loop Distance < Maximum Possible		% Models with Loop Distance < Native	
	+ EPR Restraints	− EPR Restraints	+ EPR Restraints	− EPR Restraints
1FX8	91.1	76.4	0.2	0.0
1U19	99.8	80.7	0.0	0.0
2BG9	99.3	96.7	11.2	2.4
2K73	97.7	76.0	4.4	0.2
3B60	79.2	62.5	0.0	0.0
3HFX	42.6	26.7	0.0	0.0
3O0R	76.9	41.9	0.0	0.0

Discussion

EPR restraints significantly assist in obtain models with the correct topology

The results in Table 23, Figure 38, Figure 39, and Figure 40 indicate that, for large and very large MPs, the conformational search space of MP structures must be

limited in order to obtain *de novo*-folded models with native-like folds. The MembraneAbinitio protocol attempts to accomplish this by folding MPs "from the inside out." That is, a helix in the middle of the protein sequence is inserted into the implicit membrane environment first. Next, either helices N- or C-terminal to the initially inserted helix are folded into the membrane via fragment-based assembly, beginning with the helix adjacent to the starting helix. Then, the helices on the other side (in terms of sequence) are folded in the same manner (105).

While MembraneAbinitio is able to generate models with $RMSD_{100}SSE < 8Å$ for over half of the 34 proteins tested, the majority of these success cases have fewer than 200 residues and 7 TMHs. Indeed, for 12 proteins, the MembraneAbinitio protocol performs better than RosettaTMH and folding from an extended chain when EPR restraints are not used. However, when EPR restraints were used, the additional restraints result in more models having the correct fold (Table 23). This is important because MembraneAbinitio, unlike RosettaTMH, cannot take EPR restraints into account. Therefore, for MPs of more than 4 TMHs and 145 residues, it is advantageous to include structural restraints, such as those available from NMR, EPR, etc. If one does employ such restraints, the traditional folding method, ExtendedChain, appears to be better suited for medium-sized MPs, whereas RosettaTMH may be the best method for *de novo* folding larger MPs, such as GPCRs, channels, and transporters.

Optimization of RosettaTMH folding protocol may lead to further improvement

Even though Rosetta is now capable of folding MPs that have the correct fold and is sometimes able to recover intra-helical features, these models are not yet accurate enough to be used as input for full-atom refinement using the RosettaMembrane all-atom

scoring functions (130). Typically, models of approximately 2.0Å $RMSD_{100}$SSE relative to the native structure are required in order to successfully obtain atomic detail information (134, 135).

Based on the information in Figure 40, one obvious next step in protocol optimization would be to forego the rigid body sampling in Stage 1 of RosettaTMH folding. It is expected that the initial set of rigid body transformations results in less viable MP conformations (e.g., helix out of the membrane, lying too orthogonal to the membrane normal, or too far apart in 3D space). The fragment insertions in Stages 2-4 are then not able to recover the correct fold. This is supported by the lowest-RMSD models displayed in Figure 42, which show that there is a general lack of inter-helical packing and native-like placement that is not remedied by fragment insertions. Not surprisingly, the addition of EPR restraints assists in improving packing and even in the recovery of helical features (Figure 42).

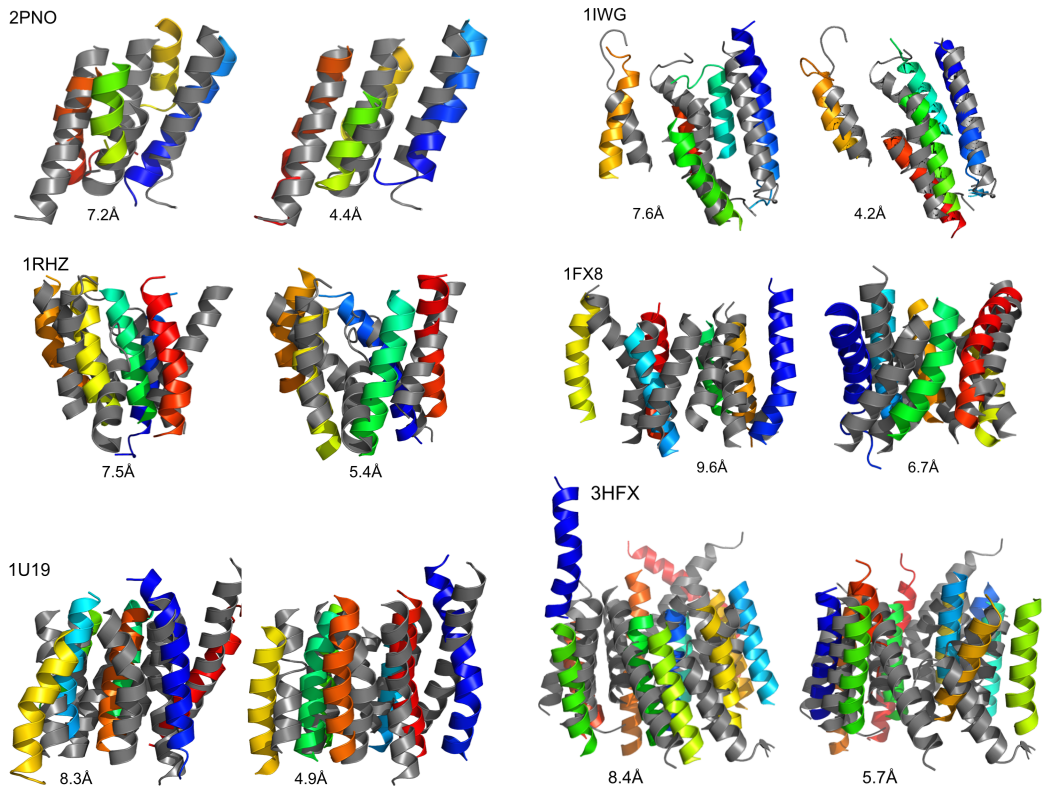

Figure 42: Most accurate model resulting from RosettaTMH folding for six proteins
The most accurate models obtained from folding with RosettaTMH without EPR restraints (left model) and with EPR restraints (right model) are colored in rainbow. The native structures are colored in gray. The $RMSD_{100}$SSE of the model compared to native is reported in angstroms.

Implementation of loop closure filter and knowledge-based potential for de novo *folding with RosettaTMH could improve inter-helix packing*

In order to create a radial fold tree for each model, the original simple fold tree must be "cut" to maintain the data structure's acyclic nature. For folding with RosettaTMH, these cutpoints are chosen within the MP loops (Figure 32). However, now that the TMHs can move independently from one another, another external force must be applied to keep the helices in relatively close proximity, as the helices appear to drift apart and not exhibit native-like packing (Figure 41 and Figure 42). One possible means

of doing this is to implement and optimize a loophash filter, which would ensure that helices that would normally be connected by a loop remain close enough in Cartesian space such that the inter-helical loop can be successfully rebuilt at a later stage.

The loophash filter is based on work recently published by Tyka, Jung, and Baker (351). In the protocol introduced by the authors, the loophash algorithm allows for extremely fast rebuilding of protein segments by rapidly determining if a loop of a given sequence length can span the distance defined by two endpoints. A hash lookup table is generated for a loop of a given sequence length, and the hashes in the table refer to specific protein segments found in a database of non-homologous proteins of known structure. In addition to the loophash, or loop closure, filter, the implementation of a loop distance KBP, such as that used by BCL::Fold (352, 353) could also be useful. While the loop closure filter would assist in ruling out models where helices could not theoretically be connected, and the loop distance KBP would provide an energetic incentive to place TMHs in more native-like conformations.

Increased sampling may be needed in order to better observe RosettaTMH's performance

Even though the RosettaTMH folding protocol remains under development, it appears to be a much more rapid means of folding MPs than MembraneAbinitio and fragment-based assembly alone (
Figure 43). This is probably a result of the lack of fragment insertions, and thus recalculation of torsion angles, during the first stage of folding. However, this decreased amount of fragment insertion may be the cause of the generation of lower-quality models. In any case, the significant speedup in model production allows for the generation of

many more models. This increased sampling speed will likely prove beneficial for obtaining higher quality models from RosettaTMH for large MPs.

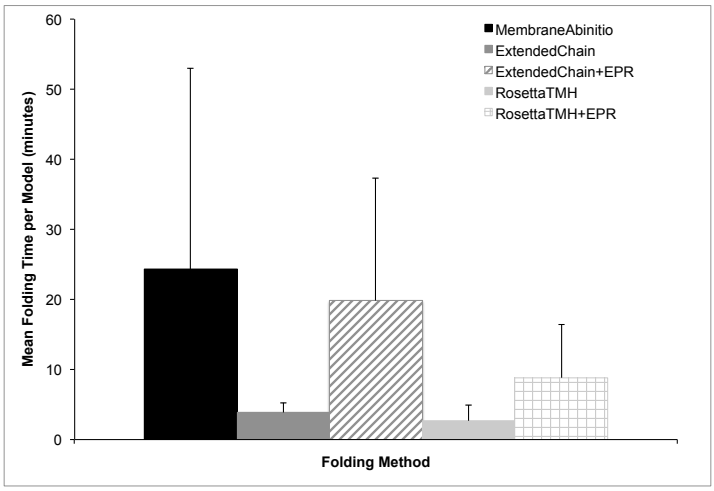

Figure 43: Average time required for *de novo* folding
The mean time in minutes (± std. dev.) required to *de novo* fold 1,000 models for 34 MPs with different Rosetta folding methods. When the use of EPR restraints is indicated (+EPR), the EPR KBP bonus weight = 1.0, and the quadratic penalty weight = 1.0.

Conclusion

RosettaTMH is a novel *de novo* folding protocol that assembles MP topologies from the rigid body movements of TMHs, followed by peptide fragment insertions. This approach, along with the significantly decreased time required to fold models, allows for increased sampling of conformational space. In addition, complicated topologies can be sampled more efficiently, which is important for the structure prediction of more complex proteins, such as GPCRs, transporters, and channels. Finally, RosettaTMH, unlike MembraneAbinitio, allows for the folding of MPs with experimental restraints. Further, while the new folding protocol alone improves sampling, the addition of experimental restraints may be necessary to obtain native-like topologies, which is especially important for determination of proteins for which there is no structure.

Availability

The RosettaTMH source code is available in the Rosetta3 master branch, which is available to developers in the RosettaCommons via https://github.com/RosettaCommons. Protocol captures used for generating the data in this paper are available in Appendix E.

Acknowledgements

The authors would like to thank Drs. Frank DiMaio, Steven Lewis, and other members of the RosettaCommons for their assistance in the development of RosettaTMH. Axel Fischer was also very helpful in providing protocols on simulating EPR distance restraints in the BCL, and Dr. Brian Weiner provided many of the Rosetta-ready input files for the benchmark set. This work was funded by NIH grant F31-GM100742, awarded to Stephanie DeLuca. Work in the Meiler laboratory is further supported through NIH (R01 GM080403, R01 MH090192, R01 GM099842, R01 DK097376) and NSF (CHE 1305874).

CHAPTER VI

CONCLUSION

Summary of this work

In an effort to understand the structural basis of disease, as well as aid in the development of novel drugs and therapeutics, structural biologists have worked tirelessly to determine the three-dimensional (3D) conformations and dynamics of biomacromolecules. This is evidenced by the explosion of structures deposited in the PDB (1, 317). However, the structural characterization of membrane proteins, referred to as MPs in this dissertation, has proven to be especially challenging. It is therefore not surprising that atomic-detail information about this important class of proteins is limited compared to that for soluble proteins (110, 354, 355). While a number of technological and methodological advances have made MP structure determination more feasible than ever before, the vast majority of MPs, including members of the ever-prominent GPCR superfamily, continue to elude both experimental and computational structural biologists.

The primary purpose of this body of work was to implement, test, and apply methods that combine the power of both experiment and computation. This hybrid approach has allowed, and will allow, for the elucidation of protein structures, including transporters, peptide GPCRs, and their ligands. The concept of computational modeling in the presence of experimental restraints was applied, via the Rosetta molecular modeling suite, to the structural determination of the prolactin releasing peptide, or PrRP (Chapter II) (37, 135), and ghrelin (Chapter III) (Vortmeier, DeLuca, *et al*., submitted), both of which are endogenous agonists of GPCRs. In these cases, the experimental

restraints were derived from NMR spectroscopic data. However, with the introduction of RosettaEPR (Chapter IV) (107, 108, 135, 251), it was shown that even the relatively broad distance information from EPR spectroscopy can be used to enhance the conformational search space of the Rosetta protein folding algorithm. Finally, RosettaEPR was combined with a novel MP *de novo* folding algorithm, RosettaTMH, which further improves sampling of native-like topologies for large MPs (Chapter V) (S.H. DeLuca, S.D. DeLuca, A. Leaver-Fay, and Meiler, submitted). Other versions of RosettaMembrane folding protocols have been reported previously (105, 115, 117, 345), but neither method could be combined with EPR or NMR experimental data to the extent that is enabled by RosettaTMH.

Improving our understanding of the structural basis of PrRP receptor activation by PrRP

The prolactin releasing peptide, or PrRP, is a peptide hormone that activates the prolactin releasing peptide receptor, which is a GPCR that is primarily located in the pituitary (28, 105). PrRP, and to some extent, its receptor, are implicated in regulating body weight homeostasis, metabolism, and energy expenditure (29-31, 33, 135). Even though two groups had reported structural ensembles of PrRP based on NMR data (1, 35, 36, 356), no coordinates of the models were made publicly available, hindering further exploration of this biologically significant interaction.

In order to overcome the obstacle of the lack of structural data, RosettaNMR (106, 167, 168, 354, 355) was used to *de novo* fold PrRP using NMR chemical shifts (CSs) and NOE distances reported in the literature (35, 37, 357). More specifically, the 13 C-terminal residues of PrRP (PrRP8-20) were folded using 3- and 9-amino acid

fragments that were generated using the 13 available CSs and the inter-proton distances resulting from backbone hydrogen NOEs. These fragments were assembled according to the Rosetta protein folding algorithm (100, 102, 135, 358, 359), during which conformations agreeing with all 38 backbone and side-chain NOE distances were energetically favored. The models were assessed according to the Rosetta soluble protein energy function during both folding and full-atom refinement (100, 105, 110, 117, 360-364).

After filtering for conformations with both low global energies and maximal satisfaction of the NMR distance restraints, an ensemble of 20 models having a backbone root mean square deviation (RMSD) of 0.83Å was obtained. As was expected based on the work by D'Ursi, *et al*. and Danho, *et al*. (28, 35, 36, 365), the final ensemble revealed an amphipathic helical structure, in which all three arginine side-chains were arranged on one side of the helix. Furthermore, Rosetta appeared to sample both α- and 3_{10}-helical conformations, indicating that the peptide exists in a dynamic equilibrium between the two. This hypothesis was supported by CD spectroscopic data collected in SDS and TFE, which were provided by the Beck-Sickinger laboratory at Leipzig University. Importantly, the final ensemble of PrRP models was deposited in the Protein Model Database (29-31, 33, 180, 366) (PM ID: 0078404). The models were also used in further computational modeling guided by structural restraints derived from receptor activation data. The peptide ensemble was docked into a comparative model of the PrRP receptor, and the combination of modeling and experimental data allowed for the elucidation of a dual-binding mode of PrRP to the PrRP receptor (19, 367-370) (Appendix A).

In addition to the determination and deposition of the 3D structure, PrRP and various analogs thereof were tested for their ability to activate the wildtype and selected point mutants of the PrRP receptor. Specifically, [R^{15}A]PrRP20, [R^{19}A]PrRP20, [F^{20}A]PrRP20, as well as PrRP4-20, PrRP8-20, PrRP14-20, and full length PrRP20 were tested and their secondary structure analyzed by CD spectroscopy in water, 100 mM SDS micelles, and 25% TFE solution. While the PrRP point mutants' EC_{50} values were markedly decreased compared to wildtype, their CD spectra were practically identical in all three solvent conditions. PrRP4-20 and PrRP8-20 also activated the wildtype receptor to a similar extend as PrRP20, but the helicity of the two shorter truncation mutants was reduced.

Further, when signal transduction assays of [$Y^{2.64}$A]PrRPR and [$W^{2.71}$A]PrRPR were performed with PrRP20 and PrRP8-20, it was found that PrRP8-20 exhibited significantly decreased agonism on the receptor mutants compared to full-length PrRP20 (197-fold and 963-fold higher EC_{50} over PrRP20 for [$Y^{2.64}$A]PrRPR and [$W^{2.71}$A]PrRPR, respectively). This is in stark contrast to PrRP20's EC_{50} values, which were 23-fold and 270-fold higher for [$Y^{2.64}$A]PrRPR and [$W^{2.71}$A]PrRPR, respectively. These results ndicate that the reduced helical character of PrRP8-20, while not affecting the truncated peptide's ability to activate the wildtype receptor, severely impaired activation for certain receptor mutants. The authors of the publication proposed that the two conserved receptor residues, $Y^{2.64}$ and $W^{2.71}$, assist PrRP in forming a more α-helical conformation upon binding to the receptor, which then allows for its activation. Even though PrRP8-20 exhibits a decreased propensity to form an α-helix, it could also activate the wildtype receptor. However, when $Y^{2.64}$ and $W^{2.71}$ were no longer available to help PrRP8-20

adopt a helical binding conformation, the truncated peptide could not stimulate the receptor (37, 371) (Chapter II).

Elucidation of ghrelin structure, dynamics, and interaction with the membrane

Ghrelin is also a peptide hormone; it is synthesized in the gut and activates the GHSR1a, or ghrelin receptor, which is primarily located in the hypothalamus (40-42, 372-376). Ghrelin, which is 28 residues in length, has an origenic effect, meaning that its presence is associated with increased food intake, but it appears to be involved in other physiological processes, such as memory, energy homeostasis, and reward mechanisms in the brain (45-49). A 3D structure of ghrelin and/or its receptor would therefore be desirable for the purposes of drug discovery, but unfortunately, but one has not been made publicly available. Therefore, in collaboration with the Huster laboratory at Leipzig University, the structural ensemble of the peptide was determined using Rosetta. In addition to providing CSs of ghrelin bound to lipid vesicles, our collaborators also performed further ssNMR studies of ghrelin that yielded information on peptide dynamics and its interaction with the membrane bilayer.

Because ghrelin is acylated at its third residue, Ser^3, and because the ssNMR CSs were collected for acylated ghrelin bound to vesicles, it was important for the modeling to take place in the Rosetta MP-specific energy function and implicit membrane environment (105, 130). In order to perform *de novo* folding in RosettaMembrane, a comparative model of the ghrelin receptor was created based on the rhodopsin crystal structure (PDB ID: 1U19 (260)). The receptor model was not of interest in this study and was only present for logistical purposes. However, it may be useful for future studies.

Ghrelin was folded using fragment-based assembly without the use of the experimental CSs. The resulting models were filtered such that Ser^3 was located within the RosettaMembrane "polar" layer of the implicit membrane, which would be expected for ghrelin's octanoyl chain help anchor it to the membrane. Next, PROSHIFT (253), SPARTA+ (261), SHIFTX (256), and SHIFTX2 (258) were used to predict CSs based on all Rosetta-generated models. For ensemble selection, the RMSD between the experimental CSs and the predicted CSs for a randomly chosen ensemble of models was computed, and the ensemble was altered in a Monte Carlo fashion until the RMSD was minimized. This process was conducted for predicted CSs resulting from all four tools mentioned above starting from models within the top 10%, 25%, 50%, 75%, and 100% of all *de novo* folded models by Rosetta energy (that contained proper placement of Ser^3). After careful analysis, it was determined that, for this peptide and dataset, PROSHIFT was the most effective CS prediction method, and the set of models selected from the Monte Carlo algorithm that was generated using the top 10% of models by Rosetta score (and passed previous filters) was chosen as the final representative conformational ensemble. This ensemble's RMSD of predicted CSs relative to the experimental CSs was 0.4 ppm. The conformations in the final ensemble exhibited a core, flanked by more flexible N- and C-terminal tails. They also exhibited strong polyproline II (PPII) helical propensity for residues 21-23 and 26-27, but this was not predicted by TALOS+ (143), which does not account for PPII helix.

Interestingly, this final ensemble was highly flexible, having a backbone RMSD of 4.0Å relative to the mean structure, according to the PSVS analysis tool (http://psvs–1_5-dev.nesg.org). This is in agreement with the NMR order parameters, which indicate

that the C-terminal region of the peptide is mobile. However, experiments measuring ghrelin binding to lipid membranes provide evidence that the octanoyl group on Ser^3 is not enough to account for the peptide/membrane interaction. Rather, because ghrelin is positively charged at the pH used for NMR studies (pH = 6), it is hypothesized that the basic residues on ghrelin also contribute. Further, spin diffusion studies showed that ghrelin binds to, but does not insert deep into, membranes via residues Ser^3 and Phe^4. The remainder of the peptide is therefore expected to be highly mobile and flexible, but not necessarily "random coil." This work has been submitted to *PLoS ONE*.

Development of RosettaEPR, a computational tool that integrates EPR distance information for the structure determination of proteins in atomic detail

Because EPR spectroscopy cannot currently yield atomic-detail information of a protein, as is possible with X-ray crystallography or NMR spectroscopy, computational methods that combine EPR data with modeling for protein structure determination would be desirable. In 2008, Alexander, *et al*. presented the first attempt at *de novo* folding soluble proteins in Rosetta using sparse EPR data (134). EPR distance data were simulated using a pseudo-spin label “motion-on-a-cone” model, in which the C_β of the MTS-based spin label was attached to a simple ellipsoid. The cone model, allowed for the relating of EPR spin label distances (d_{SL}) to protein C_β distances ($d_{C\beta}$), but distribution of simulated d_{SL}–$d_{C\beta}$ values did not align well with experimentally determined values for T4-lysozyme and αA-crystallin. The authors were nevertheless able to generate atomic-dtail models of the two soluble proteins.

Hirst, *et al*. built on Alexander and colleagues' work and introduced RosettaEPR in 2011 (135) (Chapter IV). An EPR distance knowledge-based potential (KBP) was generated, in which statistically common d_{SL} values for a given $d_{C\beta}$ were correlated with the energy of a Rosetta-folded model. During fragment assembly, the model's d_{SL}–$d_{C\beta}$ value for a residue pair of interest (i.e., experimental distance restraint provided for that residue pair) was measured, and the corresponding energy according to the KBP was added to the model's total restraint score. This was, in turn, added to the model's overall Rosetta energy.

As a proof of concept, RosettaEPR was tested on the α-helical core domain of T4-lysozyme, which consists of 107 residues. Twenty-five experimentally determined EPR distances were provided by the Mchaourab laboratory at Vanderbilt University. These data were used as structural restraints during *in silico* folding. In addition, RosettaEPR's performance was compared to folding with no restraints and to folding with "bounded" distance restraints. The upper and lower bounds of the bounded restraints were defined such that $(d_{SL} - \sigma_{SL} - 12.5\text{Å}) \leq d_{C\beta} \leq (d_{SL} + \sigma_{SL} + 2.5\text{Å})$, where σ_{SL} is the experimental error associated with each measurement. For $d_{C\beta}$ values that did not fall within the allowed range, a quadratic energy penalty similar to that used for NOE distance violations in NMR was applied to the model's restraint score. After weight optimization of the RosettaEPR KBP, the potential weighted by a factor of 4.0 was more able to recover correctly folded models ($RMSD_{C\alpha} < 7.5$Å relative to the crystal structure (PDB ID: 2LZM (317)) and models with native-like conformations ($RMSD_{C\alpha} < 3.5$Å) than when folding with bounded restraints or no restraints. Further, the correlation between Rosetta energy and model accuracy increased from 0.42 to 0.51 and then to 0.62 for

folding with no restraints, with bounded restraints, and with RosettaEPR, respectively. This supports the hypothesis that the incorporation of an energy term specifically designed for EPR distance data improved Rosetta *de novo* folding for a small, soluble, helical protein.

After establishing the RosettaEPR *de novo* folding protocol for soluble proteins, 500,000 models of T4-lysozyme were generated in the presence of the 25 aforementioned EPR distance restraints. Of the resulting models, those that fell within the top 1% by total Rosetta energy that satisfied at least 85% of the optimal EPR restraint score were chosen for full-atom refinement using Rosetta's all-atom scoring function for soluble proteins. This scoring function includes KBPs for van der Waals interactions, hydrogen bonding, and solvation (110). Of the 500,000 models folded, fewer than 1,400 were selected for refinement. During this step, the amino acid side-chains are added, as opposed to treating them as "superatoms." Next, the protein undergoes eight cycles of side-chain repacking and energy minimization, in which backbone torsion angles are sampled. The EPR restraints were not used during refinement because it was anticipated that the protein backbone would not change enough to be captured by the broad EPR KBP.

Ten all-atom models were generated for each *de novo* folded input model, resulting in over 13,000 models. The top 10% of these models according to Rosetta energy were then carried forth to the next iteration of refinement. This process was repeated until a total of eight iterations were complete. After the final cycle, the lowest-energy model had an $RMSD_{C\alpha}$ of 1.76Å relative to the crystal structure and a side-chain rotamer recover of approximately 60% over the core residues. The high accuracy of this low-scoring model is remarkable because it was produced by folding directly from the

primary sequence with the assistance of a dataset of approximately one EPR distance restraint per four residues. This is evidence that RosettaEPR can be a useful tool for protein structure determination when sparse EPR distance information is available (135).

Membrane protein structure determination made possible via the combination of a novel de novo *folding algorithm and EPR distance information*

RosettaEPR appeared to be an effective method for obtaining atomic-detail models of small, soluble proteins, with demonstrated ability on the helical domain of T4-lysozyme. However, ultimate goal of the development of RosettaEPR is to be able to apply it to the structure determination of large MPs, such as GPCRs, channels, and transporters. After thorough and rigorous testing of the currently available MP folding protocols in Rosetta, both with and without simulated EPR restraints, it was apparent that a more efficient sampling of conformational space was necessary. The originally reported RosettaMembrane folding algorithm, as well as a newer method that allowed for folding larger MPs, were more sophisticated than the default soluble protein folding protocol in Rosetta. Unfortunately, they could not be combined with experimental restraints for scoring during fragment assembly. Therefore, RosettaTMH was developed to address the lack of MP-specific Rosetta folding methods amenable to being combined with experimental data (Chapter V).

Because the Rosetta folding infrastructure has been highly optimized for fragment-based assembly, RosettaTMH is a novel approach to folding proteins in this software suite. Firstly, the RosettaTMH algorithm was implemented within the Rosetta Topology Broker framework, which was initially developed for folding of proteins with

sparse NMR data (107, 108, 251) and allows for the creation of innovative and flexible *de novo* folding methods. RosettaTMH folds helical MPs by treating individual, idealized α-helices, which are defined by the user, as rigid bodies during the first stage of folding. This increases the number and types of protein topologies sampled. Subsequent stages of folding take place via the more traditional Rosetta method of peptide fragment insertion, which allows for the potential recovery of intra-helical features, such as bends and kinks. These structural features are under-sampled by BCL::Fold and BCL::MP-Fold (115, 345). Further, RosettaTMH can be paired with experimental information. To demonstrate this, a set of 34 MPs were chosen for *de novo* folding with RosettaTMH, and EPR distance restraints were simulated using the BCL (http://bclcommons.vueinnovations.com). RosettaTMH's performance, both with and without simulated EPR restraints, was compared to that of the Rosetta MembraneAbinitio folding protocol (105), folding from an extended chain, as is done with soluble proteins, and folding from an extended chain with EPR restraints. The Rosetta methods were also compared to models previously generated by Weiner, *et al*., for benchmarking of BCL::MP-Fold (345).

The weight of the EPR restraint score was optimized on 9-protein subset of the 34-protein benchmark set. Weight optimization was performed in a similar manner to that described by Hirst, *et al*. for folding of T4-lysozyme (135). However, the process was more complex in that, in addition to the EPR KBP, a quadratic energetic penalty was used to further discourage conformations in which the $d_{SL}-d_{C\beta}$ values fell outside of the energetically favored region outlined by the RosettaEPR KBP. Therefore, each of these components for scoring restraint agreement was weighted individually. The optimal

weight combination was chosen based on the overall performance across all nine proteins. That is, for each protein, the percentage of models having an $RMSD_{100}SSE$ < 8Å was computed, and the mean of all 9 resulting values was reported. This value was computed for 49 EPR score weighting schemes, and the weighting scheme with the highest mean $value_{percent<8}$ was considered optimal.

Generally speaking, if a model's $RMSD_{100}SSE$ is less than 8Å relative to the native structure, it is considered to have the correct topology. Of the 34 benchmark MPs, RosettaTMH yielded more correctly folded models in 8 cases compared to MembraneAbinitio and in 9 cases compared to folding from an extended chain. However, upon the addition of EPR restraints, both RosettaTMH and folding from an extended chain improved topology recovery in most cases, indicating that the addition of experimental information is the driving force for achieving the correct fold. This is the case, at least, for this set of proteins and for these simulated restraints.

RosettaTMH with EPR restraints generally performed better than the other methods for proteins having more than 200 residues. Further, RosettaTMH folds proteins much more rapidly than the other Rosetta folding methods, which allows for increased sampling due to the decreased computational time required per model. In summary, RosettaTMH enables folding MPs with experimental restraints and may also be the preferred folding method in Rosetta when no restraints are available if the protein's topology is more complex than a typical helical bundle. This work is reported in detail in Chapter V has been submitted to the Rosetta special issue of *PLoS ONE*.

Implication of results

The work presented herein serves as an example of how an interdisciplinary approach to biomedical investigation can enhance the scientific community's effectiveness when it comes to research and method development. For example, in the case of the work presented on ghrelin and PrRP, the collaboration between computationalists and experimentalists led to the structural elucidation of two biologically significant peptide hormones. This information can now be used for future studies designed to explore the mechanism(s) of GPCR activation and, hopefully, the molecular basis of disease. In addition, the development of RosettaEPR, a novel protein folding method that incorporates EPR data to improve conformational sampling, would have not been possible without the collaboration of the Mchaourab laboratory, the personnel of which provided the experimental expertise to make the project a success. RosettaEPR is now available for the determination of soluble protein (and soon for MP) conformations that agree with experimental data.

By developing computational methods that take experimental information into account for the structural characterization of proteins and peptides, we can model inter-molecular interactions. These models then serve as the starting points for hypothesis generation. Experimental data resulting from those hypotheses can then be used to guide refinement of the models. This iterative process enables scientists to study biologically significant and interesting systems more in depth and with a structural biological perspective.

Towards the development of anti-obesity therapeutics

The structural characterization of two peptide hormones, PrRP and ghrelin, was possible due to the close collaboration of the Meiler laboratory with the Beck-Sickinger laboratory (PrRP) and the Huster laboratory (ghrelin). The interdisciplinary approach taken to study these two peptides has led to insight into not only their structures, but also their mechanism of binding and activation to their respective receptors. This is especially important given the ever-increasing rates of obesity, type II diabetes, and other metabolic problems, especially in the United States. Specifically, by providing the conformational ensembles of PrRP and ghrelin for other researchers to use, we may be able to better probe the means by which the peptides interact with their GPCRs and perhaps develop drugs and therapeutics that target these receptors.

The anorexigenic effect that injected PrRP has on rodents (167), as well as the observation that PrRP receptor-knockout mice incur obese phenotypes after sixteen weeks (357), points to the PrRP receptor as a possible anti-obesity drug target. Between January 2001 and March 2004, the U.S. Patent Office issued one patent, in which the inventors developed pharmacological approaches to study and develop drugs that activate (or inhibit) PrRPR (358, 359), indicating that this receptor is an attractive drug target. Publishing and providing the structure of PrRP, as well as further characterizing the peptide's ability to stimulate the PrRP receptor, enables further informing of future pharmacological studies and drug discovery.

Because ghrelin, which acts as an agonist of the ghrelin receptor, is an orexigenic hormone, compounds and peptides that are antagonists of GHSR1a would be one possible means of treating obesity. Indeed, while there are several known agonists of

GHSR1a (developed for stimulating growth hormone release), the design and optimization of ghrelin receptor antagonists is underway (360-364). (See Chollet, *et al*. for an in-depth review of agonists, antagonists, and inverse agonists of the ghrelin receptor (365).) Interestingly, ghrelin receptor agonist and inverse agonist radiotracers for positron emission tomography (PET) were developed in order to track receptor ligands i*n vivo* (366). By being able to structurally describe how ghrelin activates its receptor, it may be possible to develop, not only future drugs, but also ligands that serve as tracers for imaging studies.

Newly established de novo *folding protocols open doors for structural characterization of non-globular polypeptides*

In addition to small molecules (150-500 Da), the design of peptide ligands and peptidomimetics is playing an increasingly prominent role in the discovery of new drugs and therapeutics, especially when the drug target is a GPCR that is naturally activated by a peptide (367-370). However, unlike small-molecule ligand docking methods, there are relatively few tools for predicting the binding mode of peptides and peptidomimetics *in silico*. This is further complicated by the tendency of peptides to adopt multiple--often rapidly changing--conformations in solution. Some methods attempt to address with mild success (371). Rosetta, though optimized for *de novo* folding proteins, also allows for predicting the structures of peptides in the presence of NMR CSs and NOEs. Therefore, protocols for folding peptides in solution and in the implicit environment were developed and used to generate conformational ensembles of PrRP and ghrelin.

In this work, ghrelin posed an interesting problem because it is a membrane-bound peptide, and only sparse CSs from ssNMR were available. However, there now exists a method for *de novo* folding flexible peptides that are consistent with NMR data, which can be used for soluble or membrane-associated biomolecules. This could be especially useful for the generation of models of unstructured peptides and proteins, which serve a variety of biological functions, such as transcription regulation, translation, and cell signaling (372-376). Because intrinsically disordered proteins, or IDPs, are difficult to characterize by X-ray crystallography and NMR alone, computational structure prediction that incorporates NMR data into the modeling may be a suitable means of generating 3D conformational ensembles. This information can then be used to help provide insight into information obtained via "low-resolution" methods, such as CD spectroscopy.

RosettaEPR can be used to determine three-dimensional protein structures using sparse EPR distance datasets

The challenges associated with protein structure determination by NMR, such as molecular weight limitations on the system under study (including micelles, bicelles, etc. for MPs), can sometimes be discouraging. At the same time, crystallization remains a perpetual challenge. On the other hand, EPR spectroscopy offers several advantages. No crystallization is required; further, due to its sensitivity, only pico-moles of the protein are required, and there are no size constraints. Importantly, the protein can be studied in its native environment.

EPR is not without its own challenges, however. It cannot yield atomic-detail 3D structures, as with NMR spectroscopy and X-ray crystallography. It is also an inherently low-throughput method, in which, for each EPR measurement, a cysteine-less mutant must be made, cysteines introduced at sites of interest, constructs tested for functionality, and the protein spin labeled. Aside from the sparseness of the data resulting from the method's low-throughput nature, the paramagnetic spin label often introduces ambiguity into the experimental measurements. For example, one commonly used spin label, methanethiosulfonate (MTS) spin label, is highly flexible, having five rotatable bonds, and is relatively long--about 8.5Å from the C_β to the end of the molecule. This information must be taken into account when analyzing distance data from EPR.

RosettaEPR was developed in order to serve both experimentalists and computationalists. For the EPR spectroscopist, RosettaEPR allows for the generation of 3D models that agree with EPR distance data, which can sometimes be up to 70-80Å in magnitude (1, 335). A model can potentially be useful for interpretation of the data and how it correlates to biological function. On the other side, the addition of experimental data into computational methods is a useful means of decreasing the conformational space that needs to be sampled, thus improving the likelihood that a native-like structure is produced. Practically speaking, this means that, for a given number of generated models, a higher percentage of the models will exhibit the correct protein fold if experimental restraints are used. Therefore, RosettaEPR can be a valuable tool for structural biologists, who are interested in understanding the structure-function relationship of their systems of interest.

RosettaTMH improves Rosetta's conformational sampling of membrane protein topologies and can be used with experimental data

Even though Rosetta has been reported to predict the folds of MPs of varying sizes and levels of complexity (105, 117, 130, 354, 355), there has not been any recent development on RosettaMembrane. Further, the previously reported methods could not be used with the incorporation of experimental restraints to enhance sampling. Meanwhile, sthe default fragment-based assembly algorithm used for folding soluble proteins in Rosetta *can* be used with experimental restraints, but it is not suitable for folding MPs with high contact order, such as can be seen with α-helical MPs. On the other hand, in light of the difficulties encountered when using X-ray crystallography and NMR spectroscopy, computational methods for MP structure determination tightly integrating with experimental data are greatly needed. RosettaTMH was developed in order to meet this need.

Like other MP folding methods, including BCL::MP-Fold (37, 345), FILM3 (119, 135), and EVFold for MPs (105, 117, 118), RosettaTMH attempts to reduce conformational search space in order to improve the probability of obtaining native-like topologies. Unlike FILM3 and EVFold, however, RosettaTMH does not rely on multiple sequence alignments. Instead, it employs user-defined TMH-spanning information to divide the model into rigid bodies, which can then be translated or rotated in an extremely rapid, Monte Carlo fashion. These moves are scored according to the RosettaMembrane energy function. This is similar to what is done by BCL::MP-Fold. In contrast to the BCL, RosettaTMH then performs peptide fragment insertions, which can lead to the recovery of helical features, such as proline-induced kinks.

RosettaTMH can be combined with RosettaEPR to fold MPs in the presence of EPR distance data. It also expands upon RosettaEPR in that a more sophisticated scoring function for EPR restraints was optimized, as discussed in Chapter V and in *Summary of this Work*. Because there are only a few examples of MPs for which both experimental structures and EPR distances are available, EPR restraints were simulated using the BCL. When compared to folding with RosettaTMH alone, the addition of the EPR restraints greatly improved the algorithm's performance. This was also true for folding from an extended chain, but RosettaTMH with EPR data produced slightly better results for MPs having high contact order, which included the Na^+/galactose transporter, VSGLT (PDB ID: 2XQ2), and the carnitine transporter, CalT (PDB ID: 3HFX). More than 15% of rhodopsin models (PDB ID: 1U19) generated with RosettaTMH with EPR restraints exhibited the correct GPCR topology, which is also encouraging. These results suggest that RosettaTMH combined with EPR data--a popular structural biological technique for studying MPs--can serve as a starting point for MP structure determination. This could be especially helpful for experimentalists, who wish to have the assistance of a 3D model to interpret their data, propose new hypotheses, and postulate about protein functionality.

Future directions

Within the past decade, several new computational methods of protein and peptide structure determination have been developed and reported, especially within the Rosetta framework. The abilities of CS-Rosetta and RosettaNMR have been enhanced so that protein structures can be predicted from incompletely assigned CSs, and RosettaNMR can be used in an iterative fashion to fold proteins of up to 40 kDa in size

(35, 37, 108, 251, 263, 264, 322). RosettaEPR expanded upon the software suite's *de novo* folding capabilities to include the possibilty of folding proteins in the presence of EPR distance data. RosettaTMH has now been implemented in order to fold large, complex MPs, which is greatly assisted by the inclusion of experimental restraints. However, to understand the structural biological basis of disease and further advance molecular modeling technology, computationalists and experimentalists will likely need to work ever more closely and collaboratively moving forward. A few specific examples of where such partnerships could be advantageous are given below.

Further exploration of the structural mechanism of activation of the PrRP receptor

In Chapter II, the structure of PrRP, as well as how the propensity of the peptide to form an α-helix relates to its ability to activate the PrRP receptor, was described in detail. Notably, the less helical, truncated PrRP8-20 was much less stimulatory of the receptor when $Y^{2.64}$ or $W^{2.71}$ were mutated to alanine. The authors hypothesized that these two residues assisted the peptide ligand in forming its activating conformation, and, in their absence, PrRP8-20, unlike PrRP20, was not able to form the helical structure necessary. Even though stimulation with the full-length peptide was significantly decreased for the receptor mutants compared to wildtype, the EC_{50} was 4-5 times even greater when the same experiment was performed with PrRP8-20. Additional modeling of the full-length and truncated peptide in the wildtype and mutant receptor binding site could provide more insight into whether this hypothesis is correct. With the current knowledge concerning using Rosetta, this modeling could be performed in the implicit membrane environment, as was done with ghrelin (Chapter III).

More generally speaking, it is postulated that $Y^{2.64}$ and $W^{2.71}$, which are conserved residues in this receptor family, form a hydrophobic interaction that plays a role in causing a conformational change in the receptor (100, 102, 135, 377). One way to determine if these two residues are in close contact with one another is to perform a cross-linking experiment, in which the two residues of interest are mutated to cysteine. The general supposition is that, assuming the protein backbone and surrounding conformation remains relatively undisturbed, upon the addition of a cross-linking agent that is able to form disulfide bonds, the two residues will become cross-linked if they lie within the distance covered by the cross-linker. After separating the intra-molecularly cross-linked protein, usually by SDS-PAGE or size exclusion chromatography, the protein can be digested and analyzed via mass spectrometry. The sequences corresponding to the cross-linked peptide fragments is determined using computer software, and this information can be used to generate intra-molecular distance constraints on the protein (100, 105, 110, 117, 378). Another possibility for determining if the $Y^{2.64}$ and $W^{2.71}$ interact is to employ EPR spectroscopy, which can yield quantitative information on spin label distances.

On the modeling side, a more up-to-date comparative model of the PrRP receptor could be built, followed by extensive loop rebuilding and all-atom refinement. Further, Dr. Steven Combs in the Meiler laboratory has developed a scoring function that takes atomic orbitals based on valence shell electron pair repulsion (VSEPR) theory into account. After re-optimization of this scoring function for MPs, it could be used to refine the receptor model. In the case of soluble proteins, the new orbitals scoring function appears to recover π-π interactions better than the default Rosetta all-atom scoring

function. It is possible that this could be true for MPs as well; in this case, it may be informative for studying a possible interaction between $Y^{2.64}$ and $W^{2.71}$.

Ghrelin modeling based on experimental data collected in the presence of the ghrelin receptor

A conformational ensemble of ghrelin based on CSs from ssNMR was presented in Chapter III. This set of models, while exhibiting high flexibility, was in agreement with not only the experimental CSs, but also with results from spin diffusion experiments. However, distance restraints in the form of NOEs are highly desirable for the production of a higher quality model (or ensemble of models, rather). Unfortunately, NOEs would be likely be obtained by performing solution NMR NOESY experiments of ghrelin in detergent micelles, whereas the CSs reported in Chapter III were resultant from ssNMR in lipid vesicles. Similarly, secondary structural information of the peptide from CD spectroscopy could be informative but would also need to be collected in a micellar environment.

It would also be interesting to determine the conformation of ghrelin in the presence of the ghrelin receptor. CS information of the peptide based on ssNMR experiments--performed with both ghrelin and the ghrelin receptor in lipid vesicles--could be especially useful for understanding how the peptide changes when binding to the receptor. Further, the ensemble of ghrelin models, either from the current study or future studies, could be docked into an updated comparative model of GHSR1a, and, as with PrRP / PrRPR, could be used to propose additional mutants for analysis via signal transduction assays. (See reference (19, 28, 35, 36) and Appendix A for more information

on this process.) The resulting data can then be used to further refine the model for peptide / receptor interaction.

One of the main goals for the structural characterization of both PrRP and ghrelin is to improve our understanding of the mechanism of activation and function of their receptors. With that knowledge in hand, new therapeutics, most likely in the form of small-molecules or peptidomimetics, can be synthesized, tested *in vitro*, *in vivo*, and eventually in clinical trials for the treatment of a number of disorders, including obesity and type II diabetes in humans.

Expanding the capabilities of RosettaEPR to fold membrane proteins

The introduction of RosettaEPR (29-31, 33, 135, 180) (Chapter IV) served as a proof of concept that a) an EPR distance KBP based on the MTSSL can be used to improve sampling of native-like folds for small, soluble proteins, and b) it has the capability to be used for full-atom modeling of relatively small proteins to atomic detail accuracy. However, the reported work serves only as a starting point for the development of RosettaEPR. It has already been expanded by the addition of an all-atom representation of the MTS spin label, which can be used with the Rosetta soluble protein scoring function (19, 379), though this has not yet been tested on MPs. In order to more accurately model MPs with RosettaEPR, the statistics used to generate the EPR KBP should be re-calculated over a database of MPs. Futher, the spin label roatmer library should be tested on the leucine transporter, LeuT, for which a crystal structure of the spin labeled protein is available (37, 380). It should be noted that the EPR KBP and spin label rotamer library are specifically based on the MTS spin label. However, KBPs and

rotomaer libraries based on other spin labels, such as 2-Carboxyanthracene MTSEA amide (MTSEA), can also be created in a similar manner.

It is expected that the implementation of an EPR accessibility restraint in RosettaEPR would also improve its performance on MP modeling. This is because the spin label accessibility at a particular site can yield information on both the surrounding environment (40-42, 74, 381) and the membrane depth of the residue (45-49, 307, 381, 382). Axel Fischer and Dr. Nathan Alexander have shown that the inclusion of EPR accessibility restraints taking into account only spin label exposure (as opposed to membrane depth) significantly improves the recovery of correctly predicted MPs using BCL::MP-Fold (submitted). This would probably be the case for RosettaEPR and RosettaTMH as well.

Finally, it is important to remember that EPR distance data do not consist of single distance values. Rather, distances are reported as probability distributions, which can sometimes be broad, resembling a hill instead of a spike. Furthermore, when measuring inter-residue distances under different conditions, the average distance does not necessary change, but the shape of the probability distribution may shift, as was seen with LeuT (105, 130, 336, 383). Therefore, instead of a single model, it would be better to predict an ensemble of models consistent with the distribution of inter-residue distances observed experimentally. This is now possible thanks in large part to the work of Samuel DeLuca (Meiler laboratory) and Dr. Mathew O'Meara (Stoichet laboratory) in implementing MySQL database infrastructure into the Rosetta software suite.

Improvement of RosettaTMH and testing RosettaTMH's prediction accuracy using real experimental data

The implementation of RosettaTMH, which was explained in detail in Chapter V, provides a foundation for the *de novo* folding of other types of proteins (e.g., β-barrels, α-β proteins, etc.) using rigid body sampling and fragment-based assembly. However, the main focus of this work is on the accurate determination of 3D α-helical MP structures. More specifically, it would be most interesting to explore the synergistic effect that arises from pairing this new sampling method with EPR data.

One of the main challenges with the *de novo* folding of native-like MP topologies via rigid body assembly is the tendency of the secondary structural elements (SSEs) to move increasingly farther away from one another. This results in poorly packed structures. In order to prevent this from occurring, a loophash filter could be implemented in RosettaTMH, the intention of which would be to assist in generating TMH arrangements that could later be connected by loops. A loop length KBP similar to that used in BCL::Fold (115, 260) and BCL::MP-Fold (253, 345) would likely improve the performance of RosettaTMH because, once helices are in relatively close proximity to one another (via the loophash filter), the new energetic term would favor conformations that resemble inter-SSE orientations observed in nature.

In order to truly and rigorously test RosettaTMH's ability to accurate predict MPs in the presence of EPR restraints, it would be important to use distances obtained from actual EPR measurements. There are only a few MPs for which such a benchmark would be possible, including the ABC transporter MsbA, the K^+ ion channel KcsA, and rhodopsin, which is a GPCR. This is because there are EPR data and at least one

experimentally determined 3D structure, for these proteins (74, 145-147, 260, 261, 308, 309, 384-386). This benchmark would also require that RosettaTMH be able to fold symmetric MPs, which it does not currently do. The Rosetta Symmetric FoldAndDock algorithm (256, 302) allows for the folding of symmetric homo-oligomeric soluble proteins. It is expected that this framework can also be used for MPs.

A blind test, in which the answer (i.e., correct structure) is not known, would be an excellent means of ascertaining the ability of RosettaTMH to fold MPs. This would probably require collaboration with at least one research group, who would be willing to provide EPR experimental data. Ideally, a set of experimental data would be used to generate models with RosettaTMH. These models would then be cross-validated with additional EPR experiments. If there is no X-ray or NMR structure of the MP of interest in the pipeline, an iterative approach to model validation similar to that used for the PrRP / PrRPR system (19, 258) (Appendix A) could be taken. This would be especially exciting for the field of MP structural biology because it would be a novel means of determining the 3D structures of MPs using the power of both computational methods and EPR spectroscopy.

Concluding remarks

Structural biologists have traditionally been categorized according to their method of choice for studying biomolecules. That is, one is often referred to as an X-ray Crystallographer, an NMR Spectroscopist, an Electron Microscopist, a Computationalist, etc. This dissertation serves as an example of how the field is moving away from these individualistic titles in the direction of the more generally and aptly named Structural

Biologist. By leveraging the advantages of multiple techniques and methodologies for structural elucidation, as was demonstrated by the herein reported work, the structural biological community can progress to a better understanding of how 3D structure affects inter-molecular interactions, dynamics, and ultimately, biological function. Dr. Stephen Harrison at Harvard University articulated such a vision for structural biology in a commentary published in *Nature Structural and Molecular Biology*:

> "...[S]tructural biology must seek to understand information transfer in terms of its underlying molecular agents by analyzing the molecular hardware that executes the information-transfer software. Unlike most man-made computers, the hardware and software of physiological regulation co-evolved. The possibilities for storage, retrieval, transfer and destruction of information are not independent of the molecular devices that execute these functions." (143, 297)

Thus, a marriage of structural and systems biology appears to be a promising means of reaching this goal, and many researchers are already making great strides to this end. The work of Andrej Sali at the University of California-San Francisco in the structure determination of macromolecular assemblies by combining modeling with multiple types of experimental data serves as a prime example (135, 387). This dissertation work, which has focused on using computational, EPR, and NMR hybrid techniques, is just one small step towards achieving a full structural characterization of physiological processes.

The experiments described in this chapter would be only first steps leading towards a better understanding of GPCR / peptide interactions and improved computational de novo folding of helical membrane proteins. However, broader, more challenging questions face the structural biological—and the general biomedical—field. In recent years, one of the main questions the protein structure prediction community asks itself is: Why are the capabilities of protein modeling methods plateauing? This is especially true for large proteins with complicated topologies, such as those that exhibit the LeuT fold. A key drawback of modeling proteins with Rosetta in its current state is that it does not take protein dynamics into account. Further, the surroundings in which the proteins reside, be they aqueous or hydrophobic, are only a statistically based implicit representation of reality. Finally, proteins do not exist in isolation, but rather in highly fluid, crowded environments. The numerous interactions that these biomolecules encounter play an important role in their conformations.

Ideally, we would model the plethora of inter-molecular interactions in full detail, including computing molecular orbitals from first principles. It is possible that this approach would render an accurate picture of what goes on at the single molecule level. Unfortunately, this is computationally intractable, and likely will be for a long time to come. Scientists interested in predicting molecular structure have worked around this by including empirical data and statistics in their methods. The problem is that these approaches are difficult to dissect, troubleshoot, and improve. Therefore, we do not have a robust, analytical explanation of why, Rosetta for instance, does or does not fold certain proteins accurately. However, the same could be said for other processes and areas of

study, including protein crystallization, neuroscience and connectomics, and even dark matter and dark energy (in physics).

In order to make headway on these known unknowns, as well as even more unknown unknowns, the scientific community--and society as a whole—must work together more closely than ever. Both pure and use-inspired basic research, often referred to as Bohr's and Pasteur's quadrants, respectively, ought to remain a national priority if the United States is to maintain its competitive economic edge (437, 438). This would require significant investments in both scientific discovery and method development, which seems infeasible given the current funding and political environment. Fortunately, we can leverage already existing resources and personnel via government-university-industry partnerships (438). Moving in this direction necessitates continued and incresed effective communication across disciplines and sectors.

APPENDIX A

LIGAND-MIMICKING RECEPTOR VARIANT DISCLOSES BINDING AND ACTIVATION MODE OF PROLACTIN RELEASING PEPTIDE

This work is based on the publication (Rathmann*, Lindner*, DeLuca*, Kaufmann, Meiler, and Beck-Sickinger, 2013). *These authors contributed equally.

Summary

The prolactin-releasing peptide receptor (PrRPR) and its bioactive RF-amide peptide (PrRP20) have been investigated to explore the ligand binding mode of peptide G-protein coupled receptors (GPCR). By receptor mutagenesis we identified the conserved aspartate in the upper transmembrane helix 6 ($D^{6.59}$) of the receptor as the first position that directly interacts with arginine 19 of the ligand (R^{19}). Permutation of $D^{6.59}$ with R^{19} of PrRP20 led to $D^{6.59}R$, which turned out to be a constitutively active receptor mutant (CAM). This suggests that the mutated residue at the top of transmembrane helix 6 mimics R^{19} by interacting with additional binding partners in the receptor. Next, we set up a comparative model of this CAM because no ligand docking is required, and selected a next set of receptor mutants to find the engaged partners of the binding pocket. In an iterative process we identified two acidic residues and two hydrophobic residues that form the peptide ligand binding pocket. As all residues are localized on top or in the upper part of the transmembrane domains we clearly can show that the extracellular surface of the receptor is sufficient for full signal transduction for PrRP, rather than a deep membrane binding pocket. This contributes to the knowledge of the binding of

peptide ligands to GPCR and might facilitate the development of GPCR ligands, but also provides new targeting of CAM involved in hereditary diseases.

Introduction

Identification of direct receptor-ligand interactions for the approximately 800 identified G protein-coupled receptors (GPCR) is as challenging as it is important for drug discovery (25), as 50% of all currently available drugs target the specific manipulation of GPCR activity (24, 388). The PrRP receptor superfamily is expressed in almost all cells/tissues, is involved in a plethora of different signalling pathways, and plays an important role in a large variety of physiological processes.

The prolactin-releasing peptide receptor (PrRPR) was originally isolated from rat hypothalamus (389). PrRPR has been detected widely throughout the human and rat brain (31) and most commonly activates the G_q protein-coupled signalling pathway (390). Its eponymous endogenous ligand, the prolactin-releasing peptide (PrRP), was identified in 1998 by a reverse pharmacology approach on the basis of orphan GPCR (28, 391). PrRP features two equipotent isoforms, PrRP31 (31 residues) and an N-terminally truncated PrRP20 (20 residues) (28, 390). PrRP is an RF-amide peptide, consisting of a common carboxy-terminal arginine (R) and an amidated phenylalanine (F) motif and plays a role in energy metabolism, stress responses, circadian rhythm, analgesia, and in anorexigenic effects (391, 392). Structure-activity relationship studies of PrRP using N-terminally truncated mutants and alanine substitution within these constructs (30, 34, 36) demonstrated the biological significance of the C-terminal R and F residues, and the amidation of the C-terminus.

Site-directed mutagenesis is a powerful and widely used tool to study receptor activation. This approach alone can provide insight in the function of GPCR, but it is often used in combination with information provided by other techniques, such as crystallography or molecular modeling, in order to relate receptor function to a tertiary structure (393). The conserved $D^{6.59}$ residue of the Y receptor (YR) family was shown to interact with a specific R of either human pancreatic polypeptide or neuropeptide Y (NPY) in a subtype-specific manner (394, 395). The numbering of receptor residues has been performed as suggested by Ballesteros and Weinstein (38). PrRPR shares its phylogenic origin with Y receptors (396), leading to sequence similarities (Figure 44A) and a number of conserved residues, including $D^{6.59}$ (Figure 44C). Furthermore, the ligands of these receptors are structurally similar (35) and share a similar C-terminal sequence (Figure 44B). While the RF-amide motif was previously identified as a major requirement for PrRP-induced agonist activity (30, 34), the critical residues on the receptor remain unknown, and the ligand binding mode is still poorly understood.

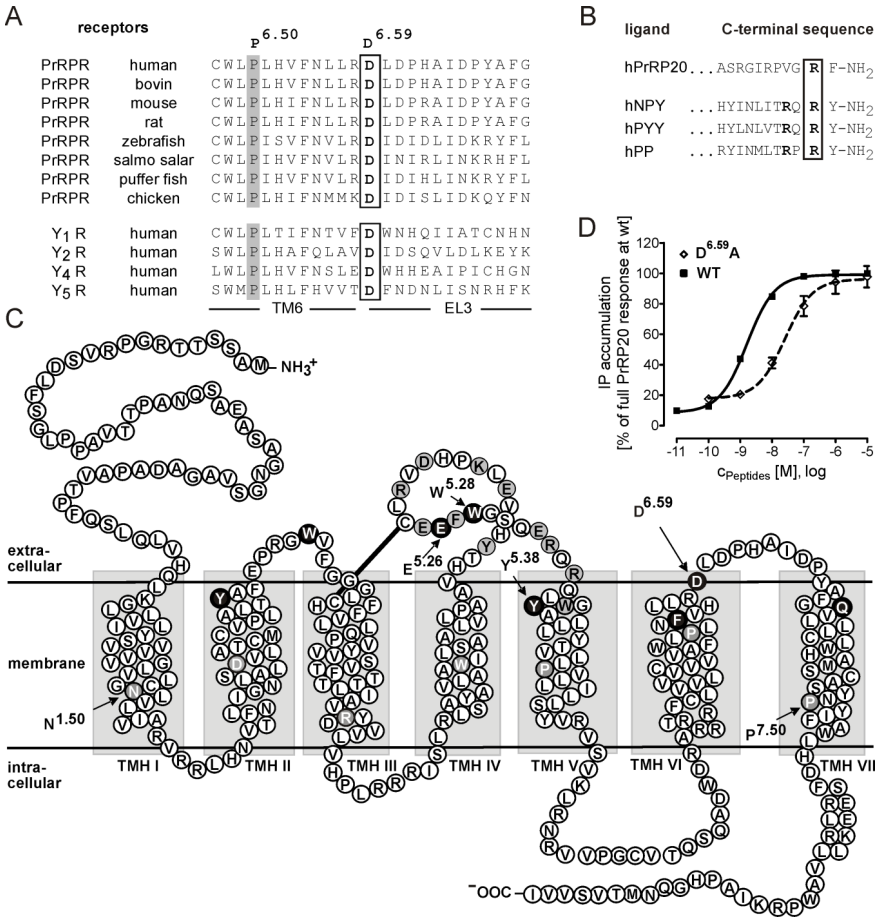

Figure 44: Identification of the conserved $D^{6.59}$ residue in the hPrRPR sequence as potential spot of interaction

A) Conservation of $D^{6.59}$ shown in the amino acid sequence alignment. The region of upper transmembrane helix (TMH) 6 and the beginning of the subsequent extracellular loop (EL) 3 of the four human Y receptor subtypes and the PrRPR is presented. Sequence alignment and description was taken from: http://www.gpcr.org/7tm/. B) Comparison of the C-terminal amino acids of the Y receptor ligands and the PrRP20. C) Snake plot representing the sequence of the human PrRPR. Residues highlighted in black were investigated as double mutants in the $D^{6.59}R$ construct. Selective alanine-scan was performed on residues pictured in grey, resulting in no functional alteration. Residues with white letters in grey correspond to the X.50 nomenclature (38). D) IP accumulating signal transduction assay performed for 1h with COS-7 cells in a concentration-response dependent manner reveals an impact of $D^{6.59}A$ PrRPR in comparison to the wt PrRP receptor. Data represent the mean ± s.e.m. of multiple independent experiments (n = 32 for hPrRPR, and n = 12 for $D^{6.59}A$ PrRPR). Receptor activity is expressed as percentage of the full response of PrRP20 at the wt PrRP receptor.

Here, we describe the first mutagenesis study of the human PrRP receptor (PrRPR). We used the extracellular region to elucidate the binding site and the molecular

mechanism of GPCR activation. Considering the relevance of the C-terminal R and F residues of PrRP for receptor binding, we applied the concept of double cycle mutagenesis approach (395, 397, 398) and identified the first direct contact point between PrRP20 and the PrRPR, consisting of the conserved $D^{6.59}$ and the R^{19} residue of PrRP20. To prove the existence of this interaction, we switched the residues involved in the salt bridge formation and created $D^{6.59}$R PrRPR and D^{19}PrRP20. This newly introduced R in the receptor variant $D^{6.59}$R might serve as surrogate for the absent R^{19} of the ligand as it led to a new type of constitutive activity. Given the lack of data of experimentally determined structures of peptide GPCR, we developed a comparative model of the human PrRPR. By combining molecular modelling with double cycle mutagenesis experiments in the framework of this constitutively active mutant (CAM), we conceived an effective strategy to explore structural determinants of ligand recognition on a molecular level. More specifically, we were able to identify $Y^{5.38}$, $W^{5.28}$, $E^{5.26}$, and to some extend $F^{6.54}$ to be involved in receptor activation and ligand binding. This combinatory approach enabled us to clarify the double binding mode of R^{19} of the peptide ligand, which has two putative interaction partners within the PrRPR, $E^{5.26}$ and $D^{6.59}$. The assembled experimental data were used to generate a model of the PrRP/receptor interaction in molecular detail. Furthermore our data describe the binding mode of a peptide ligand to GPCR by solely interacting with residues localized in the extracellular domain or upper part of the TM helices. In our approach we identified a receptor mutant with constitutive activity, which most likely relies on mimicking a direct ligand-receptor interaction. This provides knowledge on the function of an active mode of GPCR and may be applied to other peptide GPCR. More specifically, we were able to identify $Y^{5.38}$, $W^{5.28}$, $E^{5.26}$, and to

some extend $F^{6.54}$ to be involved in receptor activation and ligand binding. This combinatory approach enabled us to clarify the double binding mode of R^{19} of the peptide ligand, which has two putative interaction partners within the PrRPR, $E^{5.26}$ and $D^{6.59}$. The assembled experimental data were used to generate a model of the PrRP/receptor interaction in molecular detail. Furthermore our data describe the binding mode of a peptide ligand to GPCR by solely interacting with residues localized in the extracellular domain or upper part of the TM helices. In our approach we identified a receptor mutant with constitutive activity, which most likely relies on mimicking a direct ligand-receptor interaction. This provides knowledge on the function of an active mode of GPCR and may be applied to other peptide GPCRs.

Materials and methods

Peptide synthesis

Rink amide resin (NovaBiochem; Läufelfingen, Switzerland) was used to synthesize PrRP20, A^{19}PrRP20, D^{19}PrRP20, and A^{20}PrRP20 by automated solid phase peptide synthesis (Syro; MultiSynTech, Bochum, Germany) as previously described, using the orthogonal Fmoc/^{t}Bu (9-fluorenyl-methoxycarbonyl-*tert*-butyl) strategy (399). Purification and verification of the peptides was achieved as previously described (Table 25) (171).

Table 25: Binding affinity of single amino acid replacements of PrRP20 at the human PrRP receptor wildtype. COS-7 cells were transiently transfected with wildtype PrRP receptor

No.	Peptide	Sequence	mass (m/z)		HPLC			Binding assay[d]	
			calcd. $[M]^+$	exp. $[M+H]^+$	ACN [%]	MeOH [%]	purity [%]	IC_{50} [nM]	IC_{50}(peptide) / IC_{50}(PrRP20)
1	PrRP20	TPDINPAWYASRGIRPVGRF-NH_2	2272.6	2273.7	40.3[a]	65.5 [b]	>99	3.6 ± 0.5	**1**
2	A^{19}PrRP20	TPDINPAWYASRGIRPVG**A**F-NH_2	2187.5	2188.4	41.6[a]	70.8[c]	>99	> 10 000	**> 2 700**
3	D^{19}PrRP20	TPDINPAWYASRGIRPVG**D**F-NH_2	2231.5	2231.4	38.5[a]	67.4[b]	>99	> 10 000	**> 2 700**
4	A^{20}PrRP20	TPDINPAWYASRGIRPVGR**A**-NH_2	2196.5	2196.2	37.7[a]	61.6[b]	>99	869 ± 577	**241**

[a] 10 % to 60 % ACN (0.08 % TFA) in water (0.1 % TFA) over 30 min.
[b] 20 % to 100 % MeOH (0.08 % TFA) in water (0.1 % TFA) over 40 min.
[c] 30 % to 100 % MeOH (0.08 % TFA) in water (0.1 % TFA) over 30 min.
[d] The IC_{50} value was determined by competition assays using N [propionyl^{3}H] hPrRP20.

DNA extraction from SMS-KAN

To obtain genomic DNA from SMS-KAN cells (human neuroblastoma cells, DSMZ, Braunschweig, Germany), approximately 1 million cells were digested overnight at 55°C with 500 µl lysis buffer (1 M NaCl, 20% SDS, 0.5 M EDTA, 1 M Tris, pH 8.5 was adjusted using hydrochloric acid (HCl)) containing 50 µg proteinase K (Promega, Mannheim, Germany). Genomic DNA was extracted using phenol/chloroform and precipitated from the aqueous phase with isopropanol, washed with ethanol and then dissolved in water.

Cloning and mutagenesis of the PrRP receptors in eukaryotic expression vectors

The coding sequence of the human PrRPR was obtained by PCR amplification from the isolated genomic DNA of SMS-KAN cells and cloned into the eukaryotic expression vector pEYFP-N1 (Clontech, Heidelberg, Germany) C-terminally fused to EYFP, using the *Xho*I and *Bam*HI restriction site to result in the construct phPrRPR_EYFP-N1. The correctness of the entire coding sequence was confirmed by DNA sequencing using the dideoxynucleotide (ddNTP) termination method developed by Sanger (23). Plasmids encoding single point mutations (Table 26 and Table 28) were prepared by using the QuikChange™ site-directed mutagenesis method (Stratagene, CA, USA) with the desired mutagenic primers. For intermolecular double-cycle mutagenesis approaches, the single alanine mutated receptor constructs were investigated, using single alanine modified PrRP20 analogs. Plasmids encoding double mutations containing $Y^{2.64}A$, $W^{2.71}A$, $E^{5.26}A$, $E^{5.26}R$; $W^{5.28}A$, $D^{6.59}A$, $F^{6.54}A$ or $Q^{7.35}A$ as a second mutation, respectively, were prepared by using the QuikChange™ site-directed mutagenesis

approach with the $D^{6.59}R$ or $D^{6.59}A$ construct as template. In addition, all PrPR receptor constructs were also generated N-terminally fused to the coding sequence of the hemagglutinin (HA)-tag. The entire coding sequence of each resulting receptor mutant was proven by sequencing.

Cell culture

Cell culture material was supplied by PAA Laboratories GmbH (Pasching, Austria). Culture of COS-7 (African green monkey, kidney), HEK293 (human embryonic kidney), and SMS-KAN cells was done as recommended by the supplier (DSMZ, Braunschweig, Germany). Briefly, cells were grown as monolayers at 37°C in a humidified atmosphere of 5% CO_2 and 95% air. COS-7 cells were cultured in Dulbecco's Modified Eagle's Medium containing 10% (v/v) heat-inactivated fetal calf serum (FCS), 100 units/ml penicillin and 100 µg/ml streptomycin and HEK293 cells were grown in DMEM / Ham' F12 (1:1) without L-glutamine containing 15% (v/v) heat-inactivated FCS as previously described (395, 400). SMS-KAN cells were maintained in nutrient mixture Ham's F12 / Dulbecco's modified Eagle medium (1:1) with 15% (v/v) FCS, 4 mM glutamine, and 0.2 mM non-essential amino acids (401).

Fluorescence microscopy

HEK293 cells (1.2×10^5) were seeded into 8-well chamber slides (ibidi, Munich, Germany). The transient transfection of HEK293 cells were performed using 0.1 µg to 1µg vector DNA and 1 µl Lipofectamin™ 2000 transfection reagent (Invitrogen GmbH, Karlsruhe, Germany) according to the manufacturer's instructions. The nuclei were

visualized with Hoechst 33342 (1 µg/ml; Sigma Aldrich, Taufkirchen, Germany) for 10 min after 1h of starving with OPTI®-MEM I Reduced Serum Medium (Invitrogen GmbH, Karlsruhe, Germany). Fluorescence images were obtained using an ApoTome Imaging System with an Axio Observer microscope (Zeiss, Jena, Germany). All investigated receptors were correctly integrated in the membrane as confirmed by live-cell microscopy (Figure 45A).

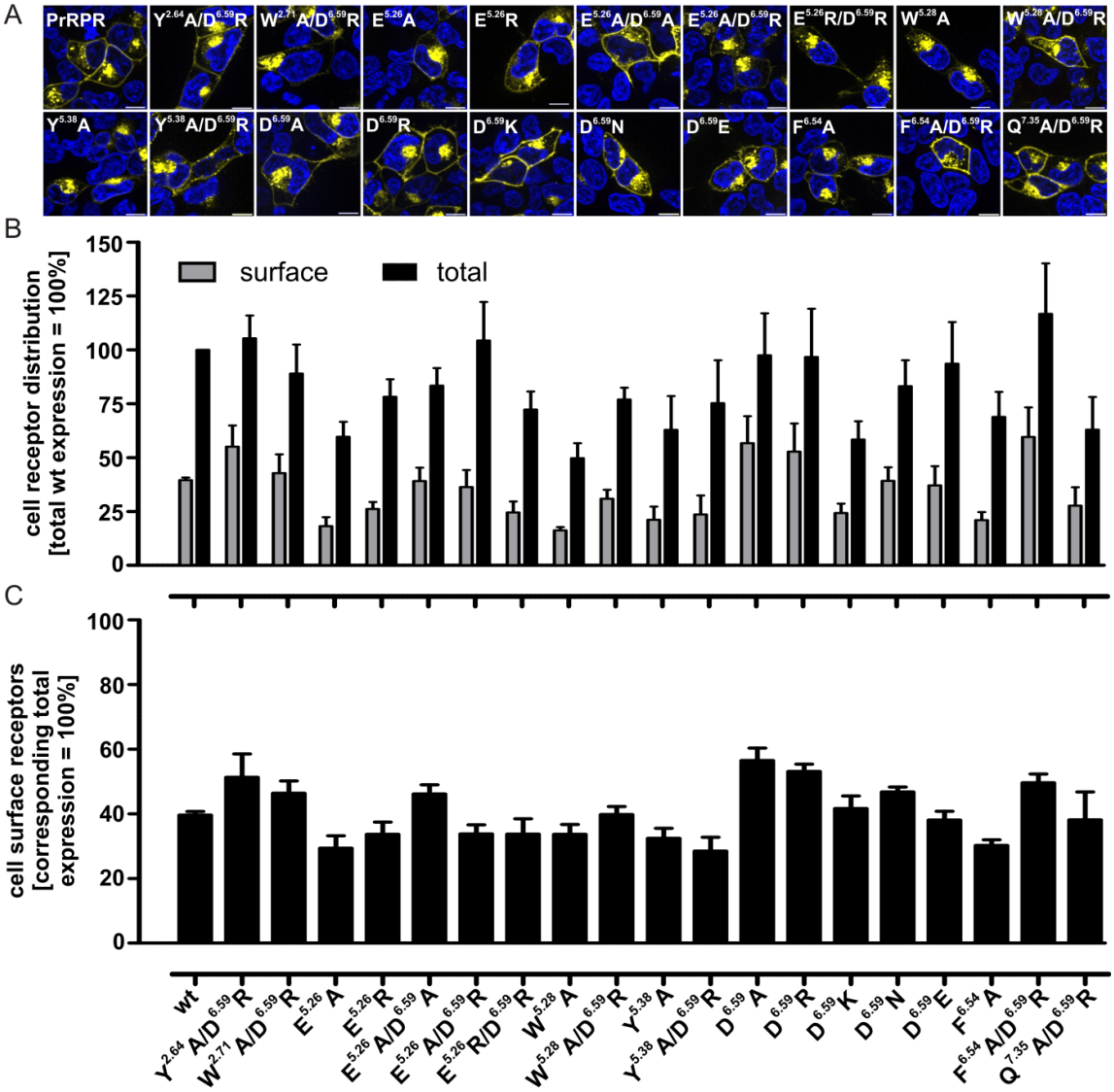

Figure 45: Surface localization of PrRPR variants in HEK293 cells
A) Cell surface expression of wt PrRPR and investigated PrRPR mutants. HEK293 cells were transiently transfected with different PrRPR mutants, C-terminally fused to eYFP. The nuclei were visualized with Hoechst 33342. Scale bars represent 10 µm. B) Quantification of cell surface and total receptors by ELISA. The amount of cell surface receptors was measured as described under *Materials and methods*. Data are shown as mean ± s.e.m. of four independent experiments, each performed in triplicate. C) Relative cell surface expression levels of each receptor construct as percentage of each total receptor expression by ELISA. Data shows the capability of the individual receptor mutants to be exported to the plasma membrane, independently from the transfection efficiency. Data is calculated from Panel B and presented as mean ± s.e.m. of four independent experiments, each performed in triplicate.

Quantification of receptor cell surface localization by cell surface ELISA

To quantify plasma-membrane receptors, a cell surface ELISA was performed using an antibody directed against the native 15 N-terminal amino acids of the PrRPR. 50,000 HEK293 cells were grown in 96-well plates and transfected with the PrRP wt receptor or its mutants after reaching 75-85% of confluence. The cells were starved with OPTI®-MEM I (30 min) 17 hours post-transfection and fixed in 4% paraformaldehyde (30 min). For immune-staining, cells were blocked with 2% BSA and permeabilized with 0.5% Triton X-100, 2% BSA in Dulbecco's Modified Eagle's Medium for 1 hour (37°C) to determine total receptor amounts, whereas surface expressed receptors were quantified without permeabilization. Incubation was performed with the primary antibody (1:2000 dilutions) for 2 hours (25°C) and followed by 1.5 hour (25°C) incubation with the secondary antibody (1:5,000). Receptors were detected by using rabbit anti-N-terminus (GPR10 antibody [N1], GTX108137, GeneTex) followed by horseradish peroxidase-conjugated goat anti-rabbit IgG (sc-2004, Santa Cruz, Heidelberg, Germany). The results were fully confirmed in a second independent ELISA set up, using a peroxidise-conjugated anti-HA-antibody (1:1000 dilutions, 12CA5, Roche, Mannheim, Germany) versus the N-terminally fused HA-tag of the generated PrRPR constructs (data not shown). Quantification of the bound peroxidase was performed as described and analysis performed with the GraphPad Prism 5.03 program (14). Values are presented as mean values ± s.e.m. of four individual experiments, measured in triplicate.

Radioligand binding studies

For radioligand binding studies, 1.5×10^6 COS-7 cells were seeded into 25 cm^2 flasks. At 60-70% confluency, cells were transiently transfected using 4 µg vector DNA and 15 µl of Metafectene™ (Biontex Laboratories GmbH, Martinsried/Planegg, Germany). Approximately 24 h after transfection binding assays were performed on intact cells using N [propionyl^{3}H] hPrRP20. Binding was determined with 1 nM N [propionyl^{3}H] hPrRP20 in the absence (total binding) or in the presence (non-specific binding) of 1 µM unlabeled hPrRP20, respectively, as described previously (172, 175). Our former evaluated protocol (176) was used to obtain N [propionyl^{3}H] hPrRP20 by selective labelling with a specific activity of 3.52 TBq/mmol and resulting in a K_d-value of 0.58 nM. Specific binding of each PrRP receptor mutant was compared to specific binding of the PrRP wt receptor. IC_{50}-values and the K_d-value were calculated with GraphPad Prism 5.03 (GraphPad Software, San Diego, USA), fitted to a one-site competition or a one-site binding model, respectively. Triplicates were measured in at least two independent experiments for the determination of IC_{50}-values, whereas one experiment in triplicate was made for K_d-value estimation.

Signal transduction assay

Signal transduction (inositol phosphate, or IP, accumulation) assays were performed as previously described with minor modifications (171). The time of incubation was increased to 3 h for the double mutants of PrRPR and reduced to 1h for measurement of concentration-response curves. To test for constitutive activity, COS-7 cells were incubated without agonist for 1 h, 3 h, and 6 h at 37°C. Each ligand-receptor

interaction was analyzed with the GraphPad Prism 5.03 program by establishing the corresponding data set from different experiments. All signal transduction assays were repeated at least twice independently and measured in duplicate. The global curve fitting function of GraphPad Prism 5.03 was asked to determine given EC_{50}-ratios. The statistical significance of relevant samples was computed by using the unpaired student's t-test, based on the means, values with $P < 0.05$ were considered to be significant.

Multiple sequence alignment

ClustalW (257) was used to align the primary sequence of the PrRPR with the sequences of mammalian Y and PrRP receptors. Next, the transmembrane regions of six GPCR of known structure were structurally aligned with Mustang (259). The profiles resulting from these first two steps were then aligned to one another with ClustalW, and the human PrRPR sequence alignment used for modelling was taken from this final profile-profile alignment. The C-terminal 310 residues of the PrRPR primary sequence were threaded onto the 3D coordinates of six available GPCR experimental structures; PDBIDs: 1U19 (260), 3CAP (402), 3DQB (403), 2RH1 (67), 2VT4 (404), 3EML (405).

Construction of the comparative models

Extracellular loop regions were reconstructed using kinematic loop closure (406) and cyclic coordinate descent (CCD) (255) as implemented in the Rosetta v3 software suite. The models were refined with the Rosetta v3 all-atom energy function. Energetically favourable models were grouped into 15 structurally similar groups by k-means clustering, and the lowest scoring models of each cluster were analysed. Models

based on the template PDB 3DQB had the lowest energy and were used to inform the mutagenesis studies.

Model refinement and peptide docking

The comparative model constructed in light of the new mutagenesis data was generated using the original multiple sequence alignment. To model the PrRPR/ligand complex, an iterative peptide docking and loop remodeling procedure was performed: Energetically favorable changes in orientation were determined using the RosettaMembrane all-atom energy function (130)**.** The PrRP8-20 model was docked into the putative binding site of the receptor while allowing remodeling of ELs 1, 2, and 3. Using the RosettaDock protocol (407)**,** translational movements of the peptide of up to 4Å were allowed in three dimensions and the peptide was allowed to rotate along its x, y, and z-axes by up to 10°. Loop regions were constructed using CCD (255). The conformational search was enhanced by conducting the modeling in the presence of loose distance restraints where models that placed $D^{6.59}$, $E^{5.26}$, $W^{5.28}$, and $Y^{5.38}$ within 10Å of R^{19} of the peptide were more energetically favorable than those that did not. The PrRP8-20 model was generated by *de novo* folding the peptide using RosettaNMR with sparse NMR chemical shift and distance data (106). Of 19,241 PrRP/receptor complex docked models, the top ten by total score were analyzed. Two of these models were considered structurally redundant, leaving eight unique models that agree with the experimental data presented herein (Figure 52).

Results

R^{19} of the endogenous ligand PrRP20 interacts with the $D^{6.59}$ of PrRPR

Based on the data of the NPY/YR system (394, 395), we hypothesized $D^{6.59}$ to be the interaction partner of R^{19} in the PrRP/PrRPR system. To test this hypothesis, charge and size prerequisites in position $D^{6.59}$ were elucidated by systematic substitution to $D^{6.59}$A, $D^{6.59}$E, $D^{6.59}$N, $D^{6.59}$R, and $D^{6.59}$K (Table 26). The expected impact on function was confirmed by the right-shifted concentration-response curve of $D^{6.59}$A, compared to the wildtype (wt) receptor after stimulation with PrRP20 (Figure 44D). The increased EC_{50}-value (26 nM) of the $D^{6.59}$A mutant confirms the importance of the $D^{6.59}$ side-chain. In addition, the results obtained for the other $D^{6.59}$ single mutants support the hypothesis of an ionic interaction; $D^{6.59}$E behaves similarly to wt, the oppositely charged $D^{6.59}$K shows strong effects in potency and the bulkier, more positively charged $D^{6.59}$R is not tolerated (Table 26). The impact of the substitutions increases as follows: E<A<N<K<R, showing that the lack of charge is a first critical component. This is followed by necessities in space and strength of the opposing charged K and R at position 6.59, showing different and increasing repulsion of the substitutions by PrRP20 stimulation (Table 26). Therefore, the charge seems to be a major prerequisite at position 6.59.

Table 26: Functional characterization of wildtype and $D^{6.59}$ PrRP receptor mutants with different PrRP analogs

IP accumulating signal transduction assay was performed for 1 hour with different concentrations of modified PrRP20 peptides to determine EC_{50}-values from concentration-response curves.

PrRPR mutants	PrRP20				A^{19}PrRP20			D^{19}PrRP20	
	EC_{50} [nM][a] (pEC_{50} ± SEM)	EC_{50}-ratio[b] (mut/wt)	Emax ± SEM [%][c]	N	EC_{50} [nM][a] (pEC_{50} ± SEM)	EC_{50}-ratio[b] (analog/wt)	N	EC_{50} [nM][a] (pEC_{50} ± SEM)	N
wt	1.66 (8.78 ± 0.04)	1	100	32	1202 (5.92 ± 0.08)	736	11	1318 (5.88 ± 0.12)	5
$D^{6.59}A$	26 (7.59 ± 0.15)	15	98 ± 7	12	166 (6.78 ± 0.17)	0.16	3	6456 (5.19 ± 0.16)	4
$D^{6.59}R$	ND[d]	ND[e]	60 ± 13	4	> 10 000 (< 5)	ND[e]	2	138 (6.86 ± 0.23)	3
$D^{6.59}K$	1380 (5.86 ± 0.20)	847	90 ± 10	3	NT	-	-	115 (6.94 ± 0.17)	2
$D^{6.59}E$	3.98 (8.4 ± 0.19)	2	106 ± 10	2	NT	-	-	NT	-
$D^{6.59}N$	36.3 (7.44 ± 0.25)	22	105 ± 20	2	NT	-	-	NT	-
$E^{5.26}A$	537 (6.27 ± 0.09)	361	81 ± 6	8	> 10 000 (< 5)	21	3	NT	-
$E^{5.26}R$	> 10 000 (< 5)	ND[e]	70 ± 6	2	ND[d]	ND[e]	2	> 10 000 (< 5)	2
$E^{5.26}A$/ $D^{6.59}A$	ND[d]	ND[e]	58 ± 7	2	ND[d]	ND[e]	2	NT	-
$E^{5.26}R$/ $D^{6.59}R$	NR	ND[e]	8 ± 2	2	NR	ND[e]	2	ND[d]	2

NT represents not tested, NR indicates no response after stimulation with 10 µM and N displays the number of individual experiments.

[a] EC_{50}-/pEC_{50}-values were calculated from the mean ± s.e.m. of N independent experiments, measured in duplicate.

[b] Efficacy was determined as percentage compared to full PrRP20 response at wt

[c] The ratio was determined using the Prism 5.03 global fitting function for EC_{50} shift determination.

[d] ND, not determined because of lack of efficacy. The plateau of the curve was not reached.

[e] ND, not determinable

The signal transduction results obtained for PrRPR stimulation with peptide analogs A^{19}PrRP20 and A^{20}PrRP20 confirmed the essential influence of the formerly described RF-amide motif with respect to binding and signaling (Table 25, Table 26, and Table 27) (30, 34, 36). Circular dichrosim (CD) spectroscopy showed that these variations have no influence on the PrRP20 overall structure, at least, not detectable by CD (data not shown).

Table 27: Functional characterization of wildtype and $D^{6.59}$ PrRP receptor mutants with A^{20}PrRP

IP accumulating signal transduction assay was performed for 1 hour with different concentrations of A^{20}PrRP to determine EC_{50}-values from concentration-response curves.

PrRPR mutants	**A^{20}PrRP20**	
	EC_{50} [nM] (pEC_{50} ± SEM)	**N**
wt	17.8 (7.75 ± 0.11)	8
$D^{6.59}$A	ND[a]	2
$D^{6.59}$R	NR	2
$E^{5.26}$R/ $D^{6.59}$R	NR	2

NR indicates no response after stimulation with 10 µM, and N displays the number of individual experiments.

[a] ND, not determined because of lack of efficacy. The plateau of the curve was not reached.

Double cycle mutagenesis suggests additional receptor region "X" critical for peptide binding.

The concentration-response curve of the $D^{6.59}$A receptor with PrRP20 reveals a 15-fold elevated EC_{50}-value (Figure 46A and Table 26), whereas the wt receptor stimulated with A^{19}PrRP20 results in a 736-fold elevated EC_{50}-value (Figure 46B and Table 26). This finding suggests that R^{19} has one or more additional interaction partner, "X," which explains the increased importance of R^{19} for receptor activity. Stimulation of the $D^{6.59}$A receptor with A^{19}PrRP20 resulted in a 0.16-fold elevated EC_{50}-value, compared to PrRP20 stimulation. This non-additive effect of the double cycle

mutagenesis experiment implies that the effects of the individual replacements are not independent of each other. Among more complicated mechanisms, such as indirect interactions of the two residues, the effect may also be due to a direct interaction between $D^{6.59}$ of PrRPR and R^{19} of PrRP20 (Figure 46C and Table 26).

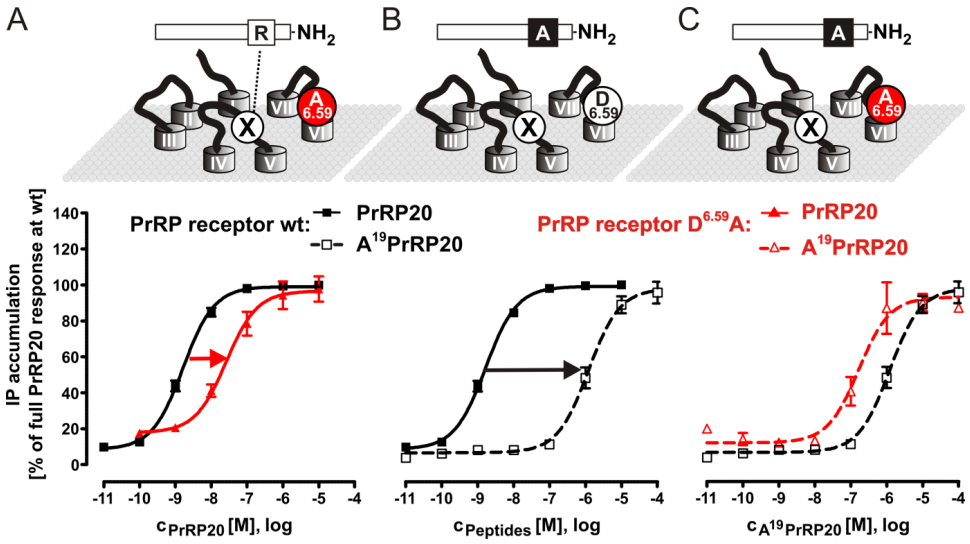

Figure 46: Functional characterization of PrRP receptor mutant $D^{6.59}A$ with PrRP20 and the modified ligand A^{19}PrRP20
Schemes represent the postulated mode of ligand binding. Due to the different relevance of $D^{6.59}$ and the R^{19}, a second contact point for R^{19} can be assumed. Complementary mutagenesis approach was used in combination with the signal transduction assay on cells, expressing the wt PrRPR or the $D^{6.59}A$ mutant in order to observe concentration-response curves. Data represent the mean ± s.e.m. of multiple independent experiments (n = 32 for hPrRPR with PrRP20, n = 12 for $D^{6.59}A$ PrRPR with PrRP20, n = 11 for hPrRPR with A^{19}PrRP20, and n = 3 for $D^{6.59}A$ PrRPR with A^{19}PrRP20). Receptor activity is expressed as percentage of full PrRP20 response at the wt PrRP receptor. A) Modification of receptor side: $D^{6.59}A$ PrRPR in comparison with wt receptor was stimulated with PrRP20. B) Exploring the ligand side: both PrRP20 and A^{19}PrRP20 were investigated using wt PrRPR. C) Complementary approach: A^{19}PrRP20 stimulation of wt and mutant receptor resulted almost matching concentration-response curves, indicating an interaction between $D^{6.59}$ of the receptor and R^{19} of the ligand.

Reciprocal mutagenesis leads to a constitutive active receptor mutant

To confirm the direct interaction between R^{19} and $D^{6.59}$, the corresponding residues were swapped (Figure 47A). The herein performed reciprocal mutagenesis

approach assumes that a lost interaction between two residues induced by single mutation to the counter amino acid can partly be recovered by a second mutation that establishes the interaction in a reverse manner. We used this method to verify the salt bridge between $D^{6.59}$ and R^{19} in the PrRP/PrRPR system by using the single peptide D^{19}PrRP20 and the $D^{6.59}$R receptor mutant (Figure 47C). The single peptide mutant D^{19}PrRP20 shows a similar effect as A^{19}PrRP20, with an increased EC_{50}-value of 1318 nM (Table 26) without impact on the efficacy (Figure 47B). We conclude that all peptide-receptor interactions that involve position R^{19} have been disrupted (Figure 46B and Figure 47B). In the reverse experiment, PrRP20 barely stimulated the $D^{6.59}$R receptor mutant with no determinable EC_{50}-value (Figure 47C). In comparison to both single mutant experiments, the activation of the $D^{6.59}$R but also $D^{6.59}$K mutant with D^{19}PrRP20 revealed a gain of function (EC_{50}-values: $D^{6.59}$R = 138 nM and $D^{6.59}$K = 115 nM, Table 26, Figure 47C, and Figure 48C), confirming the direct interaction of R^{19} and $D^{6.59}$. At the same time, the experiment provides further evidence in support of a second interaction site "X" for $D^{6.59}$R, as the EC_{50}-value is still elevated by a factor of 84 compared to the wt interaction.

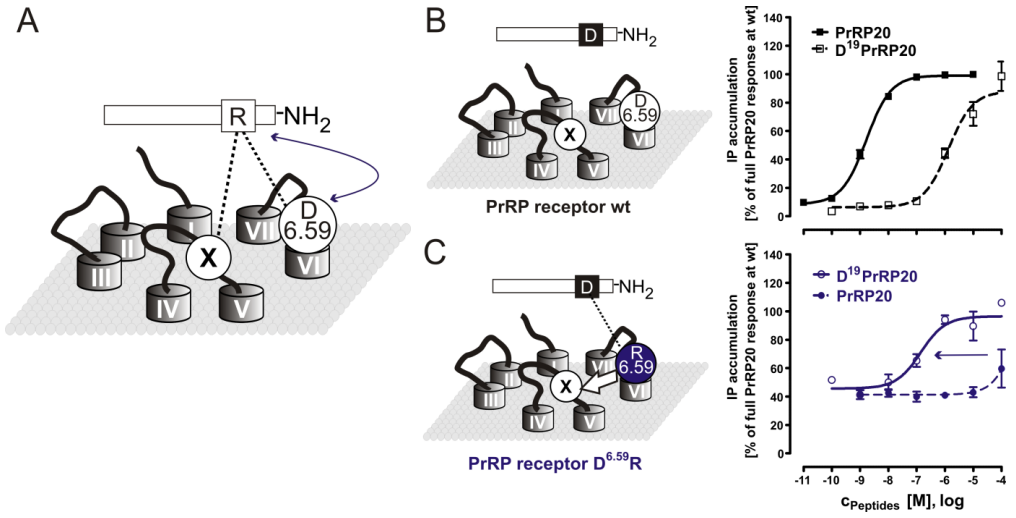

Figure 47: Reciprocal mutagenesis of the PrRPR

A) This scheme displays the assumed wt situation with the direct interaction of ligand R^{19}PrRP20 and receptor $D^{6.59}$PrRPR, as well as the second unknown interaction of the R^{19} to the receptor. B) The stimulation of wt receptor by D^{19}PrRP20 and the corresponding concentration-response curves of the signal transduction assay. C) Reciprocal mutagenesis scheme is shown with related concentration-response curves. Interestingly, $D^{6.59}$R mutant is partially basally active and can be activated by D^{19}PrRP20. The latter is due to the established D-R interaction. IP accumulation presented in Panels B and C represent the mean ± s.e.m. of multiple independent experiments (n = 32 for hPrRPR with PrRP20, n = 5 for $D^{6.59}$R PrRPR with PrRP20, n = 4 for hPrRPR with D^{19}PrRP20, and n = 3 for $D^{6.59}$R PrRPR with D^{19}PrRP20). Receptor activity is expressed as percentage of full PrRP20 response at the wt PrRP receptor.

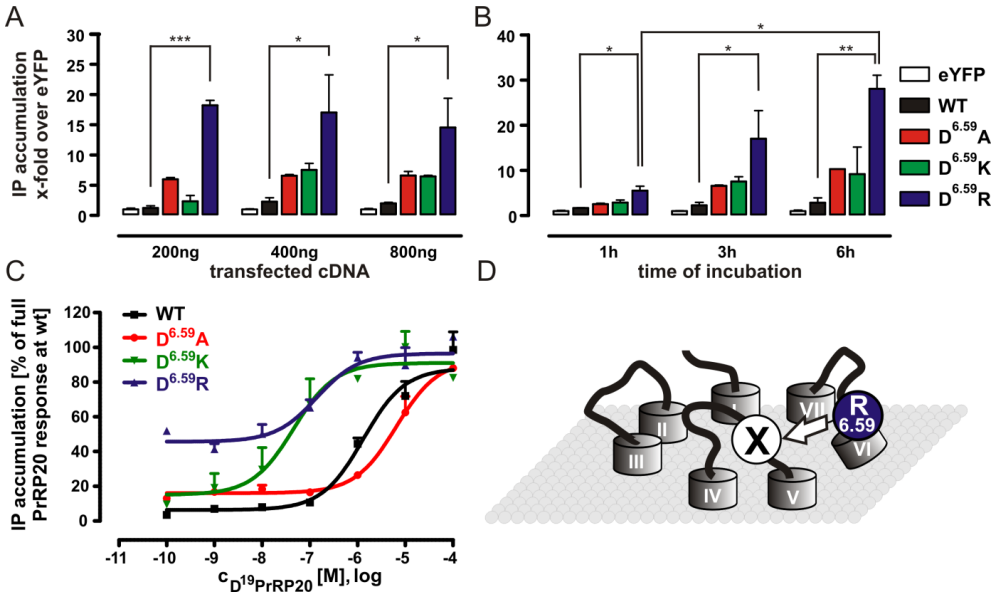

Figure 48: Investigation of the constitutive activity of $D^{6.59}R$ PrRPR mutant
A) Test of influence of transfection upon constitutive activity of wt PrRPR and $D^{6.59}$ constructs. The IP accumulation of differently transient transfected COS-7 cells expressing the various PrRPR mutants was measured without any agonist after three hours [given as x-fold over eYFP expressing cells]. [Each bar represents the mean ± s.e.m. of two different experiments; at least in triplicates; * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$] B) Constitutive activity of wt PrRPR and $D^{6.59}$ mutant was investigated in a time-dependent manner. The IP accumulation of COS-7 cells expressing the different PrRPR variants was measured without any agonist after different time periods [given as x-fold over eYFP expressing cells]. C) Concentration-response curves of $D^{6.59}$ PrRP receptor munats. Data represent the mean ± s.e.m. of multiple independent experiments (n = 5 for hPrRPR, n = 4 for $D^{6.59}$A PrRPR, n = 3 for $D^{6.59}$R PrRPR, and n = 2 for $D^{6.59}$K PrRPR). Receptor activity is expressed as percentage of full PrRP20 response at the wt PrRP receptor. D) Scheme of assumed explanation for the agonist-independent activity of the $D^{6.59}$R receptor mutant: We postulate that the $D^{6.59}$R is a CAM because $D^{6.59}$R mimics R^{19} of PrRP20 by intra-molecular interaction with a receptor region "X," inducing a partially active receptor conformation.

A novel possibility to identify the missing interaction site "X" arose because the $D^{6.59}$R receptor mutant presented a strongly increased basal activity, which is indicated by curves with higher initial IP accumulation (Figure 47C and Figure 48C). In contrast, $D^{6.59}$A and $D^{6.59}$K reveal solely slight elevated basal activity. This can be explained by more loosened constraints at this position and thus making it more susceptible for induced basal activity, whereas for $D^{6.59}$K the spatial and more charged prerequisites are missing. The observed effect of constitutive activity is independent of transient

transfection, which is a critical component. Different amounts of transfected DNA resulted in essentially similar cellular responses (Figure 46A). Finally, the constitutive activity of the $D^{6.59}R$ receptor mutant was confirmed by an increased time-dependent IP-accumulation compared to wt (Figure 46B; 1h, 3h = $P < 0.05$; 6h = $P < 0.01$). All investigated receptors were correctly integrated in the membrane as confirmed by live-cell microscopy (Figure 44A) and revealed similar cell surface levels as determined by surface ELISA (Figure 44B and Figure 44C).

Identification of "X" by modelling-guided double mutant analysis.

We hypothesize that $D^{6.59}R$ PrRPR is a CAM caused by the interaction of $D^{6.59}R$ with residue "X." $D^{6.59}R$ mimics R^{19} of PrRP20, inducing a partially active receptor conformation (Figure 48D). We further hypothesize that $D^{6.59}R/X^{X.X}A$ double mutants will lose constitutive activity and most importantly, retain activation by D^{19}PrRP20. In order to determine likely positions for "X," a comparative model of the PrRPR was constructed using the Rosetta molecular modeling software suite. Details of the modeling protocol are given in the *Materials and methods*. According to the lowest-energy model based on the semi-active opsin structure (PDBID: 3DQB (33)). $E^{5.26}$, $W^{5.28}$, $Y^{5.38}$, $F^{6.54}$, and $Q^{7.35}$ were found proximal to $D^{6.59}$ and were proposed to be potential interaction partners for $D^{6.59}R$ (Figure 49A) or for R^{19}PrRP20 when testing the wt receptor. The more distant residues, $Y^{2.64}$ and $W^{2.71}$, were chosen for control experiments.

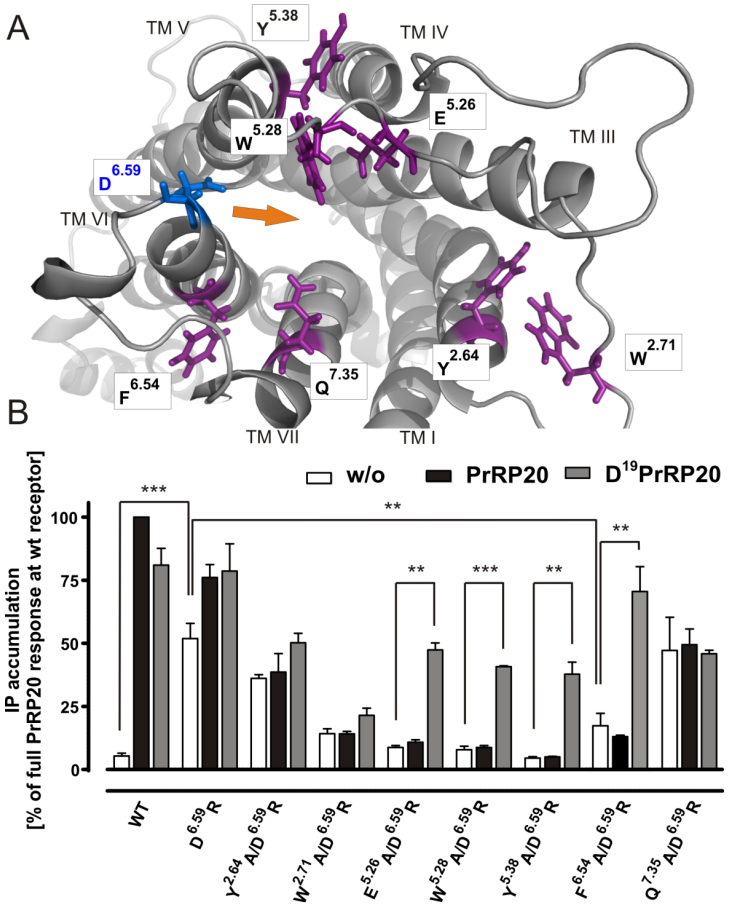

Figure 49: Molecular model of the PrRPR based on 3DQB and resulting double mutations based on the $D^{6.59}R$ PrRPR construct

A) Residues in proximity to the extracellular side are shown in purple. These were investigated in double mutational analysis with $D^{6.59}R$ PrRPR. The $D^{6.59}$ on top of TMH4 is colored in blue, and the suggested inward movement of the extracellular helical part of TMH6 is indicated by an orange dart. (B) A new approach to identify the missing interaction site, "X," arose by insertion of a second alanine substitution of assumed interacting residues to the $D^{6.59}R$ PrRPR. The second mutation is expected to diminish the basal activity but retain the capability to be activated by D^{19}PrRP20. IP accumulation assay of COS-7 cells transfected with eYFP as control and the following constructs of PrRPR: wt, $D^{6.59}R$, $Y^{2.64}A/D^{6.59}R$, $W^{2.71}A/D^{6.59}R$, $E^{5.26}A/D^{6.59}R$, $W^{5.28}A/D^{6.59}R$, $Y^{5.38}A/D^{6.59}R$, $F^{6.54}A/D^{6.59}R$, $Q^{7.35}A/D^{6.59}R$, respectively. Incubation was performed for three hours without ligand, PrRP20 or D^{19}PrRP20, and results are presented in IP accumulation as percentage of full PrRP20 response at the wt PrRP receptor. [Each bar represents the mean ± s.e.m. of at least duplicates of four different experiments; ** $P < 0.01$; *** $P < 0.001$].

With guidance from the receptor modeling data (Figure 49A), we generated and tested the double mutants $Y^{2.64}A/D^{6.59}R$, $W^{2.71}A/D^{6.59}R$, $E^{5.26}A/D^{6.59}R$, $W^{5.28}A/D^{6.59}R$ $Y^{5.38}A/D^{6.59}R$, $F^{6.54}A/D^{6.59}R$, and $Q^{7.35}A/D^{6.59}R$ of PrRPR. Interestingly, $E^{5.26}A/D^{6.59}R$, $W^{5.28}A/D^{6.59}R$, and $Y^{5.38}A/D^{6.59}R$ receptor mutants completely lost their constitutive activity in a ligand-independent signal transduction assay (Figure 49B). The IP accumulation after three hours of these unstimulated receptors dropped to a PrRPR wt level. The $F^{6.54}A/D^{6.59}R$ dropped as well but remained partially constitutively active (Figure 49B). These effects could be due to disruption of the hypothesized interaction to the $R^{6.59}$ residue or to decisive structural alterations, resulting in generally non-functional mutants. The latter situation was excluded after activation of these constructs using 10 μM D^{19}PrRP20 as an agonist (Figure 49B; $P<0.01$). In concentration-response experiments the EC_{50}-values were determined to be higher than 100 μM (Figure 50A). The fact that D^{19}PrRP20, not wt PrRP20, was able to activate these constructs re-emphasizes the direct interaction of D^{19} with $D^{6.59}R$.

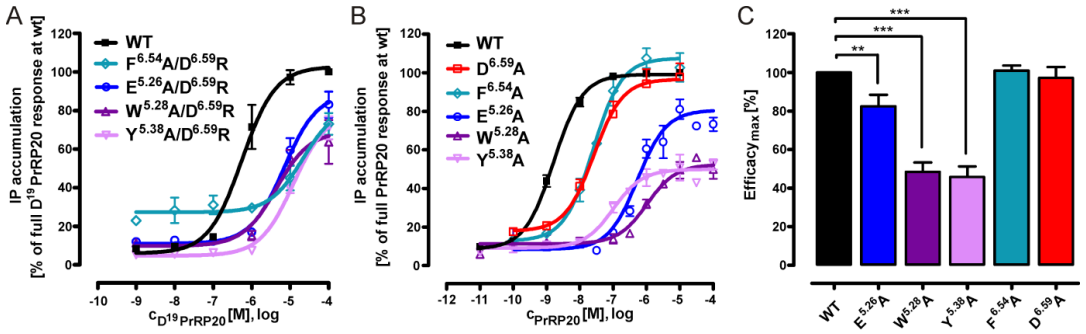

Figure 50: Functional characterization of PrRPR mutants with impact on receptor activation and ligand binding

A) COS-7 cells transfected with wt PrRPR or $E^{5.26}A/D^{6.59}R$, $W^{5.28}A/D^{6.59}R$, $Y^{5.38}A/D^{6.59}R$, $F^{6.54}A/D^{6.59}R$ receptor mutants, were stimulated for three hours with different D^{19}PrRP20 concentrations using a signal transduction assay. Data represent the mean ± s.e.m. of **5** (PrRPR), **3** ($E^{5.26}A/D^{6.59}R$, $W^{5.28}A/D^{6.59}R$, $Y^{5.38}A/D^{6.59}R$) or **2** ($F^{6.54}A/D^{6.59}R$) independent experiments, measured in duplicate. B) COS-7 cells transfected with wt (n = 32) and $E^{5.26}A$ (n = 8), $W^{5.28}A$ (n = 7), $Y^{5.38}A$ (n = 5), $D^{6.59}A$ (n = 12), and $F^{6.54}A$ (n =3) PrRPR mutants, respectively, were investigated in signal transduction assay, and data are presented in concentration-response curves as percentage of full PrRP20 response at wt PrRP receptor. Stimulation was performed for 1 hour. The height of the curves correlates with the efficacy of the mutants. Potency is given by the degree of shift to the right and its resulting EC_{50} value. C) COS-7 cells transfected with the mentioned constructs in Panel B were incubated for one hour in a signal transduction assay with 1 x 10^{-5}M (mutants) or 1 x 10^{-7}M (wt) PrRP20, and without stimulus. Results are expressed as percentage of IP accumulation compared to the PrRPR, with lowest mean of value being 0% and highest 100%. [bars represent the mean ± s.e.m of duplicates of at least 3 different experiments; * $P < 0.05$; *** $P < 0.001$].

Other double mutants, such as $Y^{2.64}A/D^{6.59}R$ or $Q^{7.35}A/D^{6.59}R$, showed slightly reduced constitutive activity but seem to be trapped in that state, as no further activation/stimulation was achieved. $W^{2.71}A/D^{6.59}R$ appears to have structural restrictions because no significant receptor activation could be observed. From the plethora of residues in the upper TMHs and ELs of PrRPR, which may interact with $D^{6.59}R$ the initial comparative models and mutational studies clearly suggested seven residues to potentially interact with $D^{6.59}R$. Of these seven potential interaction sites, we hypothesize $E^{5.26}$, $W^{5.28}$, $Y^{5.38}$, and $F^{6.54}$ to be engaged in $D^{6.59}R$-induced basal activity. Therefore, we postulate the latter residues to be involved in ligand binding and/or receptor activation. The combination of mutagenesis and comparative modelling enabled us to extract three

residues of relevance from the plethora of residues in the upper transmembrane helices (TMHs) and extracellular loops (ELs) of the PrRPR.

Confirmation of binding and activation site using single mutants.

To clarify the exact impact of the identified positions $E^{5.26}$, $W^{5.28}$, $Y^{5.38}$, and $F^{6.54}$, single alanine mutants at these positions were generated. Signal transduction studies of the single alanine mutants $E^{5.26}A$ (331-fold over wt), $W^{5.28}A$ (580-fold over wt), $Y^{5.38}A$ (61-fold over wt), and $F^{6.54}A$ (15-fold over wt) confirm the impact of residues $E^{5.26}$, $W^{5.28}$, $Y^{5.38}$, and $F^{6.54}$ on ligand binding (Table 28 and Figure 50B). Their distribution in EL2 and TMH5 suggests that this region plays a significant role in ligand binding. Therefore, EL2 and TMH5 were studied systematically to identify additional interaction sites that might have been missed due to inaccuracies of the comparative model. All charged (R, K, E, D) and aromatic (W, F, Y) residues between positions 4.65 and 5.40 were substituted to alanine (Table 28). None of the tested mutants resulted in significantly increased EC_{50}-values (Table 28 and Figure 50B). This demonstrates that the model-guided intramolecular mutagenesis experiment, at least in this setting, was more effective than alanine scanning in selecting the critical interaction partners.

Table 28: Signal transduction of the selected alanine of PrRP receptor mutants from extracellular loop 2 and top TMH5

IP accumulating signal transduction assay was performed for 1 hour with different concentrations of modified PrRP20 peptides to determine EC_{50}-values from concentration-response curves.

PrRPR mutant	E_{max} ± SEM [%] [a]	*P* [b]	pEC_{50} ± SEM[c]	EC_{50} [nM][c]	EC_{50}-ratio (mut/wt)[d]	N
Wt	100	-	8.78 ± 0.04	1.66	1	32
$Y^{4.65}A$	63 ± 22	ns	8.03 ± 0.32	9.3	6	2
$E^{4.68}A$	93 ± 8	ns	8.19 ± 0.19	6.4	4	3
$K^{4.70}A$	111 ± 35	ns	8.41 ± 0.41	3.9	2	2
$D^{4.73}A$	146 ± 41	ns	8.75 ± 0.49	1.78	1	2
$R^{4.75}A$	87 ± 15	ns	8.32 ± 0.37	4.8	3	3
$E^{5.25}A$	124 ± 10	ns	7.99 ± 0.13	10	6	3
$E^{5.26}A$	81 ± 5	0.0094	6.26 ± 0.10	549	331	8
$F^{5.27}A$	122 ± 50	ns	8.14 ± 0.49	7.2	4	2
$W^{5.28}A$	48 ± 5	< 0.0001	6.02 ± 0.14	954	580	7
$E^{5.32}A$	114 ± 11	ns	8.62 ± 0.14	2.4	1	2
$R^{5.33}A$	115 ± 15	ns	8.57 ± 0.20	2.7	2	2
$R^{5.35}A$	81 ± 4	0.0122	8.35 ± 0.32	4.5	3	2
$Y^{5.38}A$	46 ± 6	< 0.0001	6.99 ± 0.14	102	61	5
$W^{5.40}A$	101 ± 38	ns	8.78 ± 0.49	1.7	1	2
$D^{6.59}A$	97 ± 6	ns	7.59 ± 0.15	26	15	12
$F^{6.54}A$	101 ± 3	ns	7.61 ± 0.10	25	15	3

N represents the number of independent experiments.

[a] Efficacy was determined as percentage compared to full PrRP20 response at wt.

[b] Significance P was estimated using the unpaired t-test (ns represents no significantly different means with $P \geq 0.05$).

[c] EC_{50}-/pEC_{50}-values were calculated from the mean ± s.e.m. of N independent experiments, measured in duplicate.

[d] The ratio was determined using the Prism 5.03 function of dose-response EC_{50} shift determination by global fitting.

To verify the obtained results of potency of the PrRP wt receptor and its mutants, the cellular expression levels in the plasma membrane were investigated, because recently a constitutive internalization of the PrRP receptor has been reported (408). Binding studies of transiently transfected COS-7 cells revealed a sufficient number of surface wt receptors per cell (~95,000), calculated from the obtained B_{max}-value (445

Bq), the specific activity (3.52 x 10^{15} Bq/mol) and cell number (6.6 x 10^{5}). All PrRP receptor constructs with impact on potency were shown to be surface exposed and quantified by surface ELISA (Figure 44). The deviation from the wt PrRPR surface expression levels (wt = 39.6 ± 1.1%) varies from 16.3% ($W^{5.28}A$) to 59.6% ($F^{6.54}A/D^{6.59}A$).

However, these differences, basically resulting from transient transfection, reveal minor effects in the IP accumulation signaling assay set up, as the receptor mutant $F^{6.54}A$ (20.9 ± 3.7%) shows reduced total surface expression levels (Figure 44B) but full wt like efficacy (Figure 49B/C). Additionally, all PrRPR mutants are properly exported to the cell surface in comparable amounts as the wt receptor (39.6%, Figure 44C). Therefore, the herein obtained results of potency of agonists at their receptor constructs do not result from altered expression or export levels.

A reduced efficacy was observed in the concentration-response dependent signal transduction assay for $W^{5.28}A$ and $Y^{5.38}A$ ($P<0.001$) and – with decreased impact – also for $E^{5.26}A$ ($P<0.0094$, Figure 50C and Table 28). In summary, our findings support a binding mechanism in which $E^{5.26}$, in addition to $D^{6.59}$, directly engage R^{19} of PrRP20 through ionic interactions. $F^{6.54}$ might contribute to the overall global conformation of the binding pocket and positioning of TMH 6, as its single mutation is less invasive but still is in distance for direct ligand interactions. We further suggest that $W^{5.28}$ and $Y^{5.38}$ are possibly in direct contact with the ligand and are indeed critical for receptor activation and the transmission of an external signal into the cell.

Exploration of second interaction partner and dual binding mode at R^{19}.

We generated the $E^{5.26}A/D^{6.59}A$ double mutant of the receptor, which lacks both putative binding partners to the R^{19}. In addition, the reciprocal PrRPR mutants, $E^{5.26}R/D^{6.59}R$ and $E^{5.26}R$, were generated to test the interaction by swapping the putative binding residues. The $E^{5.26}A$ and the $E^{5.26}A/D^{6.59}A$ receptor mutants were investigated in a double cycle mutagenesis study, where they were stimulated with A^{19}PrRP20 and wt PrRP20 (Table 26 and Figure 51A). The $E^{5.26}A$ mutant stimulated with A^{19}PrRP20 resulted in a strongly increased EC_{50}-value higher than 10 μM, 21-fold shifted compared to PrRP20 stimulation (537 nM). The enhanced EC_{50}-value can be explained by the disruption of the second R^{19} interaction to receptor residue $D^{6.59}$. Indeed, this effect agrees with a similar impact of the $D^{6.59}A$ mutation (15-fold shifted; Table 26), which also diminished the direct interaction to the R^{19} of the ligand to a similar extent (Figure 46A and Figure 51A). Furthermore, the stimulation of the $E^{5.26}A/D^{6.59}A$ receptor mutant with either PrPR20 or A^{19}PrRP20 resulted in matching curves. As no additional loss in potency was observed compared to the $E^{5.26}A$ mutant tested with A^{19}PrRP20 (Figure 51B), the experiment provides evidence that $E^{5.26}$ is involved in binding to R^{19}.

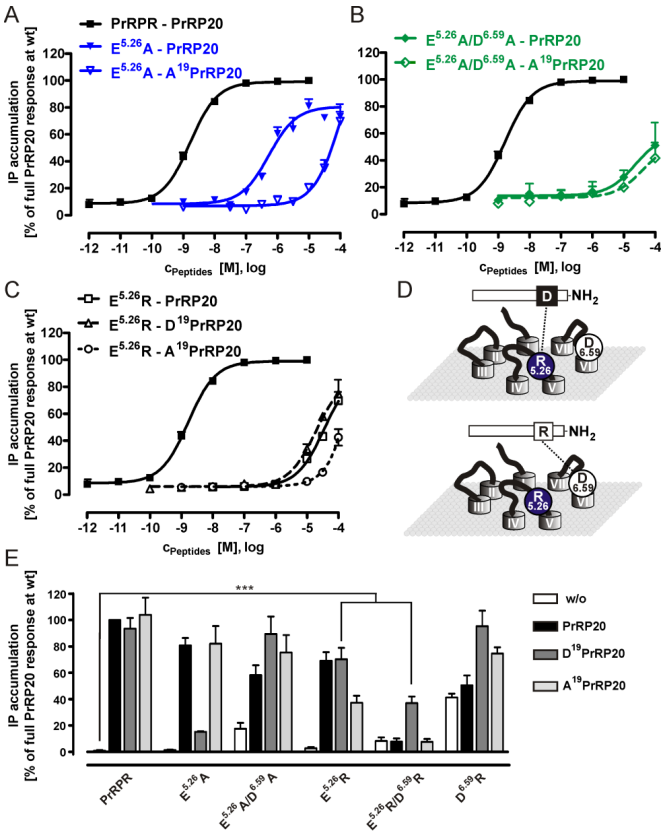

Figure 51: Stimulation analysis of $E^{5.26}$ mutants reveals a preferential activation of R mutants by the reciprocal ligand D^{19}PrRP20

Functional investigation of PrRPR mutants $E^{5.26}$A, $E^{5.26}$R, and $E^{5.26}$A/$D^{6.59}$A with the ligands PrRP20, A^{19}PrRP20, or D^{19}PrRP20. The signal transduction assay was performed in COS-7 cells expressing the wt PrRPR or $E^{5.26}$A, $E^{5.26}$R, or $E^{5.26}$A/$D^{6.59}$A mutants to observe concentration-response curves. Results of two independent experiments, each performed in duplicate, are presented as mean ± s.e.m of duplicates. A) $E^{5.26}$A PrRPR was stimulated with both PrRP20 and A^{19}PrRP20 and demonstrated an equipotent loss in potency compared to the $D^{6.59}$A PrRPR mutation (Figure 46A). Additionally, this panel highlights the direct interaction between R^{19} and $D^{6.59}$. B) Stimulation with of the $E^{5.26}$A/$D^{6.59}$A receptor with A^{19}PrRP20 or PrRP20 revealed no further loss in potency and a slightly decreased efficacy compared to the $E^{5.26}$A PrRPR. This indicates that $E^{5.26}$ might be the second binding partner of R^{19}. C) Functional characterization of the reciprocal $E^{5.26}$R PrRPR mutant using R^{19}-modified PrRP20 analogues. D) The scheme shows the assumed interplay of attraction and repulsion for the reciprocal interaction of the ligands R^{19}PrRP20 and D^{19}PrRP20 with the $E^{5.26}$R PrRP receptor mutant from Panel C. E) IP accumulation assay of COS-7 cells transfected with eYFP as control and the following constructs of PrRPR: wt, $E^{5.26}$A, $E^{5.26}$A/$D^{6.59}$A, $E^{5.26}$R, $E^{5.26}$R/$D^{6.59}$R, $D^{6.59}$R, respectively. Incubation was performed for one hour using 100 µM of PrRP20, D^{19}PrRP20, A^{19}PrRP20, and without ligand. [Each bar represents the mean ± s.e.m. of at least duplicates of 2 different experiments; *** $P < 0.001$].

Next, the capability of receptor mutants $E^{5.26}A$, $E^{5.26}A/D^{6.59}A$, $E^{5.26}R$, $E^{5.26}R/D^{6.59}R$, $D^{6.59}A$, and wt PrRPR to transmit signalling was tested (Figure 51E). Importantly, the reciprocal receptor mutants $E^{5.26}R$ and $E^{5.26}R/D^{6.59}R$ were significantly and best activated by D^{19}PrRP20 (both: $P<0.001$). In fact, $E^{5.26}R/D^{6.59}R$ was solely activated by D^{19}PrRP20. Finally, the $E^{5.26}R$ mutant was stimulated with PrRP20, A^{19}PrRP20, and D^{19}PrRP20 in a concentration-response experiment (Figure 51C). This receptor mutant behaved similarly, when stimulated by PrRP20 and D^{19}PrRP20 (both: EC_{50}-value >10 μM). Along with the experiments testing D^{19}PrRP20 stimulation of wt PrRPR, we demonstrate an approximately equal repulsive effect of R^{19} to $E^{5.26}R$ or D^{19} to $D^{6.59}$ (Figure 51D). This strengthens our hypothesis of a dual binding mode of R^{19} to $E^{5.26}$ and $D^{6.59}$.

Comparative model of PrRP/receptor complex provides structural information on mode of binding.

The $R^{19}/E^{5.26}$ and $R^{19}/D^{6.59}$ contacts as restraints, a *de novo*-folded model of PrRP8-20 based on reported NMR data (35) was docked into an ensemble of comparative models of the PrRPR. The conformation of the EL regions was constructed simultaneously with ligand docking to accurately capture conformational changes induced by the peptide. Details of the modeling procedures are given in the *Materials and methods* and Appendix C. The lowest-energy Rosetta model features salt bridges between $D^{6.59}$, $E^{5.26}$, and R^{19}. $W^{5.28}$ and $Y^{5.38}$ form π-stacking interactions that may be indicative of a "toggle-switch" mechanism (Figure 52A) (409). $F^{6.54}$ appears to further apart from R^{19} but might contribute to the positioning of TMH 6 via intra-molecular interactions and is

in distance for π-stacking interactions with the F^{20} of PrRP20. Additional interactions between peptide and receptor hold the peptide in an optimal binding conformation deeply buried in the upper TMH segments and supported by the ELs from above.

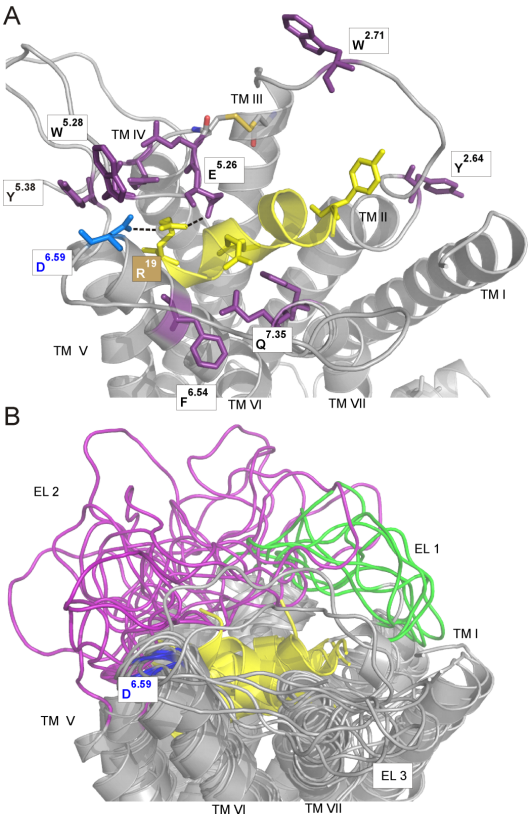

Figure 52: Comparative model of PrRPR docked to the thirteen C-terminal residues of PrRP20

A) Selected comparative model generated by Rosetta in the presence of the PrRP ligand to support experimental data. The figure displays an ensemble of low-energy PrRP/receptor models generated in Rosetta, that agrees well with experimental data. Residue $D^{6.59}$ is colored in blue, the peptide is presented in yellow, and residues in vicinity to PrRP are in purple. B) The eight non-redundant low-energy comparative models of the PrRP/receptor complex. These eight models were generated in the presence of structural constraints derived from the mutagenesis data described (see main text) and are considered energetically favorable according to the Rosetta v3 all-atom scoring function. The peptide is highlighted in yellow, $D^{6.59}$ of the receptor in blue, EL1 of the receptor in green, and EL2 of the receptor in magenta.

Discussion

We have evolved a strategy to interrogate detailed molecular mechanisms of GPCR activation by combining reciprocal, double cycle, and intramolecular double mutagenesis with computational modelling. We apply this technique effectively to PrRPR and its CAM, $D^{6.59}$R PrRPR, identifying distinct receptor residues involved in activation and/or ligand binding.

This is the first comprehensive mutational study of the extracellular and transmembrane regions of the PrRPR. The double cycle mutagenic approach suggests the interaction (direct or indirect) between residues $D^{6.59}$ and R^{19} and provides a first anchor point for receptor/ligand investigations. Interacting residues can be characterized by reciprocal mutagenesis, as shown before in an intramolecular study with the $D^{2.61}R/R^{7.39}D$ swap in the gastrin-releasing peptide receptor (410) or the $D^{6.44}/N^{7.49}$ residues of the thyrotropin (TSH) receptor (411). By applying this method to the PrRP/PrRPR system, the salt bridge of $D^{6.59}$ to R^{19} was verified, and more importantly, by generating the $D^{6.59}$R receptor, we identified the first CAM of the PrRPR. Up to now, numerous CAM were generated and investigated in a plethora of previous studies, emphasizing the increasing importance of CAMs. For example, CAM of the human angiotensin II type 1 receptor with $N^{3.35}$Gly (412), the $ß_{1B}$ (413)/ $ß_2$-adrenergic receptor (414, 415), the cannabinoid receptor 1 (416), muscarinic m_1 (417) and m_5 receptors (418), among others, have been found. Interestingly, more than sixty naturally occurring CAM GPCR are known so far (419) and are often related to human disorders (420). Consequently, GPCR activated in an agonist-independent manner are of emerging importance for drug development(388).

CAM more readily undergo transition between active and inactive conformations due to removed conformational constraints of the inactive form (421). Because $D^{6.59}R$ in PrRPR is located at the top of TMH6, we hypothesize that this helix is involved in receptor activation via an inward movement of the upper helical region (Figure 48D). Similarly to the PrRPR $D^{6.59}R$ CAM, mutant-induced receptor activity was observed in the $S^{6.58}Y/T^{6.59}P$ double mutant of m_5 muscarinic receptors (422). These data indicate that the top of TMH6 is directly involved in the switch between the active and the inactive state of several GPCR and that the interaction with the ligand stabilizes the receptor in this active conformation – a notion that supports the "global toggle switch model" (377, 423). This model suggests that activation results from an inward movement of the extracellular ends of TMHs 6 and 7 toward TMH3, concomitant with a movement of the intracellular part of the TMHs in the opposite direction, which enables signaling via G-protein coupling. PrRPR represents an excellent model system to further investigate this hypothesis and gain insights to receptor activating mechanisms.

Previous work on the TSH receptors showed the effects of spatially distant double mutants on constitutive activity (424, 425). However, we focus on the investigation of the molecular vicinity surrounding $D^{6.59}$, as we suggest that specific inter-residue interactions of the generated CAM occur. To take advantage of the $D^{6.59}R$ CAM to elucidate the mechanism of ligand binding and PrRPR activation, we established an effective combination of intramolecular double and inter-molecular reciprocal mutagenic approaches to study PrRPR activation by wt PrRP20, A^{19}PrRP20, and D^{19}PrRP20. With guidance from the PrRPR comparative model, seven possible interacting residues were considered (Figure 49A), and the double mutants $E^{5.26}A/D^{6.59}R$, $W^{5.28}A/D^{6.59}R$,

$Y^{5.38}A/D^{6.59}R$, and $F^{6.54}A/D^{6.59}R$ revealed an involvement of these residues in receptor activation. Importantly, these receptor mutants were significantly activated by D^{19}PrRP20 but not by wt PrRP20 (Figure 49B), proving that the receptor mutants were not miss-folded and that D^{19} on the ligand is still able to interact with $D^{6.59}R$. CAM are thought to mimic, at least partially, the active conformation of the wt receptor and to spontaneously adopt a structure able to activate G-proteins (426). Therefore, we hypothesize that in D^{19}PrRP20, residue D^{19} takes over the role of the destroyed intra-molecular interaction of the double mutants, reactivating the "silenced" CAM. The conformation of a basally silenced GPCR might impair its intrinsic capacity for signaling compared to the wt receptor. Notably, further mutations within EL2/TMH5 had no considerable impact on receptor potency, in contrast to all three positions identified via intramolecular interactions (Table 28). This demonstrates the precision and usefulness of the modeling-guided double mutational approach to identify interacting residues in close proximity to the ligand.

In contrast, the $W^{2.71}A/D^{6.59}R$ control turned out to be deficient in signaling. This is expected and in agreement with the high conservation of $W^{2.70}/W^{2.71}$ in most peptide GPCR, e.g. in the NPY receptor system (14). Furthermore, $W^{2.71}$ is located in the structurally relevant WxGF-motif, which is suggested to be a key component in the activation mechanism in many GPCR in the rhodopsin family (427). Recent investigations on TMH2 of the CAM $N^{3.35}G$ hAT1 suggested TMH2 to pivot, bringing the top of TMH2 closer to the binding pocket (428). Our results obtained for the conserved $Y^{2.64}$ on top of TMH2 do not support such a spatial approach to $D^{6.59}$ and thus

to the binding pocket. This reflects the divergence of GPCR activation and accentuates that the detailed mode of activation is not a common mechanism.

The results obtained from studies of the $E^{5.26}A$ mutation lead to the conclusion that this residue is predominantly responsible for ligand binding. Our initial double cycle mutagenic experiments at $D^{6.59}$ support a more complex double binding role for R^{19} of PrRP20, which appears to be in contact with two sites on PrRPR. Accordingly, we suggest $E^{5.26}$ to be the second binding partner for peptide residue R^{19} (Figure 51D). The extensive mutagenic studies of residue $E^{5.26}$ strongly indicate the participation in binding to R^{19} and the constitutive activity of $D^{6.59}R$ supports the hypothesis of a second R-specific interaction site in PrRPR that can be satisfied by the $D^{6.59}R$ but not the $D^{6.59}K$ mutant. A similar dual binding mode for arginine was recently reported for gonadotropin-releasing hormone (GnRH) receptor (212). This has been supported by other studies, where substitution of R^{19} to lysine, citruline (Cit), α-amino-4-guanidino-butyric acid (Agb), or α-amino-3-guanidino-propionic acid (Agp) on the peptide lead to reduced binding affinities (36). Interestingly, the tight ensemble of models that is in agreement with the experimental data presented herein exhibits variability in ELs 1 and 2 while still maintaining the contacts between $D^{6.59}$ and $E^{5.26}$ with R^{19}. Given this structural variability in our models, we emphasize that the presented approach is an iterative process, where initial models can be used to guide experimental design, and the resulting data allow for model refinement. The current PrRP/receptor model can only be considered valid in the light of the functional data. However, it provides insight into possible structural mechanisms of peptide/receptor interactions and receptor activation.

$W^{5.28}A$ and $Y^{5.38}A$ also showed lowered ligand potency, but both mutants revealed a strongly decreased ability to transmit signals compared to the wt receptor (Table 28). This effect may result from intramolecular structural alteration due to the lack of aromaticity at the $Y^{5.38}A$ site. Mutational studies reported for the nearby $Y^{5.39}$ residue in both cannabinoid receptors (CB_1 and CB_2) revealed that the aromaticity at this position is crucial (429). The PrRP/receptor model places $W^{5.28}$ in close proximity to $Y^{5.38}$ (Figure 52A). In this model, the residues form stacking interactions, but this remains to be proven experimentally. We speculate that, due to the effects observed for potency and efficacy, $W^{5.28}$ and $Y^{5.38}$ are related to receptor activation. In contrast, $F^{6.54}A$ mutant reveals full wt efficacy accompanied with reduced potency. From the docked modeling data, we speculate that this residue contributes to the correct conformation of the binding pocket and might interact with the F^{20} of the PrPR20.

Evolutionary and structural studies revealed that the PrRPR belongs to the family of RF-amide peptide receptors, consisting of five discovered groups: the neuropeptide FF (NPFF) group, the prolactin-releasing peptide (PrRP) group, the gonadotropin-inhibitory hormone (GnIH) group, the kisspeptin group, and the 26RFa group (430-432). However, further phylogenic investigations revealed that the PrPRR shares an ancient receptor with the NPY receptors (396). The human PrRPR possesses high sequence identity with the human NPY_2R, particularly in the upper and middle regions of TMH 4, TMH 5, and TMH 6. It is suggested that the PrRPR family began co-evolving with ancestral PrRP/C-RF-amide peptide with a redundant NPY binding receptor (396). This explains the importance of the conserved $D^{6.59}$ residue and in turn, might have been responsible for the development of a double binding mode for R^{19} in the PrRPR/PrRP system. It could be

speculated that other RF-amide receptors evolved similar binding modes for the crucial arginine within the RF-amide motif, especially for the closely related 26RF-amide receptor. In contrast, for the well investigated Y-receptor family, a double binding mode was not identified, neither for R^{33} at Y_2/Y_5R nor for R^{35} at Y_1/Y_4R (394, 395). However, the second interaction might occur via the second arginine 33 or 35, respectively.

Regarding medical and physiological implications, the expression of CAM can entail oncogenic effects, such as tumor formation in nude mice (433). A variety of diseases are known to be triggered by elevated basal activity, including autosomal dominant hypocalcaemia (434) and ovarian hyperstimulation syndrome (435). Our findings provide insight into the harmful potential of CAM and demonstrate the need for applicable drugs that are able to diminish mutation-induced receptor activity. We are confident that our technique is a promising tool to investigate residues relevant for ligand binding and receptor activation because a CAM is used as a template. Our approach paves the way for obtaining specific structure/function information on a molecular level, which is of indispensible value, as no crystal structure for a peptide GPCR currently exists. This method will hopefully contribute to the elucidation of the structural mechanisms of harmful CAM and help to develop and increase the number of inverse-agonist drugs that target these receptors.

Acknowledgement

The authors thank Kristin Löbner and Christina Dammann for their technical assistance in peptide synthesis, Janet Schwesinger for sequencing, and Regina Reppich-Sacher for recording mass spectra. They would also like to thank members of the

ROSETTACOMMONS, Elizabeth Dong, David Nannemann, Steven Combs, and Anette Kaiser for their assistance and insight provided concerning the molecular modelling.

APPENDIX B

PROTOCOL CAPTURE FOR CHAPTER II: THE ACTIVITY OF PROLACTIN RELEASING PEPTIDE CORRELATES WITH ITS HELICITY

This appendix contains the protocol capture for the modeling work published in (DeLuca*, Rathmann*, Beck-Sickinger, and Meiler, 2013), some of which is found in the manuscript's Supplemental Information. *These authors contributed equally.

Computational details

All models were generated by independent simulations using Vanderbilt University's Center for Structural Biology computing cluster and the university's Advanced Computing Center for Research and Education (ACCRE). Computations were performed on a combination of AMD Opteron and Intel Nehalem processor nodes. The time required to fold one model of the 13 C-terminal residues of PrRP20 was less than 10 seconds. The time required for a single round of high-resolution refinement of one model was less than 1 minute. All modeling was performed using Rosetta trunk revision 36905.

Input files

Before any modeling was performed, truncated peptide to residues 8 to 20 and renumbered 1-13.

FASTA file

```
>PrRP8-20 Sequence
WYASRGIRPVGRF
```

Chemical shift file for fragment generation (must end in .chsft)

```
# AA Res       C         CA         CB      HA          N
  W    1 9999.00    9999.00    9999.00    4.34    9999.00
  Y    2 9999.00    9999.00    9999.00    4.03    9999.00
  A    3 9999.00    9999.00    9999.00    4.14    9999.00
  S    4 9999.00    9999.00    9999.00    4.31    9999.00
  R    5 9999.00    9999.00    9999.00    4.18    9999.00
  G    6 9999.00    9999.00    9999.00    3.86    9999.00
  I    7 9999.00    9999.00    9999.00    4.12    9999.00
  R    8 9999.00    9999.00    9999.00    4.65    9999.00
  P    9 9999.00    9999.00    9999.00    4.46    9999.00
  V   10 9999.00    9999.00    9999.00    4.14    9999.00
  G   11 9999.00    9999.00    9999.00    3.94    9999.00
  R   12 9999.00    9999.00    9999.00    4.05    9999.00
  F   13 9999.00    9999.00    9999.00    4.60    9999.00
```

Constraints file for fragment generation with NOEs (must end in .cst)

```
# (NOTE:  No side-chain protons were taken into account)
NMR_v3.0
data set used in DUrsi et al PrRP820 strict NOE definitions
38
      1  H      2  H         5.00       0.00     medium
      2  H      3  H         5.00       0.00     medium
      3  H      4  H         5.00       0.00     medium
      4  H      5  H         5.00       0.00     medium
      6  H      7  H         5.00       0.00     medium
      7  H      8  H         5.00       0.00     medium
     10  H     11  H         3.00       0.00     strong
     11  H     12  H         3.00       0.00     strong
      1  HA     2  H         5.00       0.00     medium
      2  HA     3  H         5.00       0.00     medium
      3  HA     4  H         5.00       0.00     medium
      4  HA     5  H         5.00       0.00     medium
      5  HA     6  H         5.00       0.00     medium
      6  #HA    7  H         5.00       0.00     medium
      7  HA     8  H         5.00       0.00     medium
      9  HA    10  H         5.00       0.00     medium
     10  HA    11  H         3.00       0.00     strong
     11  #HA   12  H         5.00       0.00     medium
     12  HA    13  H         3.00       0.00     strong
      1  #HB    2  H         5.00       0.00     medium
      2  #HB    3  H         5.00       0.00     medium
      4  #HB    5  H         5.00       0.00     medium
      5  #HB    6  H         5.00       0.00     medium
      7  HB     8  H         3.00       0.00     strong
      9  #HB   10  H         5.00       0.00     medium
     10  HB    11  H         5.00       0.00     medium
     12  #HB   13  H         3.00       0.00     strong
      2  H      4  H         5.00       0.00     weak
      3  H      5  H         5.00       0.00     weak
```

```
4  H      6  H         5.00       0.00     weak
2  HA     4  H         5.00       0.00     weak
3  HA     5  H         5.00       0.00     weak
4  HA     6  H         5.00       0.00     weak
9  HA    11  H         5.00       0.00     weak
7  HA    10  H         5.00       0.00     weak
9  HA    12  H         5.00       0.00     weak
5  HA     8  #HB       5.00       0.00     weak
7  HA    10  HB        5.00       0.00     weak
```

Constraint file for folding

```
# (NOTE: For de novo folding, side-chains are not taken into account.
Therefore, any distance restraints between side-chain protons were changed to
CB and the upper bound (ub) was increased from 5Å to 7Å for weak restraints
and from 3Å to 5Å for strong restraints.)
# type   atom1  res1 atom2  res2   function      lb      ub     sd  comment comment
AtomPair     H     1     H     2    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H     2     H     3    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H     3     H     4    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H     4     H     5    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H     6     H     7    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H     7     H     8    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H    10     H    11    BOUNDED    0.00    3.00    1.0    NOE    strong
AtomPair     H    11     H    12    BOUNDED    0.00    3.00    1.0    NOE    strong
AtomPair    HA     1     H     2    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     2     H     3    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     3     H     4    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     4     H     5    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     5     H     6    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair   1HA     6     H     7    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     7     H     8    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     9     H    10    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA    10     H    11    BOUNDED    0.00    3.00    1.0    NOE    strong
AtomPair   1HA    11     H    12    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA    12     H    13    BOUNDED    0.00    3.00    1.0    NOE    strong
AtomPair    CB     1     H     2    BOUNDED    0.00    7.00    1.0    NOE    medium
AtomPair    CB     2     H     3    BOUNDED    0.00    7.00    1.0    NOE    medium
AtomPair    CB     4     H     5    BOUNDED    0.00    7.00    1.0    NOE    medium
AtomPair    CB     5     H     6    BOUNDED    0.00    7.00    1.0    NOE    medium
AtomPair    CB     7     H     8    BOUNDED    0.00    5.00    1.0    NOE    strong
AtomPair    CB     9     H    10    BOUNDED    0.00    7.00    1.0    NOE    medium
AtomPair    CB    10     H    11    BOUNDED    0.00    7.00    1.0    NOE    medium
AtomPair    CB    12     H    13    BOUNDED    0.00    5.00    1.0    NOE    strong
AtomPair     H     2     H     4    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair     H     3     H     5    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair     H     4     H     6    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     2     H     4    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     3     H     5    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     4     H     6    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     9     H    11    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     7     H    10    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     9     H    12    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     5    CB     8    BOUNDED    0.00    7.00    1.0    NOE    weak
AtomPair    HA     7    CB    10    BOUNDED    0.00    7.00    1.0    NOE    weak
```

Constraint file for full-atom refinement

```
# (NOTE: For full-atom refinement, side-chains are taken into account. Therefore, any distance restraints between side-chain protons were not altered, and the upper bound (ub) was 5Å for weak restraints and 3Å for strong restraints.)
# type   atom1  res1 atom2  res2   function      lb      ub     sd  comment comment
AtomPair    1H     1     H     2    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H     2     H     3    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H     3     H     4    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H     4     H     5    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H     6     H     7    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H     7     H     8    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair     H    10     H    11    BOUNDED    0.00    3.00    1.0    NOE    strong
AtomPair     H    11     H    12    BOUNDED    0.00    3.00    1.0    NOE    strong
AtomPair    HA     1     H     2    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     2     H     3    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     3     H     4    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     4     H     5    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     5     H     6    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair   1HA     6     H     7    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     7     H     8    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA     9     H    10    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA    10     H    11    BOUNDED    0.00    3.00    1.0    NOE    strong
AtomPair   1HA    11     H    12    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HA    12     H    13    BOUNDED    0.00    3.00    1.0    NOE    strong
AtomPair   1HB     1     H     2    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair   1HB     2     H     3    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair   1HB     4     H     5    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair   1HB     5     H     6    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HB     7     H     8    BOUNDED    0.00    3.00    1.0    NOE    strong
AtomPair   1HB     9     H    10    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair    HB    10     H    11    BOUNDED    0.00    5.00    1.0    NOE    medium
AtomPair   1HB    12     H    13    BOUNDED    0.00    3.00    1.0    NOE    strong
AtomPair     H     2     H     4    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair     H     3     H     5    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair     H     4     H     6    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     2     H     4    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     3     H     5    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     4     H     6    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     9     H    11    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     7     H    10    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     9     H    12    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     5   1HB     8    BOUNDED    0.00    5.00    1.0    NOE    weak
AtomPair    HA     7    HB    10    BOUNDED    0.00    5.00    1.0    NOE    weak
```

De novo *folding options file*

```
-abinitio
        -increase_cycles 2.5
        -rg_reweight 0.0
-fold_cst
        -force_minimize
-constraints
        -cst_file PrRP8-20.cst
        -cst_weight 1.0
```

```
        -viol
        -viol_level 101
-residues
        -patch_selectors CENTROID_HA
-in
        -path
                -database ./rosetta_database
        -file
                -frag3 aa3mers.txt
                -frag9 aa9mers.txt
                -fasta PrRP8-20.fasta
-out
        -nstruct 10000
        -prefix PrRP_
        -output
        -sf PrRP_fold.sc
        -file
                -silent PrRP.out
                -silent_struct_type binary
-overwrite
```

Full-atom refinement options file

```
-relax
        -sequence
-constraints
        -cst_fa_file PrRP_fa.cst
        -cst_fa_weight 1
        -viol
        -viol_level 101
-in
        -path
                -database rosetta_database
        -file
                -l pdb_list.txt
                -fullatom
                    -residue_type_set centroid
-out
        -output
        -nstruct 10
        -file
                -silent PrRP_fa.out
                -silent_struct_type binary
                -scorefile PrRP_fa.fasc
                -fullatom
-residues
        -patch_selectors CTERM_AMIDATION
-overwrite
```

Command lines

Fragment generation

```
rosetta/rosetta_fragments/make_fragments.pl -id PrRP_ -nosam PrRP_.fasta >&
make_fragments.log &
```

De novo *folding*

```
rosetta/rosetta_source/bin/AbinitioRelax.default.linuxgccrelease –database
rosetta/rosetta_database @fold.options
```

Full-atom refinement

```
mpiexec rosetta/rosetta_sourse/bin/relax.default.linuxgccrelease
@refinement.options
```

APPENDIX C

PROTOCOL CAPTURE FOR CHAPTER III: INTEGRATING SOLID STATE NMR AND COMPUTATIONAL MODELING TO INVESTIGATE THE STRUCTURE AND DYNAMICS OF MEMBRANE-ASSOCIATED GHRELIN

This appendix contains the protocol capture for the modeling work in the manuscript submitted to PLoS ONE of the same title by (Vortmeier*, DeLuca*, Cholet, Scheidt, Beck-Sickinger, Meiler, and Huster.) *These authors contributed equally. Further details are also available in Chapter III, and more detailed information on comparative modeling in Rosetta can be found in reference (252).

Computational details

All models were generated by independent simulations using Vanderbilt University's Center for Structural Biology computing cluster and the university's Advanced Computing Center for Research and Education (ACCRE). Computations were performed on a combination of AMD Opteron and Intel Nehalem processor nodes. All Rosetta-related protocols were conducted using Rosetta version 3.4.

Comparative modeling

FASTA file of GHSR1a

```
>gi|38455410|ref|NP_940799.1| growth hormone secretagogue receptor
MWNATPSEEPGFNLTLADLDWDASPGNDSLGDELLQLFPAPLLAGVTATCVALFVVGIAGNLLTMLVVSR
FRELRTTTNLYLSSMAFSDLLIFLCMPLDLVRLWQYRPWNFGDLLCKLFQFVSESCTYATVLTITALSVE
RYFAICFPLRAKVVVTKGRVKLVIFVIWAVAFCSAGPIFVLVGVEHENGTDPWDTNECRPTEFAVRSGLL
TVMVWVSSIFFFLPVFCLTVLYSLIGRKLWRRRRGDAVVGASLRDQNHKQTVKMLAVVVFAFILCWLPFH
VGRYLFSKSFEPGSLEIAQISQYCNLVSFVLFYLSAAINPILYNIMSKKYRVAVFRLLGFEPFSQRKLST
```

```
LKDESSRAWTESSINT
```

Transmembrane span prediction

```
# Used HMMTOP, TMHMM, JUFO9D, and OCTOPUS servers. Also ran Meiler lab's
YUFOPM (Jeff Mendenhall):
yufopm GHSR1.fasta
```

Threading of GHSR1a sequence on template structure

```
# See Chapter III for sequence alignment information.
/sb/meiler/scripts/sequence_util/thread_pdb_from_alignment.py --template
$TEMPLATE_NAME_FROM_ALIGNMENT --target $TARGET_NAME_FROM_ALIGNMENT --chain A -
-align_format clustal
```

Preparation for making fragments

```
make_fragments.pl –id GHSR1 –nofrags GHSR1.fasta
```

Generating fragments with Rosetta fragment picker

```
-in:file:fasta          GHSR1.fasta
-in:path:database       rosetta-3.4/rosetta_database
-in:file:vall     rosetta-3.4/rosetta_tools/fragment_tools/vall.jul19.2011.gz
-frags:n_candidates     1000
-frags:n_frags          200
-frags:frag_sizes       3 9
-out:file:frag_prefix   GHSR1_
-frags:scoring:config   GHSR1.cfg
-in:file:checkpoint     GHSR1.checkpoint
-frags:write_ca_coordinates
-frags:describe_fragments GHSR1_score
-frags:ss_pred GHSR1.psipred_ss2 psipred GHSR1.jufo_ss jufo GHSR1.rdb sam
```

Generation of lipophilicity file

```
rosetta_source/src/apps/public/membrane_abinitio/run_lips.pl <fasta file>
<span file> <path to blastpgp> <path to nr database> <path to alignblast.pl
script>
```

GHSR1 disulfide definition

```
116    198
```

GHSR1 spanfile

```
TM region consensus for Homo sapiens GHSR1a
7 366
antiparallel
n2c
  43    66    43    66
  80   100    80   100
 120   140   120   140
 162   181   162   181
 212   233   212   233
 263   282   263   282
 304   325   304   325
```

Fill in density by building loops

```
#options file
-database /blue/meilerlab/apps/rosetta/rosetta-3.4/rosetta_database
-loops:timer #output time spent in seconds for each loop modeling job
-loops:fa_input #input structures are in full atom format
-in:fix_disulf GHSR1.disulfide #read disulfide connectivity information
-in:file:spanfile GHSR1.span
-in:file:lipofile GHSR1.lips4
-loops:relax fastrelax #does a minimization of the structure in the torsion
space
-loops:extended true #force phi-psi angles to be set to 180 degrees
independent of loop input file (recommended for production runs)
-loops:frag_sizes 9 3 1
-loops:frag_files GHSR1.200.9mers GHSR1.200.3mers none
-loops:remodel quick_ccd
-loops:refine refine_kic
-out:file:silent_struct_type binary #output file type
-membrane:no_interpolate_Mpair # membrane scoring specification
-membrane:Menv_penalties # turn on membrane penalty scores
-score:weights membrane_highres_Menv_smooth.wts

# command line
rosetta-3.4/rosetta_source/bin/loopmodel.default.linuxgccrelease -database
rosetta-3.4/rosetta_database @fill_gaps.options -s GHSR1_on_"$TEMPLATE".pdb -
loops:input_pdb GHSR1_on_"$TEMPLATE".pdb -loops:loop_file
GHSR1_on_"$TEMPLATE"_init.loops -out:file:silent
GHSR1_on_"$TEMPLATE"_fillgaps.out -out:file:scorefile
GHSR1_on_"$TEMPLATE"_fillgaps.sc -nstruct 25
```

Filter for building ECLs

Filtered out models based on template 1U19 so that only took models with no chainbreaks within the top 10 by total score.

Rebuilding extracellular loops

```
#Options file
-database rosetta-3.4/rosetta_database
-loops:timer #output time spent in seconds for each loop modeling job
-loops:fa_input #input structures are in full atom format
-in:fix_disulf GHSR1.disulfide #read disulfide connectivity information
-in:file:spanfile GHSR1.span
-in:file:lipofile GHSR1.lips4
-in:detect_disulf true #NEW
-loops:relax fastrelax #does a minimization of the structure in the torsion
space
-loops:extended true #force phi-psi angles to be set to 180 degrees
independent of loop input file (recommended for production runs)
-loops:frag_sizes 9 3 1
-loops:frag_files GHSR1.200.9mers GHSR1.200.3mers none
-loops:ccd_closure
-loops:remodel quick_ccd
-loops:refine refine_kic
-ex1
-ex2
-relax:membrane #set up membrane environment for relax
-relax:fast
-out:file:silent_struct_type binary #output file type
-out:file:fullatom #output file will be fullatom
-membrane:no_interpolate_Mpair # membrane scoring specification
-membrane:Menv_penalties # turn on membrane penalty scores
-score:weights membrane_highres_Menv_smooth.wts

#Command
rosetta-3.4/rosetta_source/bin/loopmodel.default.linuxgccrelease -database
rosetta-3.4/rosetta_database @rebuild_ecl.options -s
GHSR1_on_"$TEMPLATE"_"$RANK".pdb -loops:input_pdb
GHSR1_on_"$TEMPLATE"_"$RANK".pdb -loops:loop_file GHSR1_on_"$TEMPLATE".loops -
out:file:silent GHSR1_on_"$TEMPLATE"_0"$RANK"_rebuild_ecl.out -
out:file:scorefile GHSR1_on_"$TEMPLATE"_0"$RANK"_rebuild_ecl.sc -nstruct 20
```

Selecting final model

The final model was selected by choosing the lowest scoring model overall.

Folding of ghrelin in the Rosetta membrane environment

FASTA file

```
>GHSRg_renumber
APLLAGVTATCVALFVVGIAGNLLTMLVVSRFRELRTTTNLYLSSMAFSDLLIFLCMPLDLVRLWQYRPWNFGDLLCK
LFQFVSESCTYATVLTITALSVERYFAICFPLRAKVVVTKGRVKLVIFVIWAVAFCSAGPIFVLVGVEHENGTDPWDT
NECRPTEFAVRSGLLTVMVWVSSIFFFLPVFCLTVLYSLIGRKLWRRRRGDAVVGASLRDQNHKQTVKMLAVVVFAFI
LCWLPFHVGRYLFSKSFEPGSLEIAQISQYCNLVSFVLFYLSAAINPILYNIMSKKYRVAVFRLLGFGSSFLSPEHQR
VQQRKESKKPPAKLQPR
```

Making fragments

See above section on making fragments for the receptor.

Spanfile

```
TM region consensus for Homo sapiens GHSR1 with ghrelin
7 329
antiparallel
n2c
   4    27     4    27
  41    61    41    61
  81   101    81   101
 123   142   123   142
 173   194   173   194
 224   243   224   243
 265   286   265   286
```

Lipophilicity file

Generated as before but with the spanfile directly above.

Rigid file (for Topology Broker)

```
RIGID 1 301
```

Topology broker setup file

```
CLAIMER MembraneTopologyClaimer
END_CLAIMER
CLAIMER RigidChunkClaimer
NO_USE_INPUT_POSE
PDB receptor.pdb
```

```
REGION_FILE  GHSRg.rigid
END_CLAIMER
```

Options file for de novo *folding*

```
-in
       -file
              -native receptor.pdb
              -fasta GHSRg.fasta
              -frag3 GHSRg.200.3mers
              -frag9 GHSRg.200.9mers
              -spanfile GHSRg.span
              -lipofile GHSRg.lips4
-residues
       -patch_selectors CENTROID_HA
-broker
       -setup GHSRg.tpb
-run
       -protocol broker
-score
       -find_neighbors_3dgrid
       -weights membrane_highres_Menv_smooth
-membrane
       -no_interpolate_Mpair
       -Menv_penalties
-abinitio
       -membrane
       -rg_reweight 0.00
       -stage2_patch score_membrane_s2.wts_patch
       -stage3a_patch score_membrane_s3a.wts_patch
       -stage3b_patch score_membrane_s3b.wts_patch
       -stage4_patch score_membrane_s4.wts_patch
-relax
       -membrane
       -fast
-ex1
-ex2
-out
       -output
       -file
              -fullatom
              -silent_struct_type binary
-overwrite
```

De novo *command line*

```
rosetta-3.4/rosetta_source/bin/minirosetta.static.linuxgccrelease -database
rosetta-3.4/rosetta_database/ @fold_GHSRg.flags -out::nstruct ${NSTRUCT} -
out:file:silent GHSRg.out -out:file:scorefile output/GHSRg.sc
```

Analysis and ensemble selection

Filter by proximity to membrane

```
#XML file

<dock_design>
    <SCOREFXNS> #defines non-standard score functions
    </SCOREFXNS>
    <FILTERS>
        <MembraneDepth name="membrane_depth" residue=304 depth_lb=48
depth_ub=60/> ### this covers polar region. Membrane is 0-60 (inner to outer)
    </FILTERS>
    <MOVERS>
    </MOVERS>
    <PROTOCOLS>
        <Add filter_name=membrane_depth/>
    </PROTOCOLS>
</dock_design>

#command

Rosetta/main/source/bin/rosetta_scripts.mpi.linuxgccrelease -database
Rosetta/main/database/ -in:file:l pdb.ls -parser:protocol test.xml -
out:file:score_only -in:file:spanfile GHSRg.span -in:file:lipofile GHSRg.lips4
-membrane:no_interpolate_Mpair -membrane:Menv_penalties -out:file:scorefile
MembraneDepth.sc -out:no_nstruct_label >& MembraneDepth.log &
```

Run PROSHIFT and format for further analysis

```
proshift.exe ${pdb} ${pdb}.pro 303 6

grep SHIFT ${pro} | grep -v PROSHIFT > ${pro}.cs
```

Run SPARTA+ and format for further analysis

```
sparta/SPARTA+/sparta+ -in input.pdb -ref GHSRg.tab -out outfile.out -outS
outfile.outstruct -outCS outfile.outcs -offset

tail -n55 outfile.outcs > outfile.outcs.tmp

cat outfile.outcs.tmp | awk '{print("SHIFT    "NR"\t"$3"\t"$2"
A\t"$1"\t"$6"\tppm +- "$9" ppm")}' > final.sparta.cs.out
```

Run SHIFTX

```
shiftx/./shiftx 1 ${pdb} ${pdb}.shiftx >& ${pdb}_shiftx.log
head -n30 ${pdb}.shiftx | tail -n28 > ${pdb}.shiftx.tmp
ls *.shiftx.tmp > shiftx.ls
./shiftx_to_proshift.py -i shiftx.ls --suffix cs –shiftx
```

Run SHIFTX2

```
python shiftx2-v107-linux/shiftx2.py -i ${pdb} -f TABULAR -p 6 -t 303 >&
${pdb}_shiftx2.log
head -n30 ${pdb}.cs | tail -n28 > ${pdb}.cs.shiftx2
ls *.shiftx2 > shiftx2.ls
./shiftx_to_proshift.py -i shiftx2.ls --suffix cs --shiftx2
```

Compare predicted chemical shifts to experimental chemical shifts

```
# list all predicted chemical shift files (one method at a time)
ls *.cs > cs.ls
awk '{system("/home/hirstsj/scripts/compare_cs.py --exp_cs GHSRg.tab.sd --
pred_cs " $1 " --outfile " $1 ".out --summary --no_sd --carbon_scale_factor
0.25")}' cs.ls

ls *.cs.out.summary >> all_outputs.ls

foreach file (`cat all_outputs.ls`)
       grep -H -v MaxDiff ${file} | awk
'{split($1,a,":");print(a[1]"\t"a[2]"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8
"\t"$9"\t"$10"\t"$11"\t"$12"\t"$13"\t"$14"\t"$15"\t"$16)}' | awk
'{split($1,a,".");print(a[1]"\t"$2"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8"\t"$9"\
t"$10"\t"$11"\t"$12"\t"$13"\t"$14"\t"$15"\t"$16)}' >> all_outputs.txt
end

echo "PDB   #yes    #no AvgDiff sdDiff  MaxDiff MaxRes# MaxResn MaxAtom
MaxExpCS    MaxExpLB    MaxExpUB    MaxProCS    MaxProLB    MaxProUB    RMSD"
> GHSRg_compare_cs.out

cat all_outputs.txt >> GHSRg_compare_cs.out
```

Input experimental chemical shifts

```
#    1 G    C  167.040    0.4
#    1 G   CA   40.900    0.2
    2 S    C  172.100    0.2
    2 S   CA   55.600    0.5
    2 S   CB   62.500    0.6
    3 S   CA   53.580    0.10
    3 S   CB   63.270    0.21
    3 S   HA    4.488    9999
    4 F    C  172.100    0.2
    4 F   CA   55.800    1.2
    4 F   CB   37.000    0.8
    5 L    C  174.800    0.4
    5 L   CA   51.900    0.2
    5 L   CB   40.700    0.5
    6 S    C  169.300    0.5
    6 S   CA   54.250    0.5
```

```
 6 S   CB   61.250     0.6
 7 P    C  174.900     1.2
 7 P   CA   61.250     0.5
 7 P   CB   30.800     1.7
 8 E    C  174.100     0.2
 8 E   CA   54.300     0.9
 8 E   CB   25.800     0.9
10 Q    C  177.300     9999
10 Q   CA   55.420     0.35
10 Q   CB   27.010     0.02
10 Q   HA    4.130     9999
12 V    C  174.100     0.2
12 V   CA   60.300     0.9
12 V   CB   30.000     9999
13 Q    C  173.350     0.26
13 Q   CA   53.530     0.18
13 Q   CB   26.950     0.12
13 Q   HA    4.310     9999
14 Q    C  173.500     9999
14 Q   CA   53.440     0.14
14 Q   CB   26.950     0.05
14 Q   HA    4.296     9999
18 S    C  171.760     0.16
18 S   CA   55.800     0.1
18 S   CB   61.300     0.2
18 S   HA    4.440     9999
21 P    C  177.700     9999
21 P   CA   58.980     0.09
21 P   CB   28.320     0.04
21 P   HA    4.720     9999
22 P    C  173.670     0.18
22 P   CA   60.430     0.25
22 P   CB   29.410     0.03
22 P   HA    4.440     9999
23 A    C  175.500     0.5
23 A   CA   50.500     0.6
23 A   CB   17.000     9999
27 P    C  173.320     0.03
27 P   CA   60.770     0.021
27 P   CB   29.420     0.14
27 P   HA    4.430     9999
```

Run ensemble selection script

```
./find_best_ensemble.py --ncycles 5000000 --min_ensemble_size 10 --
max_ensemble_size 30 --outfile outfile.out
/directory/to/predicted/cs/in/proshift/format/ending/in/*.cs
```

Run DSSP and compute phi/psi angles for models

```
./run_dssp.py -i pdb.ls --all all.out
```

Find polyproline II helix residues

```
Ls *.dssp > dssp.ls
foreach file ( `cat dssp.ls` )
       awk '{if(($3>=-104.0 && $3<=-46.0) && ($4>=116.0 && $4<=174.0) &&
($2=="-"))print}' ${file} > ${file}.pp2
end
```

APPENDIX D

PROTOCOL CAPTURE FOR CHAPTER IV: ROSETTAEPR: AN INTEGRATED TOOL FOR PROTEIN STRUCTURE DETERMINATION FROM SPARSE EPR DATA

This appendix contains the protocol capture for the modeling work published in (Hirst, Alexander, Mcaourab, and Meiler, 2011), some of which is found in the manuscript's Supplemental Information. Further details are also available in the main text (Chapter IV).

Computational details

All Rosetta-related protocols were conducted using Rosetta version 3 revision number 34586. All models were generated by independent simulations using Vanderbilt University's Center for Structural Biology computing cluster and the university's Advanced Computing Center for Research and Education (ACCRE). Computations were performed on a combination of AMD Opteron and Intel Nehalem processor nodes.

Input files

FASTA file

```
> 2LZM Sequence
ITKDEAEKLFNQDVDAAVRGILRNAKLKPVYDSLDAVRRCALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLAK
SRWYNQTPNRAKRVITTFRTGTWDAYKNL
```

Constraints file used in de novo *folding with RosettaEPR*

```
AtomPair        CB      32      CB      36      SPLINE  EPR_DISTANCE    16.0    4.0     0.5
AtomPair        CB      59      CB      74      SPLINE  EPR_DISTANCE    19.0    4.0     0.5
AtomPair        CB      62      CB      71      SPLINE  EPR_DISTANCE    19.0    4.0     0.5
AtomPair        CB      62      CB      74      SPLINE  EPR_DISTANCE    25.0    4.0     0.5
AtomPair        CB      63      CB      74      SPLINE  EPR_DISTANCE    14.0    4.0     0.5
AtomPair        CB      66      CB      74      SPLINE  EPR_DISTANCE    23.0    4.0     0.5
AtomPair        CB      83      CB      90      SPLINE  EPR_DISTANCE    13.0    4.0     0.5
AtomPair        CB      83      CB      94      SPLINE  EPR_DISTANCE    18.0    4.0     0.5
AtomPair        CB      8       CB      19      SPLINE  EPR_DISTANCE    21.4    4.0     0.5
AtomPair        CB      8       CB      78      SPLINE  EPR_DISTANCE    46.3    4.0     0.5
AtomPair        CB      4       CB      78      SPLINE  EPR_DISTANCE    47.2    4.0     0.5
AtomPair        CB      8       CB      29      SPLINE  EPR_DISTANCE    37.4    4.0     0.5
AtomPair        CB      4       CB      29      SPLINE  EPR_DISTANCE    37.5    4.0     0.5
AtomPair        CB      4       CB      23      SPLINE  EPR_DISTANCE    34.0    4.0     0.5
AtomPair        CB      8       CB      23      SPLINE  EPR_DISTANCE    26.5    4.0     0.5
AtomPair        CB      23      CB      78      SPLINE  EPR_DISTANCE    36.8    4.0     0.5
AtomPair        CB      31      CB      43      SPLINE  EPR_DISTANCE    6.0     4.0     0.5
AtomPair        CB      32      CB      39      SPLINE  EPR_DISTANCE    6.0     4.0     0.5
AtomPair        CB      29      CB      62      SPLINE  EPR_DISTANCE    15.0    4.0     0.5
AtomPair        CB      70      CB      94      SPLINE  EPR_DISTANCE    14.0    4.0     0.5
AtomPair        CB      70      CB      97      SPLINE  EPR_DISTANCE    13.0    4.0     0.5
AtomPair        CB      74      CB      93      SPLINE  EPR_DISTANCE    13.0    4.0     0.5
AtomPair        CB      74      CB      94      SPLINE  EPR_DISTANCE    9.0     4.0     0.5
AtomPair        CB      74      CB      97      SPLINE  EPR_DISTANCE    10.0    4.0     0.5
AtomPair        CB      77      CB      94      SPLINE  EPR_DISTANCE    9.0     4.0     0.5
```

Constraints file used in de novo *folding with bounded restraints*

```
AtomPair        CB      32      CB      36      BOUNDED 0.5     21.5    1.0     NOE     ;dist
AtomPair        CB      59      CB      74      BOUNDED 0.0     31.5    1.0     NOE     ;dist
AtomPair        CB      62      CB      71      BOUNDED 2.5     25.5    1.0     NOE     ;dist
AtomPair        CB      62      CB      74      BOUNDED 7.5     32.5    1.0     NOE     ;dist
AtomPair        CB      63      CB      74      BOUNDED 0.0     19.5    1.0     NOE     ;dist
AtomPair        CB      66      CB      74      BOUNDED 5.5     30.5    1.0     NOE     ;dist
AtomPair        CB      83      CB      90      BOUNDED 0.0     22.5    1.0     NOE     ;dist
AtomPair        CB      83      CB      94      BOUNDED 0.0     29.5    1.0     NOE     ;dist
AtomPair        CB      8       CB      19      BOUNDED 6.1     26.7    1.0     NOE     ;dist
AtomPair        CB      8       CB      78      BOUNDED 31.6    51      1.0     NOE     ;dist
AtomPair        CB      4       CB      78      BOUNDED 32.5    51.9    1.0     NOE     ;dist
AtomPair        CB      8       CB      29      BOUNDED 22.2    42.6    1.0     NOE     ;dist
AtomPair        CB      4       CB      29      BOUNDED 23      42      1.0     NOE     ;dist
AtomPair        CB      4       CB      23      BOUNDED 19.3    38.7    1.0     NOE     ;dist
AtomPair        CB      8       CB      23      BOUNDED 10.2    32.8    1.0     NOE     ;dist
AtomPair        CB      23      CB      78      BOUNDED 23.3    40.3    1.0     NOE     ;dist
AtomPair        CB      31      CB      43      BOUNDED 0.0     11.5    1.0     NOE     ;dist
AtomPair        CB      32      CB      39      BOUNDED 0.0     11.5    1.0     NOE     ;dist
AtomPair        CB      29      CB      62      BOUNDED 0.0     20.5    1.0     NOE     ;dist
AtomPair        CB      70      CB      94      BOUNDED 0.0     18.9    1.0     NOE     ;dist
AtomPair        CB      70      CB      97      BOUNDED 0.0     21.5    1.0     NOE     ;dist
AtomPair        CB      74      CB      93      BOUNDED 0.0     21.5    1.0     NOE     ;dist
AtomPair        CB      74      CB      94      BOUNDED 0.0     19.5    1.0     NOE     ;dist
AtomPair        CB      74      CB      97      BOUNDED 0.0     18.5    1.0     NOE     ;dist
AtomPair        CB      77      CB      94      BOUNDED 0.0     17.5    1.0     NOE     ;dist
```

De novo *folding options file*

10,000 T4-lysozyme models with 25 EPR distance restraints scored according to the RosettaEPR knowledge-based potential

```
-abinitio::increase_cycles 2.5
-fold_cst::force_minimize
-constraints::cst_file ./2LZM_dist_w4.cst
-constraints::cst_weight 1.0
-constraints::epr_distance
-constraints::viol
-constraints::viol_level 101
-frags::scoring
-frags::picking::selecting_rule BestTotalScoreSelector
-in::path::database minirosetta_database_r34586
-in::file::native ./2LZM_.pdb
-in::file::fasta ./2LZM_.fasta
-in::file:frag3 ./aa2LZM_03_05.200_v1_3
-in::file::frag9 ./aa2LZM_09_05.200_v1_3
-out::output
-out::prefix 2LZM_
-out::file::silent ./2LZM_.out
-out::file::silent_struct_type binary
-out::file::scorefile ./2LZM_.sc
-out::nstructs 10000
-out::show_accessed_options
```

Full-atom refinement options file

One T4-lysozyme *de novo* folded model with no distance restraints, resulting in ten new models complete with amino acid side-chains

```
-relax::sequence
-in::path::database ./minirsetta_database_r34586
-in::file::native ./2LZM_.pdb
-in::file::fullatom
-corrections::correct
-out::output
-out::prefix 2LZM_fa_
-out::file::silent ./2LZM_fa.out
-out::file::silent_struct_type binary
-out::file::scorefile ./2LZM_fa.fsc
-out::nstructs 10
-out::show_accessed_options
```

Command lines

Fragment generation

```
make_fragments.pl -id 2LZM_ -nohoms 2LZM_.fasta
```

De novo *folding*

```
/bin/AbinitioRelax.linuxgccrelease @2LZM_w4_folding.options
```

Full-atom refinement

```
/bin/relax.linuxgccrelease @2LZM_rlx.options
```

RMSD histogram distribution

```
perl Smbins_RMSD_dist_from_score.pl <file with rmsds> <rmsd col. #>
```

APPENDIX E

PROTOCOL CAPTURE FOR CHAPTER V: ROSETTATMH: MEMBRANE PROTEIN STRUCTURE ELUCIDATION BY COMBINING EPR DISTANCE RESTRAINTS WITH ASSEMBLY OF TRANSMEMBRANE HELICES

This appendix contains the protocol capture for the modeling performed for the Chapter V, which is based on the manuscript submitted to *PLoS ONE* of the same title by Stephanie DeLuca, Sam DeLuca, Andrew Leaver-Fay, and Jens Meiler

Preparation for folding

Parameter optimization / testing PDBs

```
1FX8A, 1KPLA, 1PY6A, 1U19A, 3B60A, 3GIAA, 3HD6A, 3HFXA, 3O0RB
```

Benchmark set

```
1FX8A,1IWGA,1J4NA,1KPLA,1OCCC,1OKCA,1PV6A,1PY6A,1PY7A,1RHZA,1U19A,2BG9A,2BL2A,
2BS2A,2IC8A,2K73A,2KSFA,2KSYA,2NR9A,2PNOA,2XQ2A,2XUTA,2YVXA,2ZW3A,3B60A,3GIAA,
3HD6A,3HFXA,3KCUA,3KJ6A,3O0RB,3P5NA,3SYOA,4A2NB
```

FASTA files

```
1FX8A
>BCL :A|PDBID|CHAIN|SEQUENCE
TLKGQCIAEFLGTGLLIFFGVGCVAALKVAGASFGQWEISVIWGLGVAMA
IYLTAGVSGAHLNPAVTIALWLFACFDKRKVIPFIVSQVAGAFCAAALVY
GLYYNLFFDFEQTHHIVRGSVESVDLAGTFSTYPNPHINFVQAFAVEMVI
TAILMGLILALTDDGNGVPRGPLAPLLIGLLIAVIGASMGPLTGFAMNPA
RDFGPKVFAWLAGWGNVAFTGGRDIPYFLVPLFGPIVGAIVGAFAYRKLI
GRHL

1IWGA
> 1IWGA
SIHEVVKTLVEAIILVFLVMYLFLQNFRATLIPTIAVPVVLLGTFAVLAAFGFSINTLTMFGMVLAIGLLVDDAIVVV
ENVERVMAEEGLPPKEATRKSMGQIQGALVGIAMVLSAVFVPMAFFGGSTGAIYRQFSITIVSAMALSVLVALILTPA
LCATMLK
```

1J4NA
>BCL :A|PDBID|CHAIN|SEQUENCE
EFKKKLFWRAVVAEFLAMILFIFISIGSALGFHYPIKSNQTTGAVQDNVK
VSLAFGLSIATLAQSVGHISGAHLNPAVTLGLLLSCQISVLRAIMYIIAQ
CVGAIVATAILSGITS

1KPLA
>BCL :A|PDBID|CHAIN|SEQUENCE
TPLAILFMAAVVGTLTGLVGVAFEKAVSWVQNMRIGALVQVADHAFLLWP
LAFILSALLAMVGYFLVRKFAPEAGGSGIPEIEGALEELRPVRWWRVLPV
KFIGGMGTLGAGMVLGREGPTVQIGGNLGRMVLDVFRMRSAEARHTLLAT
GAAAGLSAAFNAPLAGILFIIEEMRPQFRYNLISIKAVFTGVIMSSIVFR
IFN

1OCCC
>BCL :C|PDBID|CHAIN|SEQUENCE
HTPAVQKGLRYGMILFIISEVLFFTGFFWAFYHSSLAPTPELGGCWPPTG
IHPLNPLEVPLLNTSVLLASGVSITWAHHSLMEGDRKHMLQALFITITLG
VYFTLLQASEYYEAPFTISDGVYGSTFFVATGFHGLHVIIGSTFLIVCFF
RQLKFHFTSNHHFGFEAGAWYWHFVDVVWLFLYVSIYWWGS

1OKCA
>BCL :A|PDBID|CHAIN|SEQUENCE
DQALSFLKDFLAGGVAAAISKTAVAPIERVKLLLQVQHASKQISAEKQYK
GIIDCVVRIPKEQGFLSFWRGNLANVIRYFPTQALNFAFKDKYKQIFLGG
VDRHKQFWRYFAGNLASGGAAGATSLCFVYPLDFARTRLAADVGKGAAQR
EFTGLGNCITKIFKSDGLRGLYQGFNVSVQGIIIYRAAYFGVYDTAKGML
PDPKNVHIIVSWMIAQTVTAVAGLVSYPFDTVRRRMMMQSGRKGADIMYT
GTVDCWRKIAKDEGPKAFFKGAWSNVLRGMGGAFVLVLYDEI

1PV6A
>BCL :A|PDBID|CHAIN|SEQUENCE
MYYLKNTNFWMFGLFFFFYFFIMGAYFPFFPIWLHDINHISKSDTGIIFA
AISLFSLLFQPLFGLLSDKLGLRKYLLWIITGMLVMFAPFFIFIFGPLLQ
YNILVGSIVGGIYLGFCFNAGAPAVEAFIEKVSRRSNFEFGRARMFGCVG
WALGASIVGIMFTINNQFVFWLGSGCALILAVLLFFAKT

1PY6A
>BCL :A|PDBID|CHAIN|SEQUENCE
TGRPEWIWLALGTALMGLGTLYFLVKGMGVSDPDAKKFYAITTLVPAIAF
TMYLSMLLGYGLTMVPFGGEQNPIYWARYADWLFTTPLLLLDLALLVDAD
QGTILALVGADGIMIGTGLVGALTKVYSYRFVWWAISTAAMLYILYVLFF
GFTSKAESMRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEGAGIVPLNIE
TLLFMVLDVSAKVGFGLILLRSRAIFG

1PY7A
>BCL :A|PDBID|CHAIN|SEQUENCE
PIYWARYADWLFTTPLLLLDLALLVDADQGTILALVGADGIMIGTGLVGA
LTKVYSYRFVWWAISTAAMLYILYVLFFGFTSKAESMRPEVASTFKVLRN
VTVVLWSAYPVVWLIGSEGAGIV

1RHZA
>BCL :A|PDBID|CHAIN|SEQUENCE
FKEKLKWTGIVLVLYFIMGCIDVYTAGAQIPAIFEFWQTITASRIGTLIT

LGIGPIVTAGIIMQLLVGSGIIQMDLSIPENRALFQGCQKLLSIIMCFVE
AVLFVGAGAFGILTPLLAFLVIIQIAFGSIILIYLDEIVSKYGIGSGIGL
FIAAGVSQTIFVGALG

1U19A
>BCL :A|PDBID|CHAIN|SEQUENCE
EPWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVA
DLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLGGEIALWSLVVLA
IERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIPEG
MQCSCGIDYYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLVFTVKE
AAAQQQESATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSD
FGPIFMTIPAFFAKTSAVYNPVIYIMMN

2BG9A
>BCL :A|PDBID|CHAIN|SEQUENCE
PLYFVVNVIIPCLLFSFLTVLVFYLPTDSGEKMTLSISVLLSLTVFLLVI
VELIPSTSSAVPLIGKYMLFTMIFVISSIIVTVVVINTHHR

2BL2A
>BCL :A|PDBID|CHAIN|SEQUENCE
MVFAVLAMATATIFSGIGSAKGVGMTGEAAAALTTSQPEKFGQALILQLL
PGTQGLYGFVIAFLIFINLGSDMSVVQGLNFLGASLPIAFTGLFSGIAQG
KVAAAGIQILAKKPEHATKGIIFAAMVETYAILGFVISFLLVLNA

2BS2A
>BCL :C|PDBID|CHAIN|SEQUENCE
RMPAKLDWWQSATGLFLGLFMIGHMFFVSTILLGDNVMLWVTKKFELDFI
FEGGKPIVVSFLAAFVFAVFIAHAFLAMRKFPINYRQYLTFKTHKDLMRH
GDTTLWWIQAMTGFAMFFLGSVHLYIMMTQPQTIGPVSSSFRMVSEWMWP
LYLVLLFAVELHGSVGLYRLAVKWGWFDGETPDKTRANLKKLKTLMSAFL
IVLGLLTFGAYVKKGLE

2IC8A
>BCL :A|PDBID|CHAIN|SEQUENCE
ERAGPVTWVMMIACVVVFIAMQILGDQEVMLWLAWPFDPTLKFEFWRYFT
HALMHFSLMHILFNLLWWWYLGGAVEKRLGSGKLIVITLISALLSGYVQQ
KFSGPWFGGLSGVVYALMGYVWLRGERDPQSGIYLQRGLIIFALIWIVAG
WFDLFGMSMANGAHIAGLAVGLAMAFVDSLNA

2K73A
>BCL :A|PDBID|CHAIN|SEQUENCE
MLRFLNQASQGRGAWLLMAFTALALELTALWFQHVMLLKPCVLSIYERAA
LFGVLGAALIGAIAPKTPLRYVAMVIWLYSAFRGVQLTYEHTMLQLYPSP
FATSDFMVRFPEWLPLDKWVPQVFVASGDCAERQWDFLGLEMPQWLLGIF
IAYLIVAVLVVISQ

2KSFA
>BCL :A|PDBID|CHAIN|SEQUENCE
MVQIQGSVVAAALSAVITLIAMQWLMAFDAANLVMLYLLGVVVVALFYGR
WPSVVATVINVVSFDLFFIAPRGTLAVSDVQYLLTFAVMLTVGLVIGNLT
AGVRYQA

2KSYA
>BCL :A|PDBID|CHAIN|SEQUENCE

MVGLTTLFWLGAIGMLVGTLAFAWAGRDAGSGERRYYVTLVGISGIAAVA
YAVMALGVGWVPVAERTVFVPRYIDWILTTPLIVYFLGLLAGLDSREFGI
VITLNTVVMLAGFAGAMVPGIERYALFGMGAVAFIGLVYYLVGPMTESAS
QRSSGIKSLYVRLRNLTVVLWAIYPFIWLLGPPGVALLTPTVDVALIVYL
DLVTKVGFGFIALDAAATLRAEH

2NR9A
>BCL :A|PDBID|CHAIN|SEQUENCE
FLAQQGKITLILTALCVLIYIAQQLGFEDDIMYLMHYPAYEEQDSEVWRY
ISHTLVHLSNLHILFNLSWFFIFGGMIERTFGSVKLLMLYVVASAITGYV
QNYVSGPAFFGLSGVVYAVLGYVFIRDKLNHHLFDLPEGFFTMLLVGIAL
GFISPLFGVEMGNAAHISGLIVGLIWGFIDSKLRKNSLELVP

2PNOA
>BCL :A|PDBID|CHAIN|SEQUENCE
KDEVALLAAVTLLGVLLQAYFSLQVISARRAFRVSPPLTTGPPEFERVYR
AQVNCSEYFPLFLATLWVAGIFFHEGAAALCGLVYLFARLRYFQGYARSA
QLRLAPLYASARALWLLVALAALGLLAHFL

2XQ2A
>BCL :A|PDBID|CHAIN|SEQUENCE
SFIDIMVFAIYVAIIIGVGLWVSRDKKGTQKSTEDYFLAGKSLPWWAVGA
SLIAANISAEQFIGMSGSGYSIGLAIASYEWMSAITLIIVGKYFLPIFIE
KGIYTIPEFVEKRFNKKLKTILAVFWISLYIFVNLTSVLYLGGLALETIL
GIPLMYSILGLALFALVYSIYGGLSAVVWTDVIQVFFLVLGGFMTTYMAV
SFIGGTDGWFAGVSKMVDAAPGHFEMILDQSNPQYMNLPGIAVLIGGLWV
ANLYYWGFNQYIIQRTLAAKSVSEAQKGIVFAAFLALIVPFLVVLPGIAA
YVITSDPQLMASLGDIAATNLPSAANADKAYPWLTQFLPVGVKGVVFAAL
AAAIVSSLASMLNSTATIFTMDIYKEYISPDSGDHKLVNVGRTAAVVALI
IAALIAPMLGGIGQCFQYIQEYTGLVSPGILAVFLLGLFWKKTTSKGAII
GVVASIPFALFLKFMPLSMPFMDQMLYTLLFTMVVIAFTSLSTSINDDDP
KGISVTSSMFVTDRSFNIAAYGIMIVLAVLYTLFWVNADAEITLIIFGVM
AGVIGTILLISYGIK

2XUTA
>BCL :A|PDBID|CHAIN|SEQUENCE
QIPYIIASEACERFSFYGMRNILTPFLMTALLLSIPEELRGAVAKDVFHS
FVIGVYFFPLLGGWIADRFFGKYNTILWLSLIYCVGHAFLAIFEHSVQGF
YTGLFLIALGSGGIKPLVSSFMGDQFDQSNKSLAQKAFDMFYFTINFGSF
FASLSMPLLLKNFGAAVAFGIPGVLMFVATVFFWLGRKRYIHMPPEPKDP
HGFLPVIRSALLTKVEGKGNIGLVLALIGGVSAAYALVNIPTLGIVAGLC
CAMVLVMGFVGAGASLQLERARKSHPDAAVDGVRSVLRILVLFALVTPFW
SLFDQKASTWILQANDMVKPQWFEPAMMQALNPLLVMLLIPFNNFVLYPA
IERMGVKLTALRKMGAGIAITGLSWIVVGTIQLMMDGGSALSIFWQILPY
ALLTFGEVLVSATGLEFAYSQAPKAMKGTIMSFWTLSVTVGNLWVLLANV
SVKSPTVTEQIVQTGMSVTAFQMFFFAGFAILAAIVFA

2YVXA
>BCL :A|PDBID|CHAIN|SEQUENCE
HKLGAVDVPDLVYSEAGPVALWLARVRWLVILILTGMVTSSILQGFESVL
EAVTALAFYVPVLLGTGGNTGNQSATLIIRALATRDLDLRDWRRVFLKEM
GVGLLLGLTLSFLLVGKVYWDGHPLLLPVVGVSLVLIVFFANLVGAMLPF
LLRRLGVDPALVSNPLVATLSDVTGLLIYLSVARLLLE

2ZW3A
>BCL :A|PDBID|CHAIN|SEQUENCE
DWGTLQTILGGVNKHSTSIGKIWLTVLFIFRIMILVVAAKEVWGDEQADF
VCNTLQPGCKNVCYDHYFPISHIRLWALQLIFVSTPALLVAMHVAYRRHE
KKRKFIKGEIKSEFKDIEEIKTQKVRIEGSLWWTYTSSIFFRVIFEAAFM
YVFYVMYDGFSMQRLVKCNAWPCPNTVDCFVSRPTEKTVFTVFMIAVSGI
CILLNVTELCYLLIRY

3B60A
>BCL :A|PDBID|CHAIN|SEQUENCE
WQTFRRLWPTIAPFKAGLIVAGIALILNAASDTFMLSLLKPLLDDGFGKT
DRSVLLWMPLVVIGLMILRGITSYISSYCISWVSGKVVMTMRRRLFGHMM
GMPVAFFDKQSTGTLLSRITYDSEQVASSSSGALITVVREGASIIGLFIM
MFYYSWQLSIILVVLAPIVSIAIRVVSKRFRSISKNMQNTMGQVTTSAEQ
MLKGHKEVLIFGGQEVETKRFDKVSNKMRLQGMKMVSASSISDPIIQLIA
SLALAFVLYAASFPSVMDSLTAGTITVVFSSMIALMRPLKSLTNVNAQFQ
RGMAACQTLFAILDSEQEK

3GIAA
>BCL :A|PDBID|CHAIN|SEQUENCE
LKNKKLSLWEAVSMAVGVMIGASIFSIFGVGAKIAGRNLPETFILSGIYA
LLVAYSYTKLGAKIVSNAGPIAFIHKAIGDNIITGALSILLWMSYVISIA
LFAKGFAGYFLPLINAPINTFNIAITEIGIVAFFTALNFFGSKAVGRAEF
FIVLVKLLILGLFIFAGLITIHPSYVIPDLAPSAVSGMIFASAIFFLSYM
GFGVITNASEHIENPKKNVPRAIFISILIVMFVYVGVAISAIGNLPIDEL
IKASENALAVAAKPFLGNLGFLLISIGALFSISSAMNATIYGGANVAYSL
AKDGELPEFFERKVWFKSTEGLYITSALGVLFALLFNMEGVASITSAVFM
VIYLFVILSHYILIDEVGGRKEIVIFSFIVVLGVFLLLLYYQWITNRFVF
YGIIATFIGVLIFEIIYRKVTKRTFSNNMYVKS

3HD6A
>BCL :A|PDBID|CHAIN|SEQUENCE
SAWNTNLRWRLPLTCLLLQVIMVILFGVFVRYDFENEFYYRYPSFQDVHV
MVFVGFGFLMTFLQRYGFSAVGFNFLLAAFGIQWALLMQGWFHFLQDRYI
VVGVENLINADFCVASVCVAFGAVLGKVSPIQLLIMTFFQVTLFAVNEFI
LLNLLKVKDAGGSMTIHTFGAYFGLTVTRILYRRNLEQSKERQNSVYQSD
LFAMIGTLFLWMYWPSFNSAISYHGDSQHRAAINTYCSLAACVLTSVAIS
SALHKKGKLDMVHIQNATLAGGVAVGTAAEMMLMPYGALIIGFVCGIIST
LGFVYLTPFLESRLHIQDTCGINNLHGIPGIIGGIVGAVTAASDWTARTQ
GKFQIYGLLVTLAMALMGGIIVGLILRLPFWGQPSDENCFEDAVYWEMPE
GNS

3HFXA
>BCL :A|PDBID|CHAIN|SEQUENCE
PKVFFPPLIIVGILCWLTVRDLDAANVVINAVFSYVTNVWGWAFEWYMVV
MLFGWFWLVFGPYAKKRLGNEPPEFSTASWIFMMFASCTSAAVLFWGSIE
IYYYISTPPFGLEPNSTGAKELGLAYSLFHWGPLPWATYSFLSVAFAYFF
FVRKMEVIRPSSTLVPLVGEKHAKGLFGTIVDNFYLVALIFAMGTSLGLA
TPLVTECMQWLFGIPHTLQLDAIIITCWIILNAICVACGLQKGVRIASDV
RSYLSFLMLGWVFIVSGASFIMNYFTDSVGMLLMYLPRMLFYTDPIAKGG
FPQGWTVFYWAWWVIYAIQMSIFLARISRGRTVRELCFGMVLGLTASTWI
LWTVLGSNTLLLIDKNIINIPNLIEQYGVARAIIETWAALPLSTATMWGF
FILCFIATVTLVNACSYTLAMSTCREVRDGEEPPLLVRIGWSILVGIIGI
VLLALGGLKPIQTAIIAGGCPLFFVNIMVTLSFIKDAKQNWKD

3KCUA
>BCL :A|PDBID|CHAIN|SEQUENCE
KHPLKTFYLAITAGVFISIAFVFYITATTGTGTMPFGMAKLVGGICFSLG
LILCVVCGADLFTSTVLIVVAKASGRITWGQLAKNWLNVYFGNLVGALLF
VLLMWLSGEYMTANGQWGLNVLQTADHKVHHTFIEAVCLGILANLMVCLA
VWMSYSGRSLMDKAFIMVLPVAMFVASGFEHSIANMFMIPMGIVIRDFAS
PEFWTAVGSAPENFSHLTVMNFITDNLIPVTIGNIIGGGLLVGLTYWVIY
LR

3KJ6A
>BCL :A|PDBID|CHAIN|SEQUENCE
GMGIVMSLIVLAIVFGNVLVITAIAKFERLQTVTNYFITSLACADLVMGL
AVVPFGAAHILMKMWTFGNFWCEFWTSIDVLCVTASIETLCVIAVDRYFA
ITSPFKYQSLLTKNKARVIILMVWIVSGLTSFLPIQMHWYRATHQEAINC
YAEETCCDFFTNQAYAIASSIVSFYVPLVIMVFVYSRVFQEAKRQLQKID
KSEGRFHVQNLSQVEQDGRTGHGLRRSSKFCLKEHKALKTLGIIMGTFTL
CWLPFFIVNIVHVIQDNLIRKEVYILLNWIGYVNSGFNPLIYCRSPDFRI
AFQELLCLRRS

3O0RB
>BCL :B|PDBID|CHAIN|SEQUENCE
FASQAVAKPYFVFALILFVGQILFGLIMGLQYVVGDFLFPAIPFNVARMV
HTNLLIVWLLFGFMGAAYYLVPEESDCELYSPKLAWILFWVFAAAGVLTI
LGYLLVPYAGLARLTGNELWPTMGREFLEQPTISKAGIVIVALGFLFNVG
MTVLRGRKTAISMVLMTGLIGLALLFLFSFYNPENLTRDKFYWWWVVHLW
VEGVWELIMGAILAFVLVKITGVDREVIEKWLYVIIAMALISGIIGTGHH
YFWIGVPGYWLWLGSVFSALEPLPFFAMVLFAFNTINRRRRDYPNRAVAL
WAMGTTVMAFLGAGVWGFMHTLAPVNYYTHGTQLTAAHGHMAFYGAYAMI
VMTIISYAMPRLRGIGEAMDNRSQVLEMWGFWLMTVAMVFITLFLSAAGV
LQVWLQRMPADGAAMTFMATQDQLAIFYWLREGAGVVFLIGLVAYLLSF

3P5NA
>BCL :A|PDBID|CHAIN|SEQUENCE
QQNKRLITISMLSAIAFVLTFIKFPIPFLPPYLTLDFSDVPSLLATFTFG
PVAGIIVALVKNLLNYLFSMGDPVGPFANFLAGASFLLTAYAIYKNKRST
KSLITGLIIATIVMTIVLSILNYFVLLPLYGMIFNLADIANNLKVIIVSG
IIPFNIIKGIVISIVFILLYRRLANFLKR

3SYOA
>BCL :A|PDBID|CHAIN|SEQUENCE
YRYLTDIFTTLVDLKWRFNLLIFVMVYTVTWLFFGMIWWLIAYIRGDMDH
IEDPSWTPCVTNLNGFVSAFLFSIETETTIGYGYRVITDKCPEGIILLLI
QSVLGSIVNAFMVGCMFVKISQ

4A2NB
>BCL :B|PDBID|CHAIN|SEQUENCE
MNENLWKICFIVMFIIWVFVRKVYGTRAMKNKSKKKVRPNFEKSLVFLNF
IGMVFLPLTAVFSSYLDSFNINLPDSIRLFALIVTFLNIGLFTKIHKDLG
NNWSAILEIKDGHKLVKEGIYKNIRHPMYAHLWLWVITQGIILSNWVVLI
FGIVAWAILYFIRVPKEEELLIEEFGDEYIEYMGKTGRLFPK

Fragment files

The fragment files and secondary structure prediction files are the same ones used for the modeling published in reference (345). Information on how this input was generated can be found in the publication's Supplemental Information.

```
# Example using 1U19. Was repeated for all nine proteins in benchmark set
rosetta-3.4/rosetta_tools/fragment_tools/make_fragments.pl -id 1U19A -
psipredfile 1U19A.psipred_ss2 –jufofile 1U19A.jufo -nosam -verbose 1U19A.fasta
-nohoms -nojufo –nopsipred
```

Generation of spanfiles and lipophilicity files

The spanfiles and lipophilicity files are the same ones used for the modeling published in reference (345). Information on how this input was generated can be found in the publication's Supplemental Information. Span file generated using Rosetta version 3.4 octopus2span.pl script and the SPOCTOPUS prediction as input. Lipophilicty files generated using run_lips.pl script.

```
# running octopus2span.pl
/rosetta-3.4/rosetta_source/src/apps/public/membrane_abinitio/octopus2span.pl
<OCTOPUS topology file> > spanfile

# Example spanfile
TM region prediction for 1U19A.octo_topo predicted using OCTOPUS
7 278
antiparallel
n2c
   7    27     7    27
  43    63    43    63
  82   102    82   102
 120   140   120   140
 171   191   171   191
 221   241   221   241
 255   275   255   275

# running run_lips.pl
/rosetta-3.4/rosetta_source/src/apps/public/membrane_abinitio/run_lips.pl
1U19A.fasta 1U19A.span /dir/blastpgp /dir/alignblast.pl
```

Generation of Topology Broker "rigid" files for computing $RMSD_{100}SSE$

Taken from secondary structure element definitions from DSSP and are similar to that used for evaluation of models in reference (345).

```
# example rigid file for 1U19
RIGID  2     32
RIGID  39    57
RIGID  59    68
RIGID  74    107
RIGID  118   136
RIGID  168   178
RIGID  181   193
RIGID  210   245
RIGID  253   262
RIGID  269   276
```

Residues over which $RMSD_{100}SSE$ was computed

All native PDBs were renumbered starting at residue 1, as are all models folded with Rosetta. This is what the following lists of residues assume.

```
# 1FX8A
RIGID  2     29
RIGID  36    58
RIGID  64    73
RIGID  78    114
RIGID  121   130
RIGID  140   162
RIGID  173   192
RIGID  199   212
RIGID  227   253

# 1IWGA
RIGID  2     23
RIGID  31    51
RIGID  57    85
RIGID  93    122
RIGID  129   160
# 1J4NA
RIGID  2     31
RIGID  48    70
RIGID  76    85
RIGID  90    115
# 1KPLA
RIGID  2     40
RIGID  48    70
RIGID  79    87
```

```
RIGID  97    111
RIGID  117   136
RIGID  141   160
RIGID  163   174
RIGID  185   202

# 1OCCA
RIGID  3     36
RIGID  59    83
RIGID  86    113
RIGID  122   154
RIGID  163   189

# 1OKCA
RIGID  3     36
RIGID  72    98
RIGID  107   141
RIGID  175   198
RIGID  208   239
RIGID  272   290

# 1PV6A
RIGID  7     38
RIGID  42    70
RIGID  74    101
RIGID  104   136
RIGID  140   164
RIGID  166   186

# 1PY6A
RIGID  5     29
RIGID  33    58
RIGID  76    97
RIGID  101   123
RIGID  127   158
RIGID  161   187
RIGID  197   221

# 1PY7A
RIGID  4     25
RIGID  29    51
RIGID  55    86
RIGID  89    115

# 1RHZA
RIGID  2     20
RIGID  53    67
RIGID  79    107
RIGID  115   141
RIGID  147   165

# 1U19A
RIGID  2     32
RIGID  39    68
```

```
RIGID  75    108
RIGID  118   140
RIGID  168   193
RIGID  210   245
RIGID  253   277

# 2BG9A
RIGID  2     28
RIGID  33    60
RIGID  65    90

# 2BL2A
RIGID  2     36
RIGID  41    68
RIGID  75    112
RIGID  117   144

# 2BS2A
RIGID  2     33
RIGID  56    80
RIGID  101   129
RIGID  148   174
RIGID  182   216

# 2IC8A
RIGID  5     24
RIGID  58    79
RIGID  81    103
RIGID  111   127
RIGID  137   152
RIGID  161   180

# 2K73A
RIGID  12    36
RIGID  42    63
RIGID  68    96
RIGID  142   163

# 2KSFA
RIGID  7     25
RIGID  35    48
RIGID  55    66
RIGID  80    103

# 2KSYA
RIGID  3     28
RIGID  33    56
RIGID  70    91
RIGID  95    117
RIGID  122   152
RIGID  154   180
RIGID  190   222
```

```
# 2NR9A
RIGID  7      24
RIGID  60     81
RIGID  83     105
RIGID  112    129
RIGID  140    151
RIGID  163    189

# 2PNOA
RIGID  5      32
RIGID  43     73
RIGID  75     98
RIGID  101    129

# 2XQ2A
RIGID  2      21
RIGID  45     72
RIGID  74     101
RIGID  116    150
RIGID  154    172
RIGID  178    204
RIGID  247    267
RIGID  272    305
RIGID  340    378
RIGID  384    409
RIGID  415    439
RIGID  445    464
RIGID  471    493
RIGID  514    536
RIGID  538    564

# 2XUTA
RIGID  3      29
RIGID  39     69
RIGID  73     92
RIGID  97     125
RIGID  132    162
RIGID  165    187
RIGID  217    240
RIGID  246    270
RIGID  283    316
RIGID  326    345
RIGID  361    388
RIGID  397    421
RIGID  430    451
RIGID  468    487

# 2YVXA
RIGID  18     45
RIGID  61     84
RIGID  92     121
RIGID  127    155
RIGID  164    182
```

```
# 2ZW3A
RIGID  22    47
RIGID  72    105
RIGID  125   155
RIGID  184   215

# 3B60A
RIGID  15    45
RIGID  52    101
RIGID  112   154
RIGID  156   204
RIGID  214   262
RIGID  274   314

# 3GIAA
RIGID  8     35
RIGID  38    63
RIGID  82    114
RIGID  120   140
RIGID  142   170
RIGID  182   208
RIGID  215   244
RIGID  268   303
RIGID  319   335
RIGID  338   362
RIGID  371   395
RIGID  397   421

# 3HD6A
RIGID  9     29
RIGID  37    62
RIGID  67    92
RIGID  104   124
RIGID  130   155
RIGID  163   181
RIGID  197   219
RIGID  225   253
RIGID  261   281
RIGID  285   314
RIGID  322   342
RIGID  345   377

# 3HFXA
RIGID  4     20
RIGID  41    60
RIGID  77    106
RIGID  117   149
RIGID  176   209
RIGID  218   237
RIGID  244   266
RIGID  301   327
RIGID  333   364
RIGID  395   422
RIGID  436   455
```

```
RIGID  458   490

# 3KCUA
RIGID  3     28
RIGID  36    57
RIGID  79    107
RIGID  133   156
RIGID  160   177
RIGID  182   198
RIGID  219   250
# 3KJ6A
RIGID  2     24
RIGID  33    56
RIGID  75    102
RIGID  113   129
RIGID  174   197
RIGID  233   252
RIGID  274   292

# 3O0RB
RIGID  2     33
RIGID  44    75
RIGID  82    105
RIGID  132   154
RIGID  160   180
RIGID  186   221
RIGID  225   250
RIGID  258   285
RIGID  296   321
RIGID  333   362
RIGID  371   408
RIGID  417   448

# 3P5NA
RIGID  2     21
RIGID  51    69
RIGID  75    95
RIGID  100   122
RIGID  145   175

# 3SYOA
RIGID  16    45
RIGID  92    121

# 4A2NB
RIGID  3     30
RIGID  39    62
RIGID  75    99
RIGID  127   143
RIGID  146   175
```

Topology Broker setup files

```
# Using 1U19 as an example
# extended chain
CLAIMER MembraneTopologyClaimer
END_CLAIMER

# extended chain + EPR
CLAIMER MembraneTopologyClaimer
END_CLAIMER
CLAIMER ConstraintClaimer
FILE 1U19A.cst
END_CLAIMER

# RosettaTMH
CLAIMER MembraneTopologyClaimer
END_CLAIMER
CLAIMER TMHTopologySamplerClaimer
END_CLAIMER

# RosettaTMH + EPR)
CLAIMER MembraneTopologyClaimer
END_CLAIMER
CLAIMER TMHTopologySamplerClaimer
END_CLAIMER
CLAIMER ConstraintClaimer
FILE 1U19A.cst
END_CLAIMER
```

Simulating EPR distance restraints

```
# convert Rosetta native PDB file to BCL format
foreach pdb ( `cat pdb.ls` )
        sed -i '/CEN/d' ${pdb}.pdb
        bcl.exe protein:PDBConvert ${pdb}.pdb -bcl_pdb -output_prefix ${pdb}_
>& ${pdb}_bcl.log
end

# mutate.wts
bcl::storage::Table<double>    add_all add_single filter_aa_type_excl
filter_sse_size remove_single       swap distance_range_0 filter_exposure_0
weights      0     1     0     0     1     1     0     0

# score.wts
bcl::storage::Table<double> data_density aa_type_excl    seq_sep
data_set_size sse_connection   sse_size   sse_term    bipolar sse_center
triangulation_0 distance_range_0 exposure_0
weights      0     0     1     1     1     0     0     0     0     0
       10000  10000
```

```
# SSE pool for 1U19
bcl::assemble::SSEPool
HELIX    1   1 PRO A    2  GLN A   32  1                                  31
HELIX    2   2 PRO A   39  HIS A   68  1                                  30
HELIX    3   3 GLY A   74  VAL A  107  1                                  34
HELIX    4   4 GLU A  118  LEU A  140  1                                  23
HELIX    5   5 ASN A  168  GLN A  193  1                                  26
HELIX    6   6 THR A  210  THR A  245  1                                  36
HELIX    7   7 PRO A  253  MET A  276  1                                  24
END

# command line for restraint picking
bcl.exe restraint:OptimizeDataSetPairwise -fasta 1U19A.fasta -
pool_min_sse_lengths 3 0 -pool 1U19A_native.pool -distance_min_max 10 50 -
nc_limit 10 -ensembles pdb.ls -mc_number_iterations 10000 10000 -prefix 1U19A
-nmodels 10 -read_scores_optimization score.wts -read_mutates_optimization
mutate.wts -read_mutates_start mutate.wts -message_level Standard -
pymol_output -data_set_size_range 10 40 -data_set_size_fraction_of_sse_resis
0.2

# adding spin label uncertainty
bcl.exe SimulateDistanceRestraints -pdb 1U19A_bcl_format.pdb -
simulate_distance_restraints -output_file 1U19A_sim_epr.cst -
add_distance_uncertainty sl-cb_distances.histograms -restraint_list 1U19A.data
0 1 5 6 -random_seed -write_rosetta_mini_restraints
```

Restraint file format

```
# weighting EPR KBP by 10.0 and quadratic penalty by 1.0
# if have Gly in AtomPair, replace CB with 1HA or 2HA
AtomPair CB 67 CB 255 SCALARWEIGHTEDFUNC 1.0 SPLINE EPR_DISTANCE  28.9577 1.0 0.5
AtomPair CB 67 CB 255 SCALARWEIGHTEDFUNC 1.0 BOUNDED 16.9577 40.9577 1.0 NOE ;dist
```

Building loops on BCL and RosettaTMH files

```
# Loops file format – residues defined in loops are all residues not covered
by spanfile
# EXAMPLE:  1FX8A
LOOP   1      6
LOOP   26     37
LOOP   57     82
LOOP   102    141
LOOP   161    172
LOOP   192    229
LOOP   249    254

# convert BCL files to Rosetta files and make loops files for Rosetta loop
building
cd 1J4NA/pdbs
foreach pdb (`cat pdb.ls`)
```

```
        bcl.exe protein:PDBConvert ${pdb}.pdb -loop_file_rosetta CCD -
write_zero_coordinates -bcl_pdb Split -output_prefix ../loops/${pdb} >&
${pdb}_loops.log
end

# Build loops in Rosetta with options file (see fill_gaps.options) – from
Rosetta silent file
Rosetta/main/source/bin/loopmodel.mpi.linuxgccrelease -database
Rosetta/main/database/ @${DIR}/flags/fill_gaps.options -in:file:silent
${DIR}/${PDB}/test.out  -in:file:tags ${TAG000} -out:file:silent
${DIR}/${PDB}/test_loops.out -out:no_nstruct_label -out:file:scorefile
${DIR}/${PDB}/test_loops.sc

# Build loops in Rosetta with options file (see fill_gaps.options) – from a
PDB file
Rosetta/main/source/bin/loopmodel.mpi.linuxgccrelease -database
Rosetta/main/database/ @${DIR}/flags/fill_gaps.options -in:file:silent
${DIR}/${PDB}/test.pdb -out:pdb_gz –out:prefix test_ -out:no_nstruct_label -
out:file:scorefile ${DIR}/${PDB}/test_loops.sc
```

Restraint weights for folding in Rosetta

pdb	weight_new
1FX8A	10.5596
1IWGA	8.8708
1J4NA	8.3960
1KPLA	9.7660
1OCCA	9.6317
1OKCA	13.1527
1PV6A	8.6969
1PY6A	9.8134
1PY7A	8.0013
1RHZA	10.4518
1U19A	10.9355
2BG9A	6.8484
2BL2A	8.1081
2BS2A	10.9428
2IC8A	10.7754
2K73A	11.0377
2KSFA	9.1686
2KSYA	9.4516
2NR9A	11.0417
2PNOA	7.9325
2XQ2A	13.5716
2XUTA	12.7241
2YVXA	10.2313
2ZW3A	12.4219
3B60A	10.5270
3GIAA	12.2199
3HD6A	12.0958
3HFXA	14.0806
3KCUA	11.5959

```
3KJ6A  14.9617
3O0RB  11.9658
3P5NA  10.9224
3SYOA  10.9717
4A2NB  11.0417
```

Options files for de novo *folding*

```
# MembraneAbinitio
-in
       -file
              -native ${PDB}.pdb
              -fasta ${PDB}.fasta
              -frag3 aa${PDB}03_05.200_v1_3
              -frag9 aa${PDB}09_05.200_v1_3
              -spanfile ${PDB}.span
              -lipofile ${PDB}.lips4
-residues
       -patch_selectors CENTROID_HA
-score
       -find_neighbors_3dgrid
#      -use_membrane_rg  #### use this flag if using MP-specific RG score
-membrane
       -no_interpolate_Mpair
       -Menv_penalties
-abinitio
       -membrane
       -explicit_pdb_debug # if want to output at stages 0-4
       -rg_reweight ${RG_WEIGHT}
       -stage2_patch score_membrane_s2.wts_patch
       -stage3a_patch score_membrane_s3a.wts_patch
       -stage3b_patch score_membrane_s3b.wts_patch
       -stage4_patch score_membrane_s4.wts_patch
-evaluation
       -gdtmm
       -rmsd NATIVE _tm_sse ${PDB}_tm_sse_052814.txt
-out
       -output
       -file
              -output_virtual
              -silent_struct_type binary
-overwrite

# extended chain (and extended chain + EPR when CST_WEIGHT ≠ 0.0)
-in
       -file
              -native ${PDB}.pdb
              -fasta ${PDB}.fasta
              -frag3 aa${PDB}03_05.200_v1_3
              -frag9 aa${PDB}09_05.200_v1_3
              -spanfile ${PDB}.span
              -lipofile ${PDB}.lips4
```

```
-residues
	-patch_selectors CENTROID_HA
-broker
	-setup ${CSTFILE}.tpb #### will follow format of broker setup file above
-run
	-protocol broker
-score
	-find_neighbors_3dgrid
#	-use_membrane_rg #### use this flag if using MP-specific RG score
-membrane
	-no_interpolate_Mpair
	-Menv_penalties
-abinitio
	-membrane
	-explicit_pdb_debug # if want to output at stages 0-4
	-rg_reweight ${RG_WEIGHT}
	-stage2_patch score_membrane_s2.wts_patch
	-stage3a_patch score_membrane_s3a.wts_patch
	-stage3b_patch score_membrane_s3b.wts_patch
	-stage4_patch score_membrane_s4.wts_patch
-constraints
	-cst_file ${CSTFILE}
	-cst_weight ${CST_WEIGHT}
	-epr_distance
-fold_cst
	-force_minimize
	-seq_sep_stages 1.0 1.0 1.2
-evaluation
	-gdtmm
	-rmsd NATIVE _tm_sse ${PDB}_tm_sse_052814.txt
-out
	-output
	-file
		-output_virtual
		-silent_struct_type binary
-overwrite

# RosettaTMH (and RosettaTMH + EPR when cst_weight != 0.0)
-in
	-file
		-native ${PDB}.pdb
		-fasta ${PDB}.fasta
		-frag3 aa${PDB}03_05.200_v1_3
		-frag9 aa${PDB}09_05.200_v1_3
		-spanfile ${PDB}.span
		-lipofile ${PDB}.lips4
-residues
	-patch_selectors CENTROID_HA
-broker
	-setup ${CSTFILE}.tpb
	-large_frag_mover_stage1_weight 0.0
	-small_frag_mover_stage1_weight 0.0
	-rb_mover_stage1_weight 5.0
-run
```

```
	-protocol broker
-score
	-find_neighbors_3dgrid
#	-use_membrane_rg #### use this flag if using MP-specific RG score
-membrane
	-fixed_membrane
	-no_interpolate_Mpair
	-Menv_penalties
-abinitio
	-membrane
	-explicit_pdb_debug # if want to output at stages 0-4
	-rg_reweight ${RG_WEIGHT}
	-stage2_patch score_membrane_s2.wts_patch
	-stage3a_patch score_membrane_s3a.wts_patch
	-stage3b_patch score_membrane_s3b.wts_patch
	-stage4_patch score_membrane_s4.wts_patch
-constraints
	-cst_file ${CSTFILE}
	-cst_weight ${CST_WEIGHT}
	-epr_distance
-fold_cst
	-force_minimize
	-seq_sep_stages 1.0 1.0 1.2
-rigid
	-rotation 0.1
	-translation 0.5
-evaluation
	-gdtmm
	-rmsd NATIVE _tm_sse ${PDB}_tm_sse_052814.txt
-out
	-output
	-file
		-output_virtual
		-silent_struct_type binary
-overwrite


# loop building onto BCL and RosettaTMH models
-in
	-file
		-native ${DIR}/${PDB}/${PDB}.pdb
		-spanfile ${DIR}/${PDB}/${PDB}.span
		-lipofile ${DIR}/${PDB}/${PDB}.lips4
		-residue_type_set centroid
-chemical
	-patch_selectors CENTROID_HA
-score
	-find_neighbors_3dgrid
-evaluation
	-rmsd NATIVE _tm_sse ${DIR}/${PDB}/${PDB}_tm_sse_052814.txt
-membrane
	-no_interpolate_Mpair
	-Menv_penalties
-loops
```

```
	-loop_file ${DIR}/${PDB}/${PDB}.loops
	-frag_sizes 9 3 1
	-frag_files ${DIR}/${PDB}/aa${PDB}09_05.200_v1_3
${DIR}/${PDB}/aa${PDB}03_05.200_v1_3 none
	-remodel quick_ccd
	-cen_weights score_membrane
	-cen_patch ${DIR}/score_membrane_s4.wts_patch
-out
	-output
	-no_nstruct_label
	-nstruct 1
	-file
		-silent_struct_type binary # only if outputting silent files, not
pdb or pdb_gz files
		-residue_type_set centroid
-overwrite
```

Score patches

```
# score_membrane_s2.wts_patch
pair = 0.0
Mpair = 1.0
env = 0.0
Menv = 2.019
cbeta = 0.0
Mcbeta = 0.0
Menv_non_helix = 2.019
Menv_termini = 2.019
Menv_tm_proj = 2.019
Mlipo = 1.0

# score_membrane_s3a.wts_patch
pair = 0.0
Mpair = 1.0
env = 0.0
Menv = 2.019
cbeta = 0.0
Mcbeta = 0.5
Menv_non_helix = 2.019
Menv_termini = 2.019
Menv_tm_proj = 2.019
Mlipo = 1.0

# score_membrane_s3b.wts_patch
pair = 0.0
Mpair = 1.0
env = 0.0
Menv = 2.019
cbeta = 0.0
Mcbeta = 0.5
Menv_non_helix = 2.019
Menv_termini = 2.019
```

```
Menv_tm_proj = 2.019
Mlipo = 1.0

# score_membrane_s4.wts_patch
pair = 0.0
Mpair = 1.0
env = 0.0
Menv = 2.019
cbeta = 0.0
Mcbeta = 2.5
Menv_non_helix = 2.019
Menv_termini = 2.019
Menv_tm_proj = 2.019
Mlipo = 1.0
```

Command lines for folding

MembraneAbinitio

```
Rosetta/main/rosetta_source/bin/membrane_abinitio2.static.linuxgccrelease -
database /Rosetta/main/rosetta_database/ @${FLAGS} -out::nstruct ${NSTRUCT} -
out:file:silent ${OUTFILE} -out:sf ${OUTFILE}.sc
```

Extended chain (and Extended chain + EPR if CST_WEIGHT ≠ 0.0)

```
/Rosetta/main/rosetta_source/bin/minirosetta.mpi.linuxgccrelease -database
/Rosetta/main/rosetta_database/ @${FLAGS} -out::nstruct ${NSTRUCT} -
out:file:silent ${OUTFILE} -out:file:scorefile ${OUTFILE}.sc
```

RosettaTMH (and RosettaTMH + EPR if CST_WEIGHT != 0.0)

```
/Rosetta/main/rosetta_source/bin/minirosetta.mpi.linuxgccrelease -database
/Rosetta/main/rosetta_database/ @${FLAGS} -out::nstruct ${NSTRUCT} -
out:file:silent ${OUTFILE} -out:file:scorefile ${OUTFILE}.sc
```

Weighting schemes tested

RG score weights tested

Both default and MP-specific RG scores were weighted by 0.0, 0.01, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, and 2.00 when testing effect of MP-specific RG score.

EPR restraint weights tested

		Quadratic Penalty						
		0.0	1.0	10.0	20.0	30.0	40.0	50.0
EPR KBP	0.0	✓	✓	✓	✓	✓	✓	✓
	1.0	✓	✓	✓	✓	✓	✓	✓
	10.0	✓	✓	✓	✓	✓	✓	✓
	20.0	✓	✓	✓	✓	✓	✓	✓
	30.0	✓	✓	✓	✓	✓	✓	✓
	40.0	✓	✓	✓	✓	✓	✓	✓
	50.0	✓	✓	✓	✓	✓	✓	✓

Analysis of results

BCL models analysis

```
# score BCL pdbs for comparison with Rosetta models (after loop building)
bcl.exe protein:Score -pdblist bcl_pdb.ls -native 2K73A_bcl.pdb -
score_table_write 2K73A_bcl_scores_071014.tbl.tmp -weight_set  refinement.tbl
-membrane 20 10 2.5 -tm_helices 2K73A_native.pool -pool 2K73A_native.pool -
sspred JUFO9D OCTOPUS -sequence_data ./ 2K73 >& score.log

# Computing RMSD100SSE in the BCL
bcl.exe protein:Compare -reference_pdb 2K73A_bcl.pdb -pdb_list bcl_pdb.ls -
quality RMSD -atoms CA -specify_residues 2K73A_bcl_res.ls >&
2K73A_bcl_rmsd100_062114.log

#2K73A_bcl_res.ls format
'A' 12
'A' 13
'A' 14
'A' 15
'A' 16
'A' 17 ... for all residues for which to compute RMSD


# Refinement.tbl (for BCL scoring)
bcl::storage::Table<double>     aaclash aadist  aaneigh aaneigh_ent     loop
loop_closure_gradient   rgyr    sseclash        ssepack_fr      strand_fr
co_score        ss_OCTOPUS      ss_OCTOPUS_ent  ss_OCTOPUS_env  ss_JUFO9D
ss_JUFO9D_ent   ss_JUFO9D_env   ssealign        mp_helix_topology
weights 500     0.35    50      50.0    10.0    50000   5.0     500     8.0
20      0.5     20.0            20.0    20      5.0     5.0     5.0       8
500

# example native pool file – based on DSSP for the PDB file – see
*rms_tm_sse_052814.txt examples above.  These are the same residue definitions
```

```
bcl::assemble::SSEPool
HELIX    1   1 LEU A    2  VAL A   29  1                                    28
HELIX    2   2 GLN A   36  SER A   58  1                                    23
HELIX    3   3 PRO A   64  PHE A   73  1                                    10
HELIX    4   4 LYS A   78  HIS A  114  1                                    37
HELIX    5   5 VAL A  121  PHE A  130  1                                    10
HELIX    6   6 PHE A  140  THR A  162  1                                    23
HELIX    7   7 LEU A  173  LEU A  192  1                                    20
HELIX    8   8 PRO A  199  ALA A  212  1                                    14
HELIX    9   9 TYR A  227  HIS A  253  1                                    27
END
```

Generation of $RMSD_{100}$SSE histograms

```
# format for files for input into rmsd_to_rmsd100.py
# doesn't matter how many fields are between field 1 and description, but
"SCORE:" must be first, and "description" last, also total score is "score"
SCORE: score  rms_tm_sse   file   description
SCORE: 103.718      14.536 1FX8A_s01_b001_0000 S_0001
SCORE: 51427.301    17.6354      1FX8A_s01_b001_0000 S_0002
SCORE: 51299.817    15.8242      1FX8A_s01_b001_0000 S_0003
SCORE: 52004.468    18.2978      1FX8A_s01_b001_0000 S_0004
SCORE: 51368.199    14.9345      1FX8A_s01_b001_0000 S_0005
SCORE: 297.978      14.0197      1FX8A_s01_b001_0000 S_0006
SCORE: 51497.388    15.4662      1FX8A_s01_b001_0000 S_0007
SCORE: 180.629      13.0461      1FX8A_s01_b001_0000 S_0008
SCORE: 516.341      16.3426      1FX8A_s01_b001_0000 S_0009

# convert Rosetta-computed RMSD values to RMSD100
./rmsd_to_rmsd100.py --membrane --silent=${file} -n ${nres} --
outfile=${file}.rmsdSSE100 --rms_tag=tm_sse

# Generate histograms and summary
perl ~/scripts/Smbins_RMSD_dist_from_score.pl ${file}.rmsdSSE100 5 | awk
'{print($2"\t"$4)}' | head -n21 > ${file}.rmsdSSE100.txt
```

Calculating enrichment

```
Usage: compute_overall_performance.py [options]  # parses file as field 0 =
pdb and field 4 = weight - 1FX8A_tmh_s01_b001_0009_scores_rms_tm_sse_072214.sc
Options:
  -h, --help            show this help message and exit
  --filelist=FILELIST   filelist
  --metric=METRIC       header of column wanting to average
  --score_fraction=SCORE_FRACTION
                        what fraction of models do you consider for TP etc
  --quality_fraction=QUALITY_FRACTION
                        what fraction of models do you consider for TP etc
  --p_ratio=RATIO       (p+n)/p. this sets your max enrichment
  --enrichment_output=ENRICHMENT_OUTPUT
                        file to output enrichment values for each pdb and
```

```
                      weight
  --outfile=OUTFILE   outfile

./compute_overall_performance.py --filelist list --metric rms_tm_sse --
score_fraction 0.1 --quality_fraction 0.1 --p_ratio 10 --enrichment_output
test.enrch --outfile test.out
```

Calculate loops fulfillment

```
USAGE:  <pdb_filename> <restraint_filename> <# restraints>
# restraint min is 0.00 restraint max = 3.8 * (nres-1)
restraint file format: <chain> <atom> <res> <atom> <res> <min> <max>
```

Calculating contact order

Downloaded script from Baker laboratory website (http://depts.washington.edu/bakerpg/contact_order/)

```
# options:  -c = cutoff, default is 6; -a = absolute contact order
./contactOrder.pl –c 8 –a 1U19A.pdb
```

APPENDIX F

PROTOCOL CAPTURE FOR APPENDIX A: LIGAND-MIMICKING RECEPTOR VARIANT DISCLOSES BINDING AND ACTIVATION MODE OF PROLACTIN RELEASING PEPTIDE

This appendix contains the protocol capture for the modeling work published in (Rathmann*, Lindner*, DeLuca*, Kaufmann, Meiler, and Beck-Sickinger, 2012), some of which is found in the manuscript's Supplemental Information. *These authors contributed equally. Further details are also available in Appendix A, and more detailed information on comparative modeling in Rosetta can be found in reference (252).

Computational details

All models were generated by independent simulations using Vanderbilt University's Center for Structural Biology computing cluster and the university's Advanced Computing Center for Research and Education (ACCRE). Computations were performed on a combination of AMD Opteron and Intel Nehalem processor nodes. All Rosetta-related protocols were conducted using Rosetta version 3.4.

Input files

FASTA file for making fragments for loop building

```
> PrRPR residues 58-347
QLKGLIVLLYSVVVVVGLVGNCLLVLVIARVRRLHNVTNFLIGNLALSDVLMCTACVPLTLAYAFEPRGW
VFGGGLCHLVFFLQPVTVYVSVFTLTTIAVDRYVVLVHPLRRRISLRLSAYAVLAIWALSAVLALPAAVH
TYHVELKPHDVRLCEEFWGSQERQRQLYAWGLLLVTYLLPLLVILLSYVRVSVKLRNRVVPGCVTQSQAD
WDRARRRRTFCLLVVVVVVVFAVCWLPLHVFNLLRDLDPHAIDPYAFGLVQLLCHWLAMSSACYNPFIYAW
LHDSFREELRKLLV
```

Spanfile required for RosettaMembrane

```
TM region prediction for PrRPR_112210.octopus predicted using OCTOPUS
7 294
antiparallel
n2c
   4    24     4    24
  40    60    40    60
  79    99    79    99
 119   139   119   139
 169   189   169   189
 220   240   220   240
 257   277   257   277
```

Disulfide file (defines disulfide bond that want to maintain)

```
# bridge between C134 (in TM3 close to ECL) C211 (ECL2)
77 154
```

XML file for docking

The XML file gives the parser, or RosettaScripts, instructions for how to run the protocol. The protocol we used for this study is as follows:

1. Import constraints $D^{6.59}$-R^{19}, $E^{5.26}$-R^{19}, $W^{5.28}$-R^{19}, $Y^{5.38}$-R^{19} (last three have 50% confidence)
2. Docking perturbation (4Å translation, 10 degree rotation)
3. Fast relax with only 1 iteration
4. Filter $D^{6.59}$-R^{19} with 100% confidence
5. Rebuild EL2
6. Rebuild EL3
7. Rebuild EL1
8. Full fast relax
9. Filter by disulfide linkage with 100% confidence (residues 134 and 211)
10. Filter by $D^{6.59}$-R^{19}

```
# XML file
<dock_design>
       <SCOREFXNS> #defines non-standard score functions, weight
       ROSETTAMEMBRANE scores by 10x
              <mem_cen_cst weights=score_membrane>
                     <Reweight scoretype=atom_pair_constraint weight=10/>
              </mem_cen_cst>
              <mem_fa_cst  weights=membrane_highres_Menv_smooth>
                     <Reweight scoretype=atom_pair_constraint weight=10/>
```

```
            </mem_fa_cst>
      </SCOREFXNS>
      <FILTERS>
            <DisulfideFilter name=disulfide targets=77,154 confidence=1.0/>
            <ResidueDistance name=D659_R19 res1_res_num=245 res2_res_num=306
      distance=10.0 confidence=1.0/>
      </FILTERS>
      <TASKOPERATIONS>
            <InitializeFromCommandline name=ifcl/>
            <RestrictToRepacking name=rtrp/>
      </TASKOPERATIONS>
      <MOVERS>
      <Docking name=dock score_low=mem_cen_cst score_high=mem_fa_cst
fullatom=1 local_refine=1 optimize_fold_tree=1 conserve_foldtree=0 design=0
task_operations=ifcl/>
      <LoopRemodel name=loop3 loop_start_res_num=246 loop_end_res_num=258
hurry=0 protocol=ccd perturb_score=mem_cen_cst refine_score=mem_fa_cst
perturb=1 refine=1 design=0 />
      <LoopRemodel name=loop2 loop_start_res_num=139 loop_end_res_num=169
hurry=0 protocol=ccd perturb_score=mem_cen_cst refine_score=mem_fa_cst
perturb=1 refine=1 design=0 />
      <LoopRemodel name=loop1 loop_start_res_num=65 loop_end_res_num=74
hurry=0 protocol=ccd perturb_score=mem_cen_cst refine_score=mem_fa_cst
perturb=1 refine=1 design=0 />
      <FastRelax name=fastrlx_all repeats=1 scorefxn=mem_fa_cst />
      <FastRelax name=fastrlx_r1 repeats=1 scorefxn=mem_fa_cst />
      <PackRotamersMover name=repack scorefxn=mem_fa_cst
task_operations=rtrp/>
      <ConstraintSetMover name=fa_cst cst_file=dock_fa.cst />
      <ConstraintSetMover name=lowres_cst cst_file=dock.cst />
      </MOVERS>
      <APPLY_TO_POSE>
      </APPLY_TO_POSE>
      <PROTOCOLS>
            <Add mover_name=fa_cst/>
            <Add mover_name=lowres_cst/>
            <Add mover_name=dock/>
            <Add mover_name=fastrlx_r1/>
            <Add filter_name=D659_R19/>
            <Add mover_name=loop2/>
            <Add mover_name=loop3/>
            <Add mover_name=loop1/>
            <Add mover_name=fastrlx_all/>
            <Add filter_name=disulfide/>
            <Add filter_name=D659_R19/>
      </PROTOCOLS>
</dock_design>
```

Constraints file for docking

```
AtomPair CB 306 CB 245 BOUNDED 0.00 10.0 1.0 NOE loose
AtomPair CB 306 CB 156 BOUNDED 0.00 10.0 1.0 NOE loose
```

```
AtomPair CB 306 CB 158 BOUNDED 0.00 10.0 1.0 NOE loose
AtomPair CB 306 CB 168 BOUNDED 0.00 10.0 1.0 NOE loose
```

Constraints file for full-atom refinement

```
AtomPair CB 306 CB 245 BOUNDED 0.00 10.0 1.0 NOE loose
AtomPair CB 306 CB 156 BOUNDED 0.00 10.0 1.0 NOE loose
AtomPair CB 306 CB 158 BOUNDED 0.00 10.0 1.0 NOE loose
AtomPair CB 306 CB 168 BOUNDED 0.00 10.0 1.0 NOE loose
```

Options file for relaxing threaded models

```
-relax
	-membrane
#	-default_repeats
#	-fastrelax_repeats 2
-in
	-path
		-database /blue/meilerlab/home/hirstsj/mini/minirosetta_database
	-file
		-l
/blue/meilerlab/home/hirstsj/GPCRs/PrRPR/PrRPm_rlx_initial_1.ls
		-spanfile
/blue/meilerlab/home/hirstsj/GPCRs/PrRPR/PrRPR_112210.span
		-fullatom
-score
	-weights
/blue/meilerlab/home/hirstsj/mini/minirosetta_database/scoring/weights/membran
e_highres_Menv_smooth.wts
-membrane
	-normal_cycles 100
	-normal_mag 15
	-center_mag 2
-out
	-output
	-nstruct 1000
	-file
		-silent
/blue/meilerlab/home/hirstsj/GPCRs/PrRPR/mutate_templates/rlx_models/PrRPR_rlx
_112310_1.out
		-silent_struct_type binary
		-scorefile
/blue/meilerlab/home/hirstsj/GPCRs/PrRPR/mutate_templates/PrRPR_rlx_112310_1.f
asc
		-fullatom
-overwrite
```

Options file for building loops into threaded model

```
##Build Initial Loops with Fragments
-loops
```

```
	-timer #output time spent in seconds for each loop modeling job
	-fast #reduce the number of cycles used during loop building. remove for
production runs.
	-frag_sizes 9 3 1 #This option is paired with the option -
loops:frag_files- indicates fragment sizes
	-frag_files
/blue/meilerlab/home/hirstsj/GPCRs/PrRPR/aaPrRPm09_05.200_v1_3
/blue/meilerlab/home/hirstsj/GPCRs/PrRPR/aaPrRPm03_05.200_v1_3 none
	-fa_input #input structures are in full atom format
#	-input_pdb mutate_templates/PrRPR_3EML_renumber.pdb
#	-loop_file build_initial.loops #loop definition file
	-relax fastrelax
	-build_initial #build missing density
	-ccd_closure
	-random_loop
	-remodel quick_ccd
-packing #rotamer library flags
	-ex1aro
	-ex1
	-ex2
	-repack_only
-in
	-path
		-database /blue/meilerlab/home/hirstsj/mini/minirosetta_database/
	-fix_disulf /blue/meilerlab/home/hirstsj/GPCRs/PrRPR/PrRPm_disulf.txt
#read disulfide connectivity information
	-file
		-fullatom
		-psipred_ss2
/blue/meilerlab/home/hirstsj/GPCRs/PrRPR/PrRPm.psipred_ss2
		-spanfile
/blue/meilerlab/home/hirstsj/GPCRs/PrRPR/PrRPR_112210.span
-out
#	-prefix PrRPm_3EML_initial_112310_
	-output
	-pdb
	-overwrite
	-nstruct 100 #recommended 1000
	-file
		-fullatom
#		-silent PrRPm_initial_112310.out
		-silent_struct_type binary
-max_inner_cycles 30
-outer_cycles 1
-membrane
	-normal_cycles 100
	-normal_mag 15
	-center_mag 2
-score
	-weights
/blue/meilerlab/home/hirstsj/mini/minirosetta_database/scoring/weights/membran
e_highres_Menv_smooth.wts
-overwrite
```

Options file for peptide docking and loop building

```
# This file contains 2 chains (A and B) with chain B being the peptide. The
receptor and peptide were renumbered starting from 1
-s input.pdb
-nstruct 500 # number of models to build
-out:output  # output files
-out:file:fullatom  # output in full atom detail
-out:silentoutput_file.out# structure is stored in internal coordinates
instead of Cartesian coordinates
-out:silent_struct_type binary   # binary silent output is more compressed and
extracted more robustly
-out:scorefile scores.fasc# output a scorefile, which doesn't store
coordinates, only scores
-jd2:ntrials 5      # use the new job distributor
-parser:protocol parser_protocol.xml   # The protocol we use is actually run
by RosettaScripts, aka the parser
-docking:dock_pert 4 10   # during docking perturbation, allow for 4Å
translation and 10 degree rotation
-residues:patch_selectors CTERM_AMIDATION     # use this option if you want to
amidate the C-terminus of the peptide
-max_inner_cycles 30
-outer_cycles 1
# ROSETTAMEMBRANE options
-membrane:normal_cycles 100 # number of cycles to search for membrane normal
-membrane:normal_mag 15   # options for the angle allowance of normal and
center search
-membrane:center_mag 2
-in:file:psipred_ss2 secondary_structure.psipred_ss2 # secondary structure
prediction input in psipred format
-in:file:spanfile tmh.span# membrane spanning regions of receptor predicted by
OCTOPUS and in spanfile format
-in:file:fix_disulf disulf.txt   # file containing residue pairs between which
there is a disulfide bond # loop building options
-loops:timer
-loops:fast  # reduce number of loop building trials, or cycles
-loops:frag_sizes 9 3 1   # fragment sizes used for CCD loop building, but not
using 1mers
-loops:frag_files aaTest_09_05.200_v1_3 aaTest_03_05.200_v1_3 none
-loops:fa_input     # fullatom input
-loops:relaxfastrelax     # do a "fast" relax, which consists of iterative
rounds of side-chain repacking and all atom minimization
-loops:remodel quick_ccd
-packing:ex1 # include extra rotamers for side-chain repacking
-packing:ex2
-packing:repack_only
-packing:linmem_ig 10
-overwrite   # overwrite existing output files having the same name
```

Command lines

Fragment generation for loop building

```
make_fragments.pl –id <fasta_id> input.fasta
```

Threading for comparative modeling

```
# Example of template is 2VT4.  Other templates used were 2RH1, 3CAP, 1U19,
3DQB, and 3EML.
awk -v loop_file_prefix=PrRPR_2VT4 -v generate_loop_file="top" -v templatepdb=
2vt4A.pdb -v pdb_chain=A -v
blc_alignmentfile=NPY_RFamide_classA_profile_profile.blc -v tempseq=18 -v
alignseq=11 -f /sb/meiler/scripts/kaufmann_awk/awk_library.txt -f
/blue/meilerlab/apps/scripts/create_template_from_blc.awk -f
/sb/meiler/scripts/kaufmann_awk/aa_transform.txt > PrRPR_2VT4.pdb
```

Peptide docking and loop building

```
mpiexec /bin/rosetta_scripts.mpistatic.linuxgccrelease -database
rosetta_database/ @dock_PrRP.options -s $START_PDB -out:file:silent $OUTFILE -
out:file:scorefile $SCOREFILE
```

Analysis and selection of models

Computing distance matrix for clustering using BCL

```
bcl.exe Quality -quality RMSD -atom_list CA -pdb_list pdb.ls -aaclass AACaCb
```

Clustering with the BCL (436)

```
# Had tried clustering at 3, 3.5, 4, and 5A cutoffs and decided on 3.7A

/bcl.exe Cluster -distance_input_file filter_distRMSD.txt -input_format
TableLowerTriangle -output_format Rows Centers -output_file cluster_cutoff3-
7.txt -linkage Average -distance_definition less -output_pymol 1000 25 100
10000 10 dendogram.py -remove_nodes_below_size 100 -
remove_internally_similar_nodes 4 -pymol_label_output_string -
pymol_scale_node_with_size
```

Find the geometric centers of the leaf clusters

```
awk '{if($14==1)print}' Centers.txt | sort -nrk10 > leaves_centers.txt
```

Find if top scoring models in leaf clusters

```
foreach pdb ( `cat top10_by_score.ls` )
       grep $pdb cluster_co3.7.Rows.txt | grep "Leaf : 1"
end
```

BIBLIOGRAPHY

1. Berman, HM, Westbrook, J, Feng, Z, Gilliland, G, Bhat, TN, Weissig, H, Shindyalov, IN, Bourne, PE (2002) The Protein Data Bank. *Nucl Acids Res* 58:899–907.

2. Congreve M, Murray CW, Blundell TL (2005) Keynote review: Structural biology and drug discovery. *Drug Discov Today* 10:895–907.

3. Blundell TL (1996) Structure-based drug design. *Nature* 384:23–26–26.

4. Sham, HL, Kempf, DJ, Molla, A, Marsh, KC, Kumar, GN, Chen, C-M, Kati, W, Stewart, K, Lal, R, Hsu, A, Betebenner, D, Korneyeva, M, Vasavanonda, S, McDonald, E, Saldivar, A, Wideburg, N, Chen, X, Niu, P, Park, C, Jayanti, V, Grabowski, B, Granneman, GR, Sun, E, Japour, AJ, Leonard, JM, Plattner, JJ, Norbeck, DW (1998) ABT-378, a highly potent inhibitor of the human immunodeficiency virus protease. *Antimicrob Agents Chem* 42:3218–3224.

5. Kim, EE, Baker, CT, Dwyer, MD, Murcko, MA, Rao, BG, Tung, RD, Navia, MA (1995) Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *J Am Chem Soc* 117:1181–1182.

6. Kaldor, SW, Kalish, VJ, Davies, JF, Shetty, BV, Fritz, JE, Appelt, K, Burgess, JA, Campanale, KM, Chirgadze, NY, Clawson, DK, Dressman, BA, Hatch, SD, Khalil, DA, Kosa, MB, Lubbehusen, PP, Muesing, MA, Patick, AK, Reich, SH, Su, KS, Tatlock, JH (1997) Viracept (Nelfinavir Mesylate, AG1343): A Potent, Orally Bioavailable Inhibitor of HIV-1 Protease. *J Med Chem* 40:3979–3985.

7. Itzstein, M, Wu, W-Y, Kok, GB, Pegg, MS, Dyason, JC, Jin, B, Phan, TV, Smythe, ML, White, HF, Oliver, SW, Colman, PM, Varghese, JN, Ryan, DM, Woods, JM, Bethell, RC, Hotham, VJ, Cameron, JM, Penn, CR (1993) Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* 363:418–423.

8. Kim, CU, Lew, W, Williams, MA, Wu, H, Zhang, L, Chen, X, Escarpe, PA, Mendel, DB, Laver, WG, Stevens, RC (1998) Structure−activity relationship studies of novel carbocyclic influenza neuraminidase inhibitors. *J Med Chem* 41:2451–2460.

9. Pollack, VA, Savage, DM, Baker, DA, Tsaparikos, KE, Sloan, DE, Moyer, JD, Barbacci, EG, Pustilnik, LR, Smolarek, TA, Davis, JA, Vaidya, MP, Arnold, LD, Doty, JL, Iwata, KK, Morin, MJ (1999) Inhibition of epidermal growth factor receptor-associated tyrosine phosphorylation in human carcinomas with CP-358,774: Dynamics of receptor inhibition *in situ* and antitumor effects in athymic mice. *J Pharmacol Exp Therapeutics* 291:739–748.

10. Bagal, SK, Brown, AD, Cox, PJ, Omoto, K, Owen, RM, Pryde, DC, Sidders, B, Skerratt, SE, Stevens, EB, Storer, RI, Swain, NA (2013) Ion channels as therapeutic targets: A drug discovery perspective. *J Med Chem* 56:593–624.

11. Abramson J, Wright EM (2009) Structure and function of Na^+-symporters with inverted repeats. *Curr Opin Struct Biol* 19:425–432.

12. Deupi X (2012) Quantification of structural distortions in the transmembrane helices of GPCRs. *Methods Mol Biol* 914:219–235.

13. Kroeze WK, Sheffler DJ, Roth BL (2003) G-protein-coupled receptors at a glance. *J Cell Sci* 116:4867–4869.

14. Nickols HH, Conn PJ (2014) Development of allosteric modulators of GPCRs for treatment of CNS disorders. *Neurobiol Dis* 61:55–71.

15. Zhou, Z, Zhen, J, Karpowich, NK, Law, CJ, Reith, MEA, Wang, D-N (2009) Antidepressant specificity of serotonin transporter suggested by three LeuT-SSRI structures. *Nat Struct Mol Biol* 16:652–657.

16. Combs, S, Kaufmann, K, Field, JR, Blakely, RD, Meiler, J (2011) Y95 and E444 interaction required for high-affinity *S*-citalopram binding in the human serotonin transporter. *ACS Chem Neurosci* 2:75-81.

17. Manepalli S, Geffert LM, Surratt CK, Madura JD (2011) Discovery of novel selective serotonin reuptake inhibitors through development of a protein-based pharmacophore. *J Chem Info Model* 51:2417–2426.

18. Walther C, Mörl K, Beck-Sickinger AG (2011) Neuropeptide Y receptors: Ligand binding and trafficking suggest novel approaches in drug development. *J Pept Sci* 17:233–246.

19. Rathmann, D, Lindner, D, Deluca, SH, Kaufmann, KW, Meiler, J, Beck-Sickinger, AG (2012) Ligand-mimicking receptor variant discloses binding and activation mode of prolactin-releasing peptide. *J Biol Chem* 287:32181–32194.

20. Watanabe, A, Okuno, S, Okano, M, Jordan, S, Aihara, K, Watanabe, TK, Yamasaki, Y, Kitagawa, H, Sugawara, K, Kato, S (2007) Altered emotional behaviors in the diabetes mellitus OLETF type 1 congenic rat. *Brain Res* 1178:114–124.

21. Cant N, Pollock N, Ford RC (2014) CFTR structure and cystic fibrosis. *Intl J Biochem Cell Biol* 52:15–25.

22. Leth-Larsen, R, Lund, R, Hansen, HV, Laenkholm, AV, Tarin, D, Jensen, ON, Ditzel, HJ (2009) Metastasis-related plasma membrane proteins of human breast cancer cells identified by comparative quantitative mass spectrometry. *Mol Cell*

Proteomics 8:1436–1449.

23. Stevens, RC, Cherezov, V, Katritch, V, Abagyan, R, Kuhn, P, Rosen, H, Wüthrich, K (2012) The GPCR Network: A large-scale collaboration to determine human GPCR structure and function. *Nat Rev Drug Discov* 12:25–34.

24. Lagerström MC, Schiöth HB (2008) Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov* 7:339–357.

25. Hopkins AL, Groom CR (2002) The druggable genome. *Nat Rev Drug Discov* 1:727–730.

26. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? *Nat Rev Drug Discov* 5:993–996.

27. Katritch V, Cherezov V, Stevens RC (2012) Structure-function of the G- protein coupled receptor superfamily. *Ann Rev Pharmacol Toxicol* 53:531–556.

28. Hinuma, S, Habata, Y, Fujii, R, Kawamata, Y, Hosoya, M, Fukusumi, S, Kitada, C, Masuo, Y, Asano, T, Matsumoto, H, Sekiguchi, M, Kurokawa, T, Nishimura, O, Onda, H, Fujino, M (1998) A prolactin-releasing peptide in the brain. *Nature* 393:272–276.

29. Bjursell M, Lennerås M, Göransson M, Elmgren A, Bohlooly-Y M (2007) GPR10 deficiency in mice results in altered energy expenditure and obesity. *Biochem Biophys Res Commun* 363:633–638.

30. Roland, BL, Sutton, SW, Wilson, SJ, Luo, L, Pyati, J, Huvar, R, Erlander, MG, Lovenberg, TW (1999) Anatomical distribution of prolactin-releasing peptide and its receptor suggests additional functions in the central nervous system and periphery. *Endocrinology* 140:5736–5745.

31. Fujii, R, Fukusumi, S, Hosoya, M, Kawamata, Y, Habata, Y, Hinuma, S, Sekiguchi, M, Kitada, C, Kurokawa, T, Nishimura, O, Onda, H, Sumino, Y, Fujino, M (1999) Tissue distribution of prolactin-releasing peptide (PrRP) and its receptor. *Regul Pept* 83:1–10.

32. Engstrom M (2003) Prolactin releasing peptide has high affinity and efficacy at neuropeptide FF2 receptors. *J Pharmacol Exp Therapeutics* 305:825–832.

33. Ellacott, KLJ (2002) PRL-releasing peptide interacts with leptin to reduce food intake and body weight. *Endocrinology* 143:368–374.

34. Boyle, RG, Downham, R, Ganguly, T, Humphries, J, Smith, J, Travers, S (2005) Structure-activity studies on prolactin-releasing peptide (PrRP). Analogues of PrRP-(19-31)-peptide. *J Pept Sci* 11:161–165.

35. D'Ursi, AM, Albrizio, S, Di Fenza, A, Crescenzi, O, Carotenuto, A, Picone, D, Novellino, E, Rovero, P (2002) Structural studies on Hgr3 orphan receptor ligand prolactin-releasing peptide. *J Med Chem* 45:5483–5491.

36. Saini, V, Staren, DM, Ziarek, JJ, Nashaat, ZN, Campbell, EM, Volkman, BF, Marchese, A, Majetschak, M (2011) The CXC chemokine receptor 4 ligands ubiquitin and stromal cell-derived factor-1 function through distinct receptor interactions. *J Biol Chem* 286:33466-33477.

37. Deluca SH, Rathmann D, Beck-Sickinger AG, Meiler J (2013) The activity of prolactin releasing peptide correlates with its helicity. *Biopolymers* 99:314–325.

38. Ballesteros JA, Weinstein H (1995) Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein coupled receptors. *Methods Neurosci* 25:366–428.

39. Rathmann D, Pedragosa-Badia X, Beck-Sickinger AG (2013) In vitro modification of substituted cysteines as tool to study receptor functionality and structure–activity relationships. *Anal Biochem* 439:173–183.

40. Howard, AD, Feighner, SD, Cully, DF, Arena, JP, Liberator, PA, Rosenblum, CI, Hamelin, M, Hreniuk, DL, Palyha, OC, Anderson, J, Paress, PS, Diaz, C, Chou, M, Liu, KK, Mckee, KK, Pong, SS, Chaung, LY, Elbrecht, A, Dashkevicz, M, Heavens, R, Rigby, M, Sirinathsinghji, DJ, Dean, DC, Melillo, DG, Patchett, AA, Nargund, R, Griffin, PR, DeMartino, JA, Gupta, SK, Schaeffer, JM, Smith, RG, Van Der Ploeg, LH (1996) A receptor in pituitary and hypothalamus that functions in growth hormone release. *Science* 273:974–977.

41. Kojima, M, Hosoda, H, Date, Y, Nakazato, M, Matsuo, H, Kangawa, K (1999) Ghrelin is a growth-hormone-releasing acylated peptide from stomach. *Nature* 402:656–660.

42. Guan, XM, Yu, H, Palyha, OC, Mckee, KK, Feighner, SD, Sirinathsinghji, DJ, Smith, RG, Van Der Ploeg, LH, Howard, AD (1997) Distribution of mRNA encoding the growth hormone secretagogue receptor in brain and peripheral tissues. *Mol Brain Res* 48:23–29.

43. Gnanapavan, S, Kola, B, Bustin, SA, Morris, DG, McGee, P, Fairclough, P, Bhattacharya, S, Carpenter, R, Grossman, AB, Korbonits, M (2002) The tissue distribution of the mRNA of ghrelin and subtypes of its receptor, GHS-R, in humans. *J Clin Endocrinol Metabol* 87:2988–2988.

44. Tannenbaum GS, Bowers CY (2001) Interactions of growth hormone secretagogues and growth hormone-releasing hormone/somatostatin. *Endocrine* 14:021–021.

45. van der Lely AJ, Tschöp M, Heiman ML, Ghigo E (2004) Biological,

physiological, pathophysiological, and pharmacological aspects of ghrelin. *Endocr Rev* 25:426–457.

46. Dickson, SL, Egecioglu, E, Landgren, S, Skibicka, KP, Engel, JA, Jerlhag, E (2011) The role of the central ghrelin system in reward from food and chemical drugs. *Mol Cell Endocrinol* 340:80–87.

47. Jerlhag, E, Egecioglu, E, Landgren, S, Salomé, N, Heilig, M, Moechars, D, Datta, R, Perrissoud, D, Dickson, SL, Engel, JA (2009) Requirement of central ghrelin signaling for alcohol reward. *Proc Natl Acad Sci USA* 106:11318–11323.

48. Atcha, Z, Chen, W-S, Ong, AB, Wong, F-K, Neo, A, Browne, ER, Witherington, J, Pemberton, DJ (2009) Cognitive enhancing effects of ghrelin receptor agonists. *Psychopharmacology* 206:415–427.

49. Abizaid, A, Liu, Z-W, Andrews, Z-B, Shanabrough, M, Borok, E, Elsworth, JD, Roth, RH, Sleeman, MW, Picciotto, MR, Tschöp, MH, Gao, X-B, Horvath, TL (2006) Ghrelin modulates the activity and synaptic input organization of midbrain dopamine neurons while promoting appetite. *J Clin Investigation* 116:3229–3239.

50. Damian, M, Marie, J, Leyris, JP, Fehrentz, JA, Verdie, P, Martinez, J, Baneres, JL, Mary, S (2012) High constitutive activity is an intrinsic feature of ghrelin receptor protein: A study with a functional monomeric GHS-R1a receptor reconstituted in lipid discs. *J Biol Chem* 287:3630–3641.

51. Holst B, Cygankiewicz A, Jensen TH, Ankersen M, Schwartz TW (2003) High constitutive signaling of the ghrelin receptor—identification of a potent inverse agonist. *Molec Endocrinol* 17:2201–2210.

52. Sivertsen B, Holliday N, Madsen AN, Holst B (2013) Functionally biased signalling properties of 7TM receptors - opportunities for drug development for the ghrelin receptor. *Brit J Pharmacol* 170:1349–1362.

53. Cummings, DE, Purnell, JQ, Frayo, RS, Schmidova, K, Wisse, BE, Weigle, DS (2001) A preprandial rise in plasma ghrelin levels suggests a role in meal initiation in humans. *Diabetes* 50:1714–1719.

54. Tschöp, M, Weyer, C, Tataranni, PA, Devanarayan, V, Ravussin, E, Heiman, ML (2001) Circulating ghrelin levels are decreased in human obesity. *Diabetes* 50:707–709.

55. Carlini VP, Ghersi M, Schiöth HB, de Barioglio SR (2010) Ghrelin and memory: Differential effects on acquisition and retrieval. *Peptides* 31:1190–1193.

56. Diano, S, Farr, SA, Benoit, SC, McNay, EC, da Silva, I, Horvath, B, Gaskin, FS, Nonaka, N, Jaeger, LB, Banks, WA, Morley, JE, Pinto, S, Sherwin, RS, Xu, L,

Yamada, KA, Sleeman, MW, Tschöp, MH, Horvath, TL (2006) Ghrelin controls hippocampal spine synapse density and memory performance. *Nat Neurosci* 9:381–388.

57. Bednarek, MA, Feighner, SD, Pong, S-S, McKee, KK, Hreniuk, DL, Silva, MV, Warren, VA, Howard, AD, Van der Ploeg, LHY, Heck, JV (2000) Structure-function studies on the new growth hormone-releasing peptide, ghrelin: minimal sequence of ghrelin necessary for activation of growth hormone secretagogue receptor 1a. *J Med Chem* 43:4370–4376.

58. Gutierrez, JA, Solenberg, PJ, Perkins, DR, Willency, JA, Knierman, MD, Jin, Z, Witcher, DR, Luo, S, Onyia, JE, Hale, JE (2008) Ghrelin octanoylation mediated by an orphan lipid transferase. *Proceedings of the National Academy of Sciences* 105:6320–6325.

59. Yang J, Brown MS, Liang G, Grishin NV, Goldstein JL (2008) Identification of the acyltransferase that octanoylates ghrelin, an appetite-stimulating peptide hormone. *Cell* 132:387–396.

60. Ohgusu, H, Shirouzu, K, Nakamura, Y, Nakashima, Y, Ida, T, Sato, T, Kojima, M (2009) Ghrelin O-acyltransferase (GOAT) has a preference for n-hexanoyl-CoA over n-octanoyl-CoA as an acyl donor. *Biochem Biophys Res Commun* 386:153–158.

61. Silva Elipe MV, Bednarek MA, Gao Y-D (2001) ^{1}H NMR structural analysis of human ghrelin and its six truncated analogs. *Biopolymers* 59:489–501.

62. Staes, E, Absil, P-A, Lins, L, Brasseur, R, Deleu, M, Lecouturier, N, Fievez, V, des Rieux, A, Mingeot-Leclercq, M-P, Raussens, V, Préat, V (2010) Acylated and unacylated ghrelin binding to membranes and to ghrelin receptor: towards a better understanding of the underlying mechanisms. *Biochim Biophys Acta* 1798:2102–2113.

63. Beevers AJ, Kukol A (2006) Conformational flexibility of the peptide hormone ghrelin in solution and lipid membrane bound: A molecular dynamics study. *J Biomol Struct Dyn* 23:357–363.

64. Martín-Pastor, M, De Capua, A, Alvarez, CJP, Díaz-Hernández, MD, Jiménez-Barbero, J, Casanueva, FF, Pazos, Y (2010) Interaction between ghrelin and the ghrelin receptor (GHS-R1a), a NMR study using living cells. *Bioorganic & Medicinal Chemistry* 18:1583–1590.

65. Li, X, Dang, S, Yan, C, Gong, X, Wang, J, Shi, Y (2012) Structure of a presenilin family intramembrane aspartate protease. *Nature* 493:56–61.

66. Krishnamurthy H, Gouaux E (2012) X-ray structures of LeuT in substrate-free outward-open and apo inward-open states. *Nature* 481:469–474.

67. Cherezov, V, Rosenbaum, DM, Hanson, MA, Rasmussen, SGF, Thian, FS, Kobilka, TS, Choi, HJ, Kuhn, P, Weis, WI, Kobilka, BK, Stevens, RC (2007) High-resolution crystal structure of an engineered human Beta-2-Adrenergic G protein-coupled receptor. *Science* 318:1258–1265.

68. Zou, Y, Weis, WI, Kobilka, BK (2012) N-Terminal T4 lysozyme fusion facilitates crystallization of a G protein coupled receptor. *PLoS ONE* 7:e46039.

69. Caffrey M, Li D, Dukkipati A (2012) Membrane protein structure determination using crystallography and lipidic mesophases: Recent advances and successes. *Biochemistry* 51:6266–6288.

70. Kay LE (2011) Solution NMR spectroscopy of supra-molecular systems, why bother? A methyl-TROSY view. *J Magn Reson* 210:159–170.

71. Barrett, PJ, Chen, J, Cho, M-K, Kim, J-H, Lu, Z, Mathew, S, Peng, D, Song, Y, Van Horn, WD, Zhuang, T, Sönnichsen, FD, Sanders, CR (2013) The quiet renaissance of protein nuclear magnetic resonance. *Biochemistry* 52:1303–1320.

72. Maslennikov I, Choe S (2013) Advances in NMR structures of integral membrane proteins. *Curr Opin Struct Biol* 23:555–562.

73. Sahu, ID, McCarrick, RM, Troxel, KR, Zhang, R, Smith, HJ, Dunagan, MM, Swartz, MS, Rajan, PV, Kroncke, BM, Sanders, CR, Lorigan, GA (2013) DEER EPR measurements for membrane protein structures via bifunctional spin labels and lipodisq nanoparticles. *Biochemistry* 52:6627–6632.

74. Zou P, Mchaourab HS (2009) Alternating access of the putative substrate-binding chamber in the ABC transporter MsbA. *J Mol Biol* 393:574–585.

75. Mchaourab HS, Mishra S, Koteiche HA, Amadi SH (2008) Role of sequence bias in the topology of the multidrug transporter EmrE. *Biochemistry* 47:7980–7982.

76. Bordignon E, Steinhoff H-J (2007) Membrane protein structure and dynamics studied by site-directed spin-labeling ESR *ESR Spect Memb Biophys* 27:129–164.

77. Cordero-Morales, JF, Cuello, LG, Zhao, Y, Jogini, V, Cortes, DM, Roux, B, Perozo, E (2006) Molecular determinants of gating at the potassium-channel selectivity filter. *Nat Struct Mol Biol* 13:311–318.

78. Lee, J, Chen, J, Brooks, CL, Im, W (2008) Application of solid-state NMR restraint potentials in membrane protein modeling. *J Magnet Resn* 193:68–76.

79. Marassi FM, Opella SJ (2000) A solid-state NMR index of helical membrane protein structure and topology. *J Magnet Reson* 144:150–155.

80. Ding, X, Zhao, X, Watts, A (2013) G-protein-coupled receptor structure, ligand binding and activation as studied by solid-state NMR spectroscopy. *Biophys J* 450:443–457.

81. Ding Y, Yao Y, Marassi FM (2013) Membrane protein structure determination in membrana. *Acc Chem Res* 46:2182–2190.

82. Kloppmann E, Punta M, Rost B (2012) Structural genomics plucks high-hanging membrane proteins. *Curr Opin Struct Biol* 22:326–332.

83. Finn, RD, Bateman, A, Clements, J, Coggill, P, Eberhardt, RY, Eddy, SR, Heger, A, Hetherington, K, Holm, L, Mistry, J, Sonnhammer, ELL, Tate, J, Punta, M (2013) Pfam: The protein families database. *Nucl Acids Res* 42:D222–D230.

84. Käll L, Krogh A, Sonnhammer ELL (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* 21 Suppl 1:i251–7.

85. Murzin, AG, Brenner, SE, Hubbard, T (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540.

86. Bakshi K, Liyanage MR, Volkin DB, Middaugh CR (2013) Fourier transform infrared spectroscopy of peptides. *Meth Mol Biol* 1088:255–269.

87. Jas GS, Kuczera K (2004) Equilibrium structure and folding of a helix-forming peptide: Circular dichroism measurements and replica-exchange molecular dynamics simulations. *Biophys J* 87:3786–3798.

88. Williams, S, Causgrove, TP, Gilmanshin, R, Fang, KS, Callender, RH, Woodruff, WH, Dyer, RB (1996) Fast events in protein folding: Helix melting and formation in a small peptide. *Biochemistry* 35:691–697.

89. Matthes D, de Groot BL (2009) Secondary structure propensities in peptide folding simulations: A systematic comparison of molecular mechanics interaction schemes. *Biophys J* 97:599–608.

90. Gnanakaran S, Nymeyer H, Portman J, Sanbonmatsu KY, Garcia AE (2003) Peptide folding simulations. *Curr Opin Struct Biol* 13:168–174.

91. Kaufmann, KW, Dawson, ES, Henry, LK, Field, JR, Blakely, RD, Meiler, J (2009) Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies. *Prot Struct Funct Bioinfo* 74:630–642.

92. Lees-Miller, JP, Subbotina, JO, Guo, J, Yarov-Yarovoy, V, Noskov, SY, Duff, HJ (2009) Interactions of H562 in the S5 helix with T618 and S621 in the pore

helix are important determinants of hERG1 potassium channel structure and function. *Biophys J* 96:3600–3610.

93. Fortenberry, C, Bowman, EA, Proffitt, W, Dorr, B, Combs, S, Harp, J, Mizoue, L, Meiler, J (2011) Exploring symmetry as an avenue to the computational design of large protein domains. *J Am Chem Soc* 133:18026–18029.

94. Keeble, AH, Joachimiak, LA, Maté, MJ, Meenan, N, Kirkpatrick, N, Baker, D, Kleanthous, C (2008) Experimental and computational analyses of the energetic basis for dual recognition of immunity proteins by colicin endonucleases. *J Mol Biol* 379:745–759.

95. Eswar N, Eramian D, Webb B, Shen M-Y, Sali A (2008) Protein structure modeling with Modeller. *Meth Mol Biol* 426: 145–159.

96. Misura KMS, Chivian D, Rohl CA, Kim DE, Baker D (2006) Physically realistic homology models built with Rosetta can be more accurate than their templates. *Proc Natl Acad Sci* 103:5361–5366.

97. Nguyen ED, Norn C, Frimurer TM, Meiler J (2013) Assessment and challenges of ligand docking into comparative models of G-protein coupled receptors. *PLoS ONE* 8:e67302.

98. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols* 4:363–371.

99. Bennett-Lovsey RM, Herbert AD, Sternberg MJE, Kelley LA (2007) Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Prot Struct Funct Bioinfo* 70:611–625.

100. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein Structure Prediction Using Rosetta. *Meth Enzymol* 383:66–93.

101. Kaufmann KW, Lemmon GH, Deluca SL, Sheehan JH, Meiler J (2010) Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry* 49:2987–2998.

102. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268:209–225.

103. Leaver-Fay, A, Tyka, M, Lewis, SM, Lange, OF, Thompson, J, Jacak, R, Kaufman, K, Renfrew, PD, Smith, CA, Sheffler, W, Davis, IW, Cooper, S, Treuille, A, Mandell, DJ, Richter, F, Ban, Y-EA, Fleishman, SJ, Corn, JE, Kim, DE, Lyskov, S, Berrondo, M, Mentzer, S, Popović, Z, Havranek, JJ, Karanicolas, J, Das, R, Meiler, J, Kortemme, T, Gray, JJ, Kuhlman, B, Baker, D, Bradley, P (2011) Rosetta3: An object-oriented software suite for the simulation and design

of macromolecules. *Methods Enzymol* 487:545–574.

104. Meiler J, Baker D (2003) Coupled prediction of protein secondary and tertiary structure. *Proc Natl Acad Sci USA* 100:12105–12110.

105. Yarov-Yarovoy V, Schonbrun J, Baker D (2005) Multipass membrane protein structure prediction using Rosetta. *Prot Struct Funct Bioinfo* 62:1010–1025.

106. Rohl CA (2005) Protein structure estimation from minimal restraints using Rosetta. *Methods Enzymol* 394:244–260.

107. Lange, OF, Rossi, P, Sgourakis, NG , Song, Y, Lee, HW, Aramini, JM, Ertekin, A, Xiao, R, Acton, TB, Montelione, GT, Baker, D (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc Natl Acad Sci*109:10873–10878.

108. Lange OF, Baker D (2011) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Prot Struct Funct Bioinfo* 80:884–895.

109. Bradley, P, Malmström, L, Qian, B, Schonbrun, J, Chivian, D, Kim, DE, Meiler, J, Misura, KMS, Baker, D (2005) Free modeling with Rosetta in CASP6. *Prot Struct Funct Bioinfo* 61:128–134.

110. Bradley P, Misura KMS, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309:1868–1871.

111. Das, R, Qian, B, Raman, S, Vernon, R, Thompson, J, Bradley, P, Khare, S, Tyka, MD, Bhat, D, Chivian, D, Kim, DE, Sheffler, WH, Malmström, L, Wollacott, AM, Wang, C, André, I, Baker, D (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Prot Struct Funct Bioinfo* 69:118–128.

112. Xu D, Zhang Y (2012) *Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Prot Struct Funct Bioinfo* 80:1715-1735.

113. Tai C-H, Bai H, Taylor TJ, Lee B (2013) Assessment of template-free modeling in CASP10 and ROLL. *Prot Struct Funct Bioinfo* 82:57–83.

114. Marks, DS, Colwell, LJ, Sheridan, R, Hopf, TA, Pagnani, A, Zecchina, R, Sander, C (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6:e28766.

115. Karakaş, M, Woetzel, N, Staritzbichler, R, Alexander, N, Weiner, BE, Meiler, J (2012) BCL::Fold--*De novo* prediction of complex and large protein topologies by assembly of secondary structure elements. *PLoS ONE* 7:e49240.

116. Woetzel, N, Karakaş, M, Staritzbichler, R, Müller, R, Weiner, BE, Meiler, J (2012) BCL::Score—Knowledge based energy potentials for ranking protein models represented by idealized secondary structure elements. *PLoS ONE* 7:e49242.

117. Barth P, Wallner B, Baker D (2009) Prediction of membrane protein structures with complex topologies using limited constraints. *Proc Natl Acad Sci* 106:1409–1414.

118. Hopf, TA, Colwell, LJ, Sheridan, R, Rost, B, Sander, C, Marks, DS (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149:1607–1621.

119. Nugent T, Jones DT (2012) Accurate *de novo* structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci* 109:E1540–E1547.

120. Maupetit, J, Derreumaux, P, Tuffery, P (2009) PEP-FOLD: An online resource for *de novo* peptide structure prediction. *Nucl Acids Res* 37:W498–W503.

121. Raveh B, London N, Zimmerman L, Schueler-Furman O (2011) Rosetta FlexPepDock ab-initio: Simultaneous folding, docking and refinement of peptides onto their receptors. *PLoS ONE* 6:e18934.

122. Kaur H, Garg A, Raghava G (2007) PEPstr: A *de novo* method for tertiary structure prediction of small bioactive peptides. *Prot Pept Lett* 14:626–631.

123. Ishikawa K, Yue K, Dill KA (2008) Predicting the structures of 18 peptides using Geocore. *Protein Sci* 8:716–721.

124. Nicosia G, Stracquadanio G (2008) Generalized pattern search algorithm for peptide structure prediction. *Biophys J* 95:4988–4999.

125. Jang, S, Shin, S, Pak, Y (2002) Molecular dynamics study of peptides in implicit water: *Ab initio* folding of beta-hairpin, beta-sheet, and beta-beta alpha-motif. *J Am Chem Soc* 124:4976–4977.

126. Chebaro Y, Dong X, Laghaei R, Derreumaux P, Mousseau N (2009) Replica exchange molecular dynamics simulations of coarse-grained proteins in implicit solvent. *J Phys Chem B* 113:267–274.

127. Dunbrack RL Jr, Cohen FE (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* 6:1661–1681.

128. Rohl CA, Strauss CEM, Chivian D, Baker D (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Prot Struct Funct Bioinfo* 55:656–677.

129. Simons KT1, Bonneau R, Ruczinski I, Baker D. (1999) *Ab initio* protein structure prediction of CASP III targets using Rosetta. *Proteins* Suppl 3:171–176.

130. Barth P, Schonbrun J, Baker D (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc Natl Acad Sci USA* 104:15682–15687.

131. Lazaridis, T, Karplus, M (1999) Effective energy function for proteins in solution. *Proteins* 35:133–152.

132. Kortemme T, Morozov AV, Baker D (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol* 326:1239–1259.

133. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nat Biotechnol* 30:1072–1080.

134. Alexander N, Bortolus M, Al-Mestarihi A, Mchaourab H, Meiler J (2008) *De novo* high-resolution protein structure determination from sparse spin-labeling EPR data. *Structure* 16:181–195.

135. Hirst SJ, Alexander N, Mchaourab HS, Meiler J (2011) RosettaEPR: An integrated tool for protein structure determination from sparse EPR data. *J Struct Biol* 173:506–514.

136. Meiler J, Baker D (2011) Rapid protein fold determination using unassigned NMR data. *Proc Natl Acad Sci* 100:15404–15409.

137. Meiler J, Baker D (2005) The fumarate sensor DcuS: Progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. *J Magnet Resn* 173:310–316.

138. DiMaio F, Leaver-Fay A, Bradley P, Baker D, André I (2011) Modeling symmetric macromolecular structures in Rosetta3. *PLoS ONE* 6:e20450.

139. Shen, Y, Bryan, PN, He, Y, Orban, J, Baker, D, Bax, A (2010) *De novo* structure generation using chemical shifts for proteins with high-sequence identity but different folds. *Protein Sci* 19:349–356.

140. Herzog, F, Kahraman, A, Boehringer, D, Mak, R, Bracher, A, Walzthoeni, T, Leitner, A, Beck, M, Hartl, FU, Ban, N, Malmström, L, Aebersold, R (2012) Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science* 337:1348–1352.

141. Grishaev A, Guo L, Irving T, Bax A (2010) Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit

water modeling. *J Am Chem Soc* 132:15484–15486.

142. Pandit, D, Tuske, SJ, Coales, SJ, Sook, YE, Liu, A, Lee, JE, Morrow, JA, Nemeth, JF, Hamuro, Y (2012) Mapping of discontinuous conformational epitopes by amide hydrogen/deuterium exchange mass spectrometry and computational docking. *J Mol Recognit* 25:114–124.

143. Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol NMR* 44:213–223.

144. Schieborr, S, Sreeramulu, S, Elshorst, B, Maurer, M, Saxena, K, Stehle, T, Kudlinzki, D, Gande, SL, Schwalbe, H (2013) MOTOR: Model assisted software for NMR structure determination. *Prot Struct Funct Bioinfo* 81:2007–2022.

145. Perozo E (1999) Structural rearrangements underlying K^+-channel activation gating. *Science* 285:73–78.

146. Liu, Y-S, Sompornpisut, P, Perozo, E (2001). Structure of the KcsA channel intracellular gate in the open state. *Nat Struct Biol* 8:883–887.

147. Zou P, Bortolus M, Mchaourab HS (2009) Conformational cycle of the ABC transporter MsbA in liposomes: Detailed analysis using double dlectron–electron resonance spectroscopy. *J Mol Biol* 393:586–597.

148. Hubbell WL, Mchaourab HS, Altenbach C, Lietzow MA (1996) Watching proteins move using site-directed spin labeling. *Structure* 4:779–783.

149. Kaplan RS, Mayor JA, Kotaria R, Walters DE, Mchaourab HS (2000) The yeast mitochondrial citrate transport protein: determination of secondary structure and solvent accessibility of transmembrane domain IV using site-directed spin labeling. *Biochemistry* 39:9157–9163.

150. Altenbach C, Oh KJ, Trabanino RJ, Hideg K, Hubbell WL (2001) Estimation of inter-residue distances in spin labeled proteins at physiological temperatures: experimental strategies and practical limitations. *Biochemistry* 40:15471–15482.

151. Perozo E, Cortes DM, Cuello LG (1998) Three-dimensional architecture and gating mechanism of a K^+ channel studied by EPR spectroscopy. *Nat Struct Biol* 6:459–469.

152. Borbat, PP, Mchaourab, HS, Freed, JH (2002) Protein structure determination using long-distance constraints from double-quantum coherence ESR: Study of T4 lysozyme. *J Am Chem Soc* 124:5304–5314.

153. Berliner LJ, Eaton GR, Eaton SS eds. (2002) *Distance Measurements in Biological Systems by EPR* (Springer US, Boston, MA).

154. Fanucci GE, Cafiso DS (2006) Recent advances and applications of site-directed spin labeling. *Curr Opin Struct Biol* 16:644–653.

155. Herzyk P, Hubbard RE (1995) Automated method for modeling seven-helix transmembrane receptors from experimental data. *Biophys J* 69:2419–2442.

156. Jao CC, Hegdea BG, Chenb J, Hawortha IS, Langena R (2004) Structure of membrane-bound α-synuclein studied by site-directed spin labeling. *Proc Natl Acad Sci USA* 101:8331–8336.

157. Vasquez V, Sotomayor M, Cordero-Morales J, Schulten K, Perozo E (2008) A structural mechanism for MscS gating in lipid bilayers. *Science* 321:1210–1214.

158. Sompornpisut P, Liu YS, Perozo E (2001) Calculation of rigid-body conformational changes using restraint-driven Cartesian transformations. *Biophys J* 81:2530–2546.

159. Alexander, NS, Preininger, AM, Kaya, AI, Stein, RA, Hamm, HE, Meiler, J (2013) Energetic analysis of the rhodopsin–G-protein complex links the α5 helix to GDP release. *Nat Struct Mol Biol* 21:56–63.

160. Van Eps, N, Preininger, AM, Alexander, N, Kaya, AI, Meier, S, Meiler, J, Hamm, HE, Hubbell, WL (2011) Interaction of a G protein with an activated receptor opens the interdomain interface in the alpha subunit. *Proc Natl Acad Sci* 108:9420–9424.

161. Langen R, Oh KJ, Cascio D, Hubbell WL (2000) Crystal structures of spin labeled T4 lysozyme mutants: Implications for the interpretation of EPR spectra in terms of structure. *Biochemistry* 39:8396–8405.

162. Sale K, Song L, Liu Y-S, Perozo E, Fajer P (2005) Explicit treatment of spin labels in modeling of distance constraints from dipolar EPR and DEER. *J Am Chem Soc* 127:9334–9335.

163. Fajer, MI, Li, H, Yang, W, Fajer, PG (2007) Mapping electron paramagnetic resonance spin label conformations by the simulated scaling method. *J Am Chem Soc* 129:13840–13846.

164. Polyhach Y, Bordignon E, Jeschke G (2011) Rotamer libraries of spin labelled cysteines for protein studies. *Phys Chem Chem Phys* 13:2356.

165. Ibata, Y, Iijima, N, Kataoka, Y, Kakihara, K, Tanaka, M, Hosoya, M, Hinuma, S (2000) Morphological survey of prolactin-releasing peptide and its receptor with special reference to their functional roles in the brain. *Neurosci Res* 38:223–230.

166. Matsumoto, H, Noguchi, J, Horikoshi, Y, Kawamata, Y, Kitada, C, Hinuma, S, Onda, H, Nishimura, O, Fujino, M (1999) Stimulation of prolactin release by

prolactin-releasing peptide in rats. *Biochem Biophys Res Commun* 259:321–324.

167. Luckman, SM, Lawrence, CB, Celsi, F, Brennand, J (2000) Alternative role for prolactin-releasing peptide in the regulation of food intake. *Nat Neurosci* 3:645–646.

168. Bowers PM, Strauss CEM, Baker D (2000) *De novo* protein structure determination using sparse NMR data. *J Biomol NMR* 18:311–318.

169. Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* 13:289–302.

170. Schueler-Furman O, Wang C, Bradley P, Baker D (2005) Progress in modeling of protein structures and interactions. *Science* 310:638–642.

171. Findeisen M, Rathmann D, Beck-Sickinger AG (2011) Structure-activity studies of RFamide peptides reveal subtype-selective activation of neuropeptide FF1 and FF2 receptors. *Chem Med Chem* 6:1081–1093.

172. Bohme I, Morl K, Bamming D, Meyer C, Beck-Sickinger AG (2007) Tracking of human Y receptors in living cells--A fluorescence approach. *Peptides* 28:226–234.

173. Berridge MJ, Dawson RM, Downes CP, Heslop JP, Irvine RF (1983) Changes in the levels of inositol phosphates after agonist-dependent hydrolysis of membrane phosphoinositides. *Biochem J* 212:473–482.

174. Berridge MJ (1983) Rapid accumulation of inositol trisphosphate reveals that agonists hydrolyse polyphosphoinositides instead of phosphatidylinositol. *Biochem J* 212:849–858.

175. Höfliger MM, Castejon GL, Kiess W, Beck-Sickinger AG (2003) Novel cell line selectively expressing neuropeptide Y-Y2 receptors. *J Recept Signal Transduct* 23:351–360.

176. Koglin N, Lang M, Rennert R, Beck-Sickinger AG (2003) Facile and selective nanoscale labeling of peptides in solution by using photolabile protecting groups. *J Med Chem* 46:4369–4372.

177. Millhauser GL, Stenland CJ, Hanson P, Bolin KA, van de Ven FJ (1997) Estimating the relative populations of 3(10)-helix and α-helix in Ala-rich peptides: A hydrogen exchange and high field NMR study. *J Mol Biol* 267:963–974.

178. Millhauser GL (1995) Views of helical peptides: A proposal for the position of 3(10)-helix along the thermodynamic folding pathway. *Biochemistry* 34:3873–

3877.

179. Schievano E, Pagano K, Mammi S, Peggion E (2005) Conformational studies of Aib-rich peptides containing lactam-bridged side chains: Evidence of 3(10)-helix formation. *Biopolymers* 80:294–302.

180. Castrignanò, T, D'Onorio De Meo, P, Cozzetto, D, Talamo, IG, Tramontano, A (2006) The PMDB Protein Model Database. *Nucl Acids Res* 34:D306–D309.

181. Joosten, RP, te Beek, TAH, Krieger, E, Hekkelman, ML, Hooft, RWW, Schneider, R, Sander, C, Vriend, G (2010) A series of PDB related databases for everyday needs. *Nucl Acids Res* 39:D411–D419.

182. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.

183. Haridas, V. (2009). From peptides to non‐peptide alpha‐helix inducers and mimetics. *Euro J Org Chem* 2009:5112–5128.

184. Rohl CA, Doig AJ (1996) Models for the 3(10)‐helix/coil, π‐helix/coil, and α‐helix/310‐helix/coil transitions in isolated peptides. *Protein Sci* 5:1687–1696.

185. Toniolo C, Crisma M, Formaggio F, Peggion C (2002) Control of peptide conformation by the Thorpe-Ingold effect (C?-tetrasubstitution). *Biopolymers* 60:396–419.

186. Liu, Z, Chen, K, Ng, A, Shi, Z, Woody, RW, Kallenbach, NR (2004) Solvent dependence of PII conformation in model alanine peptides. *J Am Chem Soc* 126:15141–15150.

187. Woody, RW (2009) Circular dichroism spectrum of peptides in the poly(Pro)II conformation. *J Am Chem Soc* 131:8234–8245.

188. Toniolo, C, Polese, A, Formaggio, F, Crisma, M, Kamphuis, J (1996) Circular dichroism spectrum of a peptide 3(10)-helix. *J Am Chem Soc* 118:2744–2745.

189. Toniolo, C, Formaggio, F, Tognon, S, Broxterman, QB, Kaptein, B, Huang, R, Setnicka, V, Keiderling, TA, McColl, IH, Hecht, L, Barron, LD (2004) The complete chirospectroscopic signature of the peptide 310-helix in aqueous solution. *Biopolymers* 75:32–45.

190. Greenfield NJ (1996) Methods to estimate the conformation of proteins and polypeptides from circular dichroism data. *Anal Biochem* 235:1–10.

191. Cheng Y, Prusoff WH (1973) Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an

enzymatic reaction. *Biochem Pharmacol* 22:3099–3108.

192. Peeters, MC, van Westen, GJP, Guo, D, Wisse, LE, Muller, CE, Beukers, MW, Ijzerman, AP(2011) GPCR structure and activation: an essential role for the first extracellular loop in activating the adenosine A2B receptor. *FASEB J* 25:632–643.

193. Peeters MC, van Westen GJP, Li Q, Ijzerman AP (2011) Importance of the extracellular loops in G protein-coupled receptors for ligand recognition and receptor activation. *Trends Pharmacol Sci* 32:35–42.

194. Chorev, M, Gurrath, M, Behar, V, Mammi, S, Tonello, A, Peggion, E (1995) Conformation and interactions of bombolitin I analogues with SDS micelles and phospholipid vesicles: CD, fluorescence, two-dimensional NMR and computer simulations. *Biopolymers* 36:473–484.

195. Padmanabhan S, Jimenez MA, Laurents DV, Rico M (1998) Helix-stabilizing nonpolar interactions between tyrosine and leucine in aqueous and TFE solutions: ^{2}D-^{1}H NMR and CD studies in alanine-lysine peptides. *Biochemistry* 37:17318–17330.

196. Roccatano D, Colombo G, Fioroni M, Mark AE (2002) Mechanism by which 2,2,2-trifluoroethanol/water mixtures stabilize secondary-structure formation in peptides: A molecular dynamics study. *Proc Natl Acad Sci USA* 99:12179–12184.

197. Schwyzer R (1991) Peptide-membrane interactions and a new principle in quantitative structure-activity relationships. *Biopolymers* 31:785–792.

198. Sudha TS, Vijayakumar EK, Balaram P (1983) Circular dichroism studies of helical oligopeptides. Can 3(10) and alpha-helical conformations be chiroptically distinguished? *Int J Pept Prot Res* 22:464–468.

199. Wang, J, McElheny, D, Fu, Y, Li, G, Kim, J, Zhou, Z, Wu, L, Keiderling, TA, Hammer, RP (2009) A 3 10-helical pentapeptide in water: Interplay of α,α-disubstituted amino acids and the central residue on structure formation. *Biopolymers* 92:452–464.

200. Biron, Z, Khare, S, Samson, AO, Hayek, Y, Naider, F, Anglister, J (2002) A monomeric 3(10)-helix is formed in water by a 13-residue peptide representing the neutralizing determinant of HIV-1 on gp41. *Biochemistry* 41:12687–12696.

201. Hammarström LG, Gauthier TJ, Hammer RP, McLaughlin ML (2001) Amphipathic control of the 3(10)-/α-helix equilibrium in synthetic peptides. *J Pept Res* 58:108–116.

202. Chen X, Wang Q, Ni F, Ma J (2010) Structure of the full-length Shaker

potassium channel Kv1.2 by normal-mode-based X-ray crystallographic refinement. *Proc Natl Acad Sci* 107:11352–11357.

203. Ricardo Simão Vieira-Pires and João Henrique Morais-Cabral (2010) 310 helices in channels and other membrane proteins. *J Gen Physiol* 136:585–592.

204. Dastmalchi S, Church WB, Morris MB, Iismaa TP, Mackay JP (2004) Presence of transient helical segments in the galanin-like peptide evident from 1 H NMR, circular dichroism, and prediction studies. *J Struct Biol* 146:261–271.

205. Bouchayer E, Stassinopoulou CI, Tzougraki C, Marion D, Gans P (2001) NMR and CD conformational studies of the C-terminal 16-peptides of Pseudomonas aeruginosa c551 and Hydrogenobacter thermophilus c552 cytochromes. *J Pept Res* 57:39–47.

206. Armen R, Alonso DOV, Daggett V (2003) The role of α-, 3 10-, and π-helix in helix→coil transitions. *Protein Sci* 12:1145–1157.

207. Sheinerman FB, Brooks CL (1995) 310 Helixes in Peptides and Proteins As Studied by Modified Zimm-Bragg Theory. *J Am Chem Soc* 117:10098–10103.

208. Härterich S, Koschatzky S, Einsiedel J, Gmeiner P (2008) Novel insights into GPCR—Peptide interactions: Mutations in extracellular loop 1, ligand backbone methylations and molecular modeling of neurotensin receptor 1. *Bioorg Med Chem* 16:9359–9368.

209. Sautel, M, Rudolf, K, Wittneben, H, Herzog, H, Martinez, R, Munoz, M, Eberlein, W, Engel, W, Walker, P, Beck-Sickinger, AG (1996) Neuropeptide Y and the nonpeptide antagonist BIBP 3226 share an overlapping binding site at the human Y1 receptor. *Molec Pharmacol* 50:285–292.

210. Shi L, Javitch JA (2002) The binding site of aminergic G-protein coupled receptors: The transmembrane segments and second sxtracellular loop. *Ann Rev Pharmacol Toxicol* 42:437–467.

211. Hawtin, SR, Simms, J, Conner, M, Lawson, Z, Parslow, RA, Trim, J, Sheppard, A, Wheatley, M (2006) Charged extracellular residues, conserved throughout a G-protein-coupled receptor family, are required for ligand binding, receptor activation, and cell-surface expression. *J Biol Chem* 281:38478–38488.

212. Klco JM, Nikiforovich GV, Baranski TJ (2006) Genetic analysis of the first and third extracellular loops of the C5a receptor reveals an essential WXFG motif in the first loop. *J Biol Chem* 281:12010–12019.

213. Akman MS, Girard M, O'Brien LF, Ho AK, Chik CL (1993) Mechanisms of action of a second generation growth hormone-releasing peptide (Ala-His-D-beta Nal-Ala-Trp-D-Phe-Lys-NH2) in rat anterior pituitary cells. *Endocrinology*

132:1286–1291.

214. Cheng K, Chan WW, Barreto A, Convey EM, Smith RG (1989) The synergistic effects of His-D-Trp-Ala-Trp-D-Phe-Lys-NH2 on growth hormone (GH)-releasing factor-stimulated GH release and intracellular adenosine 3“,5-”monophosphate accumulation in rat primary pituitary cell culture. *Endocrinology* 124:2791–2798.

215. Wren, AM, Seal, LJ, Cohen, MA, Brynes, AE, Frost, GS, Murphy, KG, Dhillo, WS, Ghatei, MA, Bloom, SR (2001) Ghrelin enhances appetite and increases food intake in humans. *J Clin Endocrinol Metabol* 86:5992–5995.

216. Date, Y, Nakazato, M, Hashiguchi, S, Dezaki, K, Mondal, MS, Hosoda, HS, Kojima, M, Kangawa, K, Arima, T, Matsuo, H, Yada, T, Matsukura, S (2002) Ghrelin is present in pancreatic alpha-cells of humans and rats and stimulates insulin secretion. *Diabetes* 51:124–129.

217. Lin, Y, Matsumura, K, Fukuhara, M, Kagiyama, S, Fujii, K, Iida, M (2004) Ghrelin acts at the nucleus of the solitary tract to decrease arterial pressure in rats. *Hypertension* 43:977–982.

218. Filigheddu, N, Gnocchi, VF, Coscia, M, Cappelli, M, Porporato, PE, Taulli, R, Traini, S, Baldanzi, G, Chianale, F, Cutrupi, S, Arnoletti, E, Ghe, C, Fubini, A, Surico, N, Sinigaglia, F, Ponzetto, C, Muccioli, G, Crepaldi, T, Graziani, A (2007) Ghrelin and des-acyl ghrelin promote differentiation and fusion of C2C12 skeletal muscle cells. *Mol Biol Cell* 18:986–994.

219. le Roux, CW, Patterson, M, Vincent, RP, Hunt, C, Ghatei, MA, Bloom, SR (2005) Postprandial plasma ghrelin is suppressed proportional to meal calorie content in normal-weight but not obese subjects. *J Clin Endocrinol Metab* 90:1068–1071.

220. Shiiya, T, Nakazato, M, Mizuta, M, Date, Y, Mondal, MS, Tanaka, M, Nozoe, S-I, Hosoda, H, Kangawa, K, Matsukura, S (2002) Plasma ghrelin levels in lean and obese humans and the effect of glucose on ghrelin secretion. *J Clin Endocrinol Metab* 87:240–244.

221. English PJ, Ghatei MA, Malik IA, Bloom SR, Wilding JPH (2002) Food fails to suppress ghrelin levels in obese humans. *J Clin Endocrinol Metab* 87:2984–2984.

222. Hansen, TK, Dall, R, Hosoda, H, Kojima, M, Kangawa, K, Christiansen, JS, Jørgensen, JOL (2002) Weight loss increases circulating levels of ghrelin in human obesity. *Clin Exp Pharmacol Physiol* 56:203–206.

223. Poykko, SM, Kellokoski, E, Horkko, S, Kauma, H, Kesaniemi, YA, Ukkola, O (2003) Low plasma ghrelin is associated with insulin resistance, hypertension,

and the prevalence of type 2 diabetes. *Diabetes* 52:2546–2553.

224. Ukkola O (2011) Ghrelin in type 2 diabetes mellitus and metabolic syndrome. *Mol Cell Endocrinol* 340:26–28.

225. Heppner, KM, Chaudhary, N, Müller, TD, Kirchner, H, Habegger, KM, Ottaway, N, Smiley, DL, Dimarchi, R, Hofmann, SM, Woods, SC, Sivertsen, B, Holst, B, Pfluger, PT, Perez-Tilve, D, Tschöp, MH (2012) Acylation type determines ghrelin's effects on energy homeostasis in rodents. *Endocrinology* 153:4687–4695.

226. Bednarek, MA, Feighner, SD, Pong, S-S, McKee, KK, Hreniuk, DL, Silva, MV, Warren, VA, Howard, AD, Van der Ploeg, LHY, Heck, JV (2000) Structure−function studies on the new growth hormone-releasing peptide, ghrelin: Minimal sequence of ghrelin necessary for activation of growth hormone secretagogue receptor 1a. *J Med Chem* 43:4370–4376.

227. Torsello, A, Ghè, C, Bresciani, E, Catapano, F, Ghigo, E, Deghenghi, R, Locatelli, V, Muccioli, G (2002) Short ghrelin peptides neither displace ghrelin binding in vVitroNor stimulate GH release *in vivo*. *Endocrinology* 143:1968–1968.

228. Sargent DF, Schwyzer R (1986) Membrane lipid phase as catalyst for peptide-receptor interactions. *Proc Natl Acad Sci* 83:5774–5778.

229. De Ricco R, Valensin D, Gaggelli E, Valensin G (2013) Conformation propensities of des-acyl-ghrelin as probed by CD and NMR. *Peptides* 43:62–67.

230. Grossauer J, Kosol S, Schrank E, Zangger K (2010) The peptide hormone ghrelin binds to membrane-mimetics via its octanoyl chain and an adjacent phenylalanine. *Bioorg Med Chem*18:5483–5488.

231. Casey PJ (1995) Protein lipidation in cell signaling. *Science* 268:221–225.

232. Brunsveld L, Waldmann H, Huster D (2009) Membrane binding of lipidated Ras peptides and proteins — The structural point of view. *Biochim et Biophys Acta - Biomem* 1788:273–288.

233. Reuther, G, Tan, K-T, Köhler, J, Nowak, C, Pampel, A, Arnold, K, Kuhlmann, J, Waldmann, H, Huster, D (2006) Structural model of the membrane-bound C terminus of lipid-modified human N-Ras protein. *Angew Chem Int Ed* 45:5387–5390.

234. Reuther, G, Tan, K-T, Vogel, A, Nowak, C, Arnold, K, Kuhlmann, J, Waldmann, H, Huster, D (2006) The lipidated membrane anchor of full length N-Ras Protein shows an extensive dynamics as revealed by solid-state NMR spectroscopy. *J Am Chem Soc* 128:13840–13846.

235. Vogel, A, Reuther, G, Weise, K, Triola, G, Nikolaus, J, Tan, K-T, Nowak, C, Herrmann, A, Waldmann, H, Winter, R, Huster, D (2009) The lipid modifications of Ras that sense membrane environments and induce local enrichment. *Angew Chem Int Ed* 48:8784–8787.

236. Sgourakis, NG, Lange, OF, DiMaio, F, André, I, Fitzkee, NC, Rossi, P, Montelione, GT, Bax, A, Baker, D (2011) Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. *J Am Chem Soc* 133:6288–6298.

237. Vernon R, Shen Y, Baker D, Lange OF (2013) Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker. *J Biomol NMR* 57:117–127.

238. Cook GA, Dawson LA, Tian Y, Opella SJ (2013) Three-dimensional structure and interaction studies of hepatitis C virus p7 in 1,2-dihexanoyl-sn-glycero-3-phosphocholine by solution nuclear magnetic resonance. *Biochemistry* 52:5295–5303.

239. Els S, Beck-Sickinger AG, Chollet C (2010) Ghrelin receptor: High constitutive activity and methods for developing inverse agonists. *Methods Enzymol* 485:103–121.

240. Hope MJ, Bally MB, Webb G, Cullis PR (1985) Production of large unilamellar vesicles by a rapid extrusion procedure: Characterization of size distribution, trapped volume and ability to maintain a membrane potential. *Biochim Biophys Acta* 812:55–65.

241. Udenfriend, S, Stein, S, Böhlen, P, Dairman, W, Leimgruber, W, Weigele, M (1972) Fluorescamine: a reagent for assay of amino acids, peptides, proteins, and primary amines in the picomole range. *Science* 178:871–872.

242. Buser, CA, McLaughlin, S (1998) Ultracentrifugation technique for measuring the binding of peptides and proteins to sucrose-loaded phospholipid vesicles. *Methods Mol Biol* 84:267–281.

243. Huster D, Arnold K, Gawrisch K (1998) Influence of docosahexaenoic acid and cholesterol on lateral lipid organization in phospholipid mixtures. *Biochemistry* 37:17299–17308.

244. Lee CW, Griffin RG (1989) Two-dimensional $^{1}H/^{13}C$ heteronuclear chemical shift correlation spectroscopy of lipid bilayers. *Biophys J* 55:355–358.

245. Szeverenyi, NM, Sullivan, MJ, Maciel, GE (1982) Observation of spin exchange by two-dimensional fourier transform ^{13}C cross polarization-magic-angle spinning. *J Magnet Resn* 47:462–475.

246. Munowitz, MG, Griffin, RG, Bodenhausen, G, Huang, TH (1981) Two-dimensional rotational spin-echo nuclear magnetic resonance in solids: Correlation of chemical shift and dipolar interactions. *J Am Chem Soc* 103:2529–2533.

247. Huster D, Xiao L, Hong M (2001) Solid-state NMR investigation of the dynamics of the soluble and membrane-bound colicin Ia channel-forming domain. *Biochemistry* 40:7662–7674.

248. Barré P, Zschörnig O, Arnold K, Huster D (2003) Structural and dynamical changes of the bindin B18 peptide upon binding to lipid membranes. A solid-state NMR study. *Biochemistry* 42:8377–8386.

249. Huster, D, Yao, X, Hong, M (2002) Membrane protein topology probed by ^{1}H spin diffusion from lipids using solid-state NMR spectroscopy. *J Am Chem Soc* 124:874–883.

250. Kumashiro, KK, Schmidt-Rohr, K, Murphy, OJ, Ouellette, KL, Cramer, WA, Thompson, LK (1998) A novel tool for probing membrane protein structure: Solid-state NMR with proton spin diffusion and X-nucleus detection. *J Am Chem Soc* 120:5043–5051.

251. Raman, S, Lange, OF, Rossi, P, Tyka, M, Wang, X, Aramini, J, Liu, G, Ramelot, TA, Eletsky, A, Szyperski, T, Kennedy, MA, Prestegard, J, Montelione, GT, Baker, D (2010) NMR structure determination for larger proteins using backbone-only data. *Science* 327:1014–1018.

252. Combs, SA, Deluca, SL, Deluca, SH, Lemmon, GH, Nannemann, DP, Nguyen, ED, Willis, JR, Sheehan, JH, Meiler, J (2013) Small-molecule ligand docking into comparative models with Rosetta. *Nature Protocols* 8:1277–1298.

253. Meiler J (2003) PROSHIFT: Protein chemical shift prediction using artificial neural networks. *J Biomol NMR* 26:25–37.

254. Gregory, KJ, Nguyen, ED, Reiff, SD, Squire, EF, Stauffer, SR, Lindsley, CW, Meiler, J, Conn, PJ (2013) Probing the metabotropic glutamate receptor 5 (mGlu5) positive allosteric modulator (PAM) binding pocket: Discovery of point mutations that engender a "molecular switch" in PAM pharmacology. *Molec Pharmacol* 83:991–1006.

255. Canutescu A, Shelenkov A, Dunbrack R (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* 12:2001–2014.

256. Neal S (2003) Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. *J Biomol NMR* 26:215–240.

257. Thompson, JD, Higgins, DG, Gibson, TJ (1994) CLUSTAL W: Improving the

sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl Acids Res* 22:4673–4680.

258. Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: Significantly improved protein chemical shift prediction. *J Biomol NMR* 50:43–57.

259. Konagurthu, AS, Whisstock, JC, Stuckey, PJ, Lesk, AM (2006) MUSTANG: A multiple structural alignment algorithm. *Proteins Struct Funct Bioinfo* 64:559–574.

260. Okada, T, Sugihara, M, Bondar, A-N, Elstner, M, Entel, P, Buss, V (2004) The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure. *J Mol Biol* 342:571–583.

261. Shen Y, Bax A (2010) SPARTA+: A modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR* 48:13–22.

262. Bond CS, Schüttelkopf AW (2009) ALINE: a WYSIWYG protein-sequence alignment editor for publication-quality alignments. *Acta Crystallogr D Biol Crystallogr* 65:510–512.

263. Shen, Y, Lange, O, Delaglio, F, Rossi, P, Aramini, JM, Liu, G, Eletsky, A, Wu, Y, Singarapu, KK, Lemak, A, Ignatchenko, A, Arrowsmith, CH, Szyperski,, T, Montelione, GT, Baker, D, Bax, A (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105:4685–4690.

264. Shen Y, Vernon R, Baker D, Bax A (2008) *De novo* protein structure generation from incomplete chemical shift assignments. *J Biomol NMR* 43:63–78.

265. Gront D, Kulp DW, Vernon RM, Strauss CEM, Baker D (2011) Generalized fragment picking in Rosetta: Design, protocols and applications. *PLoS ONE* 6:e23294.

266. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21:1087.

267. Adzhubei AA, Sternberg MJE, Makarov AA (2013) Polyproline-II helix in proteins: structure and function. *J Mol Biol* 425:2100–2132.

268. Vogel, A, Katzka, CP, Waldmann, H, Arnold, K, Brown, MF, Huster, D (2005) Lipid modifications of a Ras peptide exhibit altered packing and mobility versus host membrane as detected by ^{2}H solid-state NMR. *J Am Chem Soc* 127:12263–12272.

269. Huster D (2005) Investigations of the structure and dynamics of membrane-associated peptides by magic angle spinning NMR. *Prog Nucl Magn Reson Spect* 46:79–107.

270. Wishart DS, Watson MS, Boyko RF, Sykes BD (1997) Automated ^{1}H and ^{13}C chemical shift prediction using the BioMagResBank. *J Biomol NMR* 10:329–336.

271. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 292:195–202.

272. Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14:846–856.

273. Huster, D, Kuhn, K, Kadereit, D, Waldmann, H, Arnold, K (2001) ^{1}H high-resolution magic angle spinning NMR spectroscopy for the investigation of a Ras lipopeptide in a lipid membrane. *Ang Chem Int Ed* 40:1056–1058.

274. Huster, D, Vogel, A, Katzka, C, Scheidt, HA, Binder, H, Dante, S, Gutberlet, T, Zschörnig, O, Waldmann, H, Arnold, K (2003) Membrane insertion of a lipidated Ras peptide studied by FTIR, solid-state NMR, and neutron diffraction spectroscopy. *J Am Chem Soc* 125:4070–4079.

275. Wimley WC, White SH (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat Struct Mol Biol* 3:842–848.

276. Krimm S (1980) The hydrophobic effect: Formation of micelles and biological membranes, *J Polym Sci B Polym Lett Ed* 18:687–687.

277. Murray, D, Ben-Tal, N, Honig, B, McLaughlin, S (1997) Electrostatic interaction of myristoylated proteins with membranes: Simple physics, complicated biology. 5:985–989.

278. Ben-Tal N, Honig B, Peitzsch RM, Denisov G, McLaughlin S (1996) Binding of small basic peptides to membranes containing acidic lipids: Theoretical models and experimental results. *Biophys J* 71:561–575.

279. Floquet, N, M'Kadmi, C, Perahia, D, Gagne, D, Bergé, G, Marie, J, Banères, J-L, Galleyrand, J-C, Fehrentz, J-A, Martinez, J (2010) Activation of the ghrelin receptor is described by a privileged collective motion: A model for constitutive and agonist-induced activation of a sub-class A G-protein coupled receptor (GPCR). *J Mol Biol* 395:769–784.

280. Kukol A (2007) The structure of ghrelin. *Vitamins and Hormones* 77:1–12.

281. Dehlin, E, Liu, J, Yun, SH, Fox, E, Snyder, S, Gineste, C, Willingham, L, Geysen, M, Gaylinn, BD, Sando, JJ (2008) Regulation of ghrelin structure and

membrane binding by phosphorylation. *Peptides* 29:904–911.

282. Foroutan A, Lazarova T, Padrós E (2011) Study of membrane-induced conformations of Substance P: Detection of extended polyproline II helix conformation. *J Phys Chem B* 115:3622–3631.

283. Chellgren BW, Creamer TP (2004) Short sequences of non-proline residues can adopt the polyproline II helical conformation. *Biochemistry* 43:5864–5869.

284. Stapley BJ, Creamer TP (2008) A survey of left-handed polyproline II helices. *Protein Sci* 8:587–595.

285. Cubellis MV, Caillez F, Blundell TL, Lovell SC (2005) Properties of polyproline II, a secondary structure element implicated in protein-protein interactions. *Prot Struct Funct Bioinfo* 58:880–892.

286. Blanch, EW, Morozova-Roche, LA, Cochran, DA, Doig, AJ, Hecht, L, Barron, LD (2000) Is polyproline II helix the killer conformation? A Raman optical activity study of the amyloidogenic prefibrillar intermediate of human lysozyme. *J Mol Biol* 301:553–563.

287. Eker F, Griebenow K, Schweitzer-Stenner R (2004) Aβ 1-28 fragment of the amyloid peptide predominantly adopts a polyproline II conformation in an acidic solution. *Biochemistry* 43:6893–6898.

288. Hicks, JM, Hsu, VL (2004) The extended left-handed helix: A simple nucleic acid-binding motif. *Prot Struct Funct Bioinfo* 55:330–338.

289. Ma K, Kan L-S, Wang K (2001) Polyproline II helix is a key structural motif of the elastic PEVK segment of titin. *Biochemistry* 40:3427–3438.

290. Rath A, Davidson AR, Deber CM (2005) The structure of "unstructured" regions in peptides and proteins: Role of the polyproline II helix in protein folding and recognition. *Biopolymers* 80:179–185.

291. Lam SL, Hsu VL (2003) NMR identification of left-handed polyproline type II helices. *Biopolymers* 69:270–281.

292. Camilloni C, De Simone A, Vranken WF, Vendruscolo M (2012) Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* 51:2224–2231.

293. Jensen, MR, Salmon, L, Nodet, G, Blackledge, M (2010). Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J Am Chem Soc 132*(4), 1270–1272.

294. Skora L, Zweckstetter M (2012) Determination of amyloid core structure using

chemical shifts. *Protein Sci* 21:1948–1953.

295. Tusnady G, Dosztanyi Z, Simon I (2004) Transmembrane proteins in the Protein Data Bank: Identification and classification. *Bioinformatics* 20:2964–2972.

296. Haley DA, Bova MP, Huang QL, Mchaourab HS, Stewart PL (2000) Small heat-shock protein structures reveal a continuum from symmetric to variable assemblies. *J Mol Biol* 298:261–272.

297. Harrison SC (2004) Whither structural biology? *Nat Struct Mol Biol* 11:12–15.

298. Bonneau R, Ruczinski I, Tsai J, Baker D (2009) Contact order and ab initio protein structure prediction. *Protein Sci* 11:1937–1944.

299. Bonneau R, Baker D (2001) Ab initio protein structure prediction: Progress and Prospects. *Annu Rev Biophys Biomol Struct* 30:173–189.

300. Bradley, P, Chivian, D, Meiler, J, Misura, KMS, Rohl, CA, Schief, WR, Wedemeyer, WJ, Schueler-Furman, O, Murphy, P, Schonbrun, J, Strauss, CEM, Baker, D (2003) Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. *Prot Struct Funct Genetics* 53:457–468.

301. Raman, S, Vernon, R, Thompson, J, Tyka, M, Sadreyev, R, Pei, J, Kim, D, Kellogg, E, DiMaio, F, Lange, O, Kinch, L, Sheffler, W, Kim, B-H, Das, R, Grishin, NV, Baker, D (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta. *Prot Struct Funct Bioinfo* 77:89–99.

302. Das, R, Andre, I, Shen, Y, Wu, Y, Lemak, A, Bansal, S, Arrowsmith, CH, Szyperski, T, Baker, D (2009) Simultaneous prediction of protein folding and docking at high resolution. *Proc Natl Acad Sci* 106:18978–18983.

303. Rohl, CA, Baker, D (2002) *De novo* determination of protein backbone structure from residual dipolar couplings using Rosetta. *J Am Chem Soc* 124:2723–2729.

304. Cartes DM, Cuello LG, Perozo E (2001) Molecular architecture of full-length KcsA role of cytoplasmic domains in ion permeation and activation gating. *J Gen Physiol* 117:165–180.

305. Hubbell WL, Altenbach C (1994) Investigation of structure and dynamics in membrane proteins using site-directed spin labeling. *Curr Opin Struct Biol* 4:566–573.

306. Koteiche HA, Mchaourab HS (2002) The determinants of the oligomeric structure in Hsp16.5 are encoded in the alpha-crystallin domain. *FEBS Lett* 519:16–22.

307. Altenbach C, Greenhalgh DA, Khorana HG, Hubbell WL (1994) A collision

gradient method to determine the immersion depth of nitroxides in lipid bilayers: application to spin-labeled mutants of bacteriorhodopsin. *Proc Natl Acad Sci* 91:1667–1671.

308. Altenbach, C, Yang, K, Farrens, DL, Farahbakhsh, ZT, Khorana, HG, Hubbell, WL (1996) Structural features and light-dependent changes in the cytoplasmic interhelical E−F loop region of rhodopsin: A site-directed spin-labeling study. *Biochemistry* 35:12470–12478.

309. Altenbach C, Klein-Seetharaman J, Hwa J, Khorana HG, Hubbell WL (1999) Structural features and light-dependent changes in the sequence 59-75 connecting helices I and II in rhodopsin: A site-directed spin-labeling study. *Biochemistry* 38:7945–7949.

310. Gross A, Columbus L, Hideg K, Altenbach C, Hubbell WL (1999) Structure of the KcsA potassium channel from *Streptomyces lividans*: A site-directed spin labeling study of the second transmembrane segment. *Biochemistry* 38:10324–10335.

311. Mchaourab HS, Lietzow MA, Hideg K, Hubbell WL (1996) Motion of spin-labeled side chains in T4-lysozyme. Correlation with protein structure and dynamics. *Biochemistry* 35:7692–7704.

312. Brown, LJ, Sale, KL, Hills, R, Rouviere, C, Song, L, Zhang, X, Fajer, PG (2002) Structure of the inhibitory region of troponin by site directed spin labeling electron paramagnetic resonance. *Proc Natl Acad Sci* 99:12765–12770.

313. Fleissner MR, Cascio D, Hubbell WL (2009) Structural origin of weakly ordered nitroxide motion in spin-labeled proteins. *Protein Sci* 18:893–908.

314. Kazmier K, Alexander NS, Meiler J, Mchaourab HS (2011) Algorithm for selection of optimized EPR distance restraints for *de novo* protein structure determination. *J Struct Biol* 173:549–557.

315. Wang, G, Dunbrack, RL (2003). PISCES: A protein sequence culling server. *Bioinformatics*, 19:1589–1591.

316. Durham E, Dorr B, Woetzel N, Staritzbichler R, Meiler J (2009) Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *J Mol Model* 15:1093–1108.

317. Weaver LH, Matthews BW (1987) Structure of bacteriophage T4-lysozyme refined at 1.7Å resolution. *J Mol Biol* 193:189–199.

318. Karplus, K, Sjölander, K, Barrett, C, Cline, M, Haussler, D, Hughey R, Holm, L, Sander, C (1997) Predicting protein structure using hidden Markov models. *Proteins* Suppl 1:134–139.

319. Dunbrack RL Jr, Karplus M (1993) Backbone-dependent rotamer library for proteins application to side-chain prediction. *J Mol Biol* 230:543–574.

320. Misura KMS, Baker D (2005) Progress and challenges in high-resolution refinement of protein structure models. *Prot Struct Funct Bioinfo* 59:15–29.

321. Dunbrack RL (2002) Rotamer libraries in the 21st century. *Curr Opin Struct Biol* 12:431–440.

322. Raman, S, Huang, YJ, Mao, B, Rossi, P, Aramini, JM, Liu, G, Montelione, GT, Baker, D (2010) Accurate automated protein NMR structure determination Using unassigned NOESY data. *J Am Chem Soc* 132:202–207.

323. O'Leary, JM, Hamilton, JM, Deane, CM, Valeyev, NV, Sandell, LJ, Downing, AK (2004) Solution structure and dynamics of a prototypical chordin-like cysteine-rich repeat (von Willebrand factor type C module) from collagen IIA. *J Biol Chem* 279:53857–53866.

324. Sanders CR, Sönnichsen F (2006) Solution NMR of membrane proteins: practice and challenges. *Magn Reson Chem* 44:S24–S40.

325. Bakheet TM, Doig AJ (2009) Properties and identification of human protein drug targets. *Bioinformatics* 25:451–457.

326. Weierstall, U, James, D, Wang, C, White, TA, Wang, D, Liu, W, Spence, JCH, Doak, B, Nelson, G, Fromme, P, Fromme, R, Grotjohann, I, Kupitz, C, Zatsepin, NA, Liu, H, Basu, S, Wacker, D, Han, GW, Katritch, V, Boutet, S, Messerschmidt, M, Williams, GJ, Koglin, JE, Seibert, MM, Klinker, M, Gati, C, Shoeman, RL, Barty, A, Chapman, HN, Kirian, RA, Beyerlein, KR, Stevens, RC, Li, D, Shah, STA, Howe, N, Caffrey, M, Cherezov, V (2014) Lipidic cubic phase injector facilitates membrane protein serial femtosecond crystallography. *Nat Commun* 5:3309-3314.

327. Liu, W, Wacker, D, Gati, C, Han, GW, James, D, Wang, D, Nelson, G, Weierstall, U, Katritch, V, Barty, A, Zatsepin, NA, Li, D, Messerschmidt, M, Boutet, S, Williams, GJ, Koglin, JE, Seibert, MM, Wang, C, Shah, STA, Basu, S, Fromme, R, Kupitz, C, Rendek, KN, Grotjohann, I, Fromme, P, Kirian, RA, Beyerlein, KR, White, TA, Chapman, HN, Caffrey, M, Spence, JCH, Stevens, RC, Cherezov, V (2013) Serial femtosecond crystallography of G protein-coupled receptors. *Science* 342:1521–1524.

328. Li D, Boland C, Walsh K, Caffrey M (2012) Use of a robot for high-throughput crystallization of membrane proteins in lipidic mesophases. *J Vis Exp*:e4000.

329. Horst, R, Stanczak, P, Stevens, RC, Wüthrich, K (2013). β2-Adrenergic receptor solutions for structural biology analyzed with microscale NMR diffusion measurements. *Angew Chem Int Ed* 52:331–335.

330. Chun, E, Thompson, AA, Liu, W, Roth, CB, Griffith, MT, Katritch, V, Kunken, J, Xu, F, Cherezov, V, Hanson, MA, Stevens, RC (2012) Fusion partner toolchest for the stabilization and crystallization of G protein-coupled receptors. *Structure* 20:967–976.

331. Baker LA, Baldus M (2014) Characterization of membrane protein function by solid-state NMR spectroscopy. *Curr Opin Struct Biol* 27C:48–55.

332. Klammt, C, Maslennikov, I, Bayrhuber, M, Eichmann, C, Vajpai, N, Chiu, EJC, Blain, KY, Esquivies, L, Kwon, JHJ, Balana, B, Pieper, U, Sali, A, Slesinger, PA, Kwiatkowski, W, Riek, R, Choe, S (2012) Facile backbone structure determination of human membrane proteins by NMR spectroscopy. *Nat Meth*. 9:834-839.

333. Tang M, Comellas G, Rienstra CM (2013) Advanced solid-state NMR approaches for structure determination of membrane proteins and amyloid fibrils. *Acc Chem Res* 46:2080–2088.

334. Ni. QZ, Daviso, E, Can, TV, Markhasin, E, Jawla, SK, Swager, TM, Temkin, RJ, Herzfeld, J, Griffin, RG (2013) High frequency dynamic nuclear polarization. *Acc Chem Res* 46:1933–1941.

335. Zou P, Mchaourab HS (2010) Increased sensitivity and extended range of distance measurements in spin-labeled membrane proteins: Q-band double electron-electron resonance and nanoscale bilayers. *Biophys J* 98:L18–L20.

336. Mchaourab HS, Steed PR, Kazmier K (2011) Toward the fourth dimension of membrane protein structure: Insight into dynamics from spin-labeling EPR spectroscopy. *Structure* 19:1549–1561.

337. Fiser A, Sali A (2003) Modeller: Generation and refinement of homology-based protein structure models. *Methods Enzymol* 374:461–491.

338. Schwede, T, Kopp, J, Guex, N, Peitsch, MC (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucl Acids Res* 31:3381–3385.

339. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols* 5:725–738.

340. Yamashita A, Singh SK, Kawate T, Jin Y, Gouaux E (2005) Crystal structure of a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporters. *Nature* 437:215–223.

341. Faham, S, Watanabe, A, Besserer, GM, Cascio, D, Specht, A, Hirayama, BA, Wright, EM, Abramson, J (2008) The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na^+/sugar symport. *Science*

321:810–814.

342. Perez C, Koshy C, Yildiz O, Ziegler C (2012) Alternating-access mechanism in conformationally asymmetric trimers of the betaine transporter BetP. *Nature*. 490:126-130.

343. Ma, D, Lu, P, Yan, C, Fan, C, Yin, P, Wang, J, Shi, Y (2012). Structure and mechanism of a glutamate-GABA antiporter. *Nature*, 483:632–636.

344. Kazmier, K, Sharma, S, Quick, M, Islam, SM, Roux, B, Weinstein, H, Javitch, JA, Mchaourab, HS (2014). Conformational dynamics of ligand-dependent alternating access in LeuT. *Nat Struct Mol Biol*, 2:472–479.

345. Weiner BE, Woetzel N, Karakaş M, Alexander N, Meiler J (2013) BCL::MP-fold: Folding membrane proteins through assembly of transmembrane helices. *Structure* 21:1107–1117.

346. Altenbach C, Klein-Seetharaman J, Cai K, Khorana H, Hubbell W (2001) Structure and function in rhodopsin: mapping light-dependent changes in distance between residue 316 in helix 8 and residues in the sequence 60-75, covering the cytoplasmic end of helices TM1 and TM2 and their connection loop CL1. *Biochemistry* 40:15493–15500.

347. Carugo O, Pongor S (2001) A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci* 10:1470–1473.

348. Feigin LA, Svergun DI (1987) *Structure Analysis by Small-Angle X-Ray and Neutron Scattering* ed Taylor GW (Springer US, Boston, MA).

349. Viklund H, Elofsson A (2008) OCTOPUS: Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 24:1662–1668.

350. Adamian L, Liang J (2006) Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Struct Biol* 6:13-29.

351. Tyka MD, Jung K, Baker D (2012) Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers. *J Comput Chem* 33:2483–2491.

352. Karakaş, M, Woetzel, N, Staritzbichler, R, Alexander, N, Weiner, BE, Meiler, J (2012) BCL::Fold - *De Novo* prediction of complex and large protein topologies by assembly of secondary structure elements. *PLoS ONE* 7:e49240.

353. Woetzel, N., Karakaş, M., Staritzbichler, R., Müller, R., Weiner, B. E., & Meiler, J. (2012) BCL::Score—Knowledge Based Energy Potentials for Ranking Protein Models Represented by Idealized Secondary Structure Elements. *PLoS ONE*

7:e49242.

354. Raman P, Cherezov V, Caffrey M (2005) The Membrane Protein Data Bank. *Cell Mol Life Sci* 63:36–51.

355. White SH (2004) The progress of membrane protein structure determination. *Protein Sci* 13:1948–1949.

356. Faham, S, Yang, D, Bare, E, Yohannan, S, Whitelegge, JP, Bowie, JU (2004) Side-chain contributions to membrane protein structure and stability. *J Mol Biol* 335:297–305.

357. Gu W, Geddes BJ, Zhang C, Foley KP, Stricker-Krongrad A (2004) The Prolactin-releasing peptide receptor (GPR10) regulates body weight homeostasis in mice. *JMN* 22:93–104.

358. Stricker-Kongra, A, Gu, W (2001) GPR10 as a target for identifying weight modulating compounds. United States Patent 6,537,765 B2.

359. Jandacek RJ, Woods SC (2004) Pharmaceutical approaches to the treatment of obesity. *Drug Discov Today* 9:874–880.

360. Zhao, H, Xin, Z, Patel, JR, Nelson, LTJ, Liu, B, Szczepankiewicz, BG, Schaefer, VG, Falls, HD, Kaszubska, W, Collins, CA, Sham, HL, Liu, G (2005) Structure–activity relationship studies on tetralin carboxamide growth hormone secretagogue receptor antagonists. *Bioorg Med Chem Lett*, 15:1825–1828.

361. Zhao, H, Xin, Z, Liu, G, Schaefer, VG, Falls, HD, Kaszubska, W, Collins, CA, Sham, HL (2004) Discovery of tetralin carboxamide growth hormone secretagogue receptor antagonists via scaffold manipulation. *J Med Chem* 47:6655–6657.

362. Xin, Z, Serby, MD, Zhao, H, Kosogof, C, Szczepankiewicz, BG, Liu, M, Hutchins, CW, Sarris, KA, Hoff, ED, Falls, HD, Lin, CW, Ogiela, CA, Collins, CA, Brune, ME, Bush, EN, Droz, BA, Fey, TA, Knourek-Segel, VE, Shapiro, R, Jacobson, PB, Beno, DWA, Turner, TM, Sham, HL, Liu G (2006) Discovery and pharmacological evaluation of growth hormone secretagogue receptor antagonists. *J Med Chem* 49:4459–4469.

363. Demange, L, Boeglin, D, Moulin, A, Mousseaux, D, Ryan, J, Bergé, G, Gagne, D, Heitz, A, Perrissoud, D, Locatelli, V, Torsello, A, Galleyrand, J-C, Fehrentz,, J-A, Martinez, J (2007) Synthesis and pharmacological *in vitro* and *in vivo* evaluations of novel triazole derivatives as ligands of the ghrelin receptor 1. *J Med Chem* 50:1939–1957.

364. Moulin, A, Demange, L, Bergé, G, Gagne, D, Ryan, J, Mousseaux, D, Heitz, A, Perrissoud, D, Locatelli, V, Torsello, A, Galleyrand, J-C, Fehrentz, J-A,

Martinez, J (2007) Toward potent ghrelin receptor ligands based on trisubstituted 1,2,4-triazole structure. 2. Synthesis and pharmacological *in vitro* and *in vivo* Evaluations. *J Med Chem* 50:5790–5806.

365. Chollet C, Meyer K, Beck-Sickinger AG (2009) Ghrelin-A novel generation of anti-obesity drug: design, pharmacomodulation and biological activity of ghrelin analogues. *J Pept Sci* 15:711–730.

366. Chollet C, Bergmann R, Pietzsch J, Beck-Sickinger AG (2012) Design, evaluation, and comparison of ghrelin receptor agonists and inverse agonists as suitable radiotracers for PET imaging. *Bioconjugate Chem* 23:771–784.

367. Hruby VJ (2002) Designing peptide receptor agonists and antagonists. *Nat Rev Drug Discov* 1:847–858.

368. Hoyer D, Bartfai T (2012) Neuropeptides and neuropeptide receptors: Drug targets, and peptide and non-peptide ligands: A tribute to Prof. Dieter Seebach. *Chem Biodiv* 9:2367–2387.

369. Bellmann-Sickert K, Beck-Sickinger AG (2010) Peptide drugs to target G protein-coupled receptors. *Trends Pharmacol Sci* 31:434–441.

370. Hruby VJ, Cai M (2013) Design of peptide and peptidomimetic ligands with novel pharmacological activity profiles. *Ann Rev Pharmacol Toxicol* 53:557–580.

371. Thomas, A, Deshayes, S, Decaffmeyer, M, Van Eyck, MH, Charloteaux, B, Brasseur, R (2006) Prediction of peptide structure: How far are we? *Prot Struct Funct Bioinfo* 65:889–897.

372. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J Mol Biol* 293:321–331.

373. Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 323:573–584.

374. Iakoucheva LM (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucl Acids Res* 32:1037–1049.

375. Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem Sci 27*(10), 527–533.

376. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6:197–208.

377. Schwartz TW, Frimurer TM, Holst B, Rosenkilde MM, Elling CE (2006)

Molecular mechanism of 7TM receptor activation--A global toggle switch model. *Ann Rev Pharmacol Toxicol* 46:481–519.

378. Sinz A (2003) Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *J Mass Spectrom* 38:1225–1237.

379. Alexander, NS, Stein, RA, Koteiche, HA, Kaufmann, KW, Mchaourab, HS, Meiler, J (2013) RosettaEPR: Rotamer library for spin label structure and dynamics. *PLoS ONE* 8:e72851.

380. Kroncke BM, Horanyi PS, Columbus L (2010) Structural origins of nitroxide side chain dynamics on membrane protein α-helical sites. *Biochemistry* 49:10045–10060.

381. Altenbach C, Froncisz W, Hemker R, Mchaourab H, Hubbell WL (2005) Accessibility of nitroxide side chains: Absolute Heisenberg exchange rates from power saturation EPR. *Biophys J* 89:2103–2112.

382. Nielsen RD, Che K, Gelb MH, Robinson BH (2005) A ruler for determining the position of proteins in membranes. *J Am Chem Soc* 127:6430–6442.

383. Claxton, DP, Quick, M, Shi, L, de Carvalho, FD, Weinstein, H, Javitch, JA, Mchaourab, HS (2010) Ion/substrate-dependent conformational dynamics of a bacterial homolog of neurotransmitter:sodium symporters. *Nat Struct Mol Biol* 17:822–829.

384. Altenbach C, Kusnetzow AK, Ernst OP, Hofmann KP, Hubbell WL (2008) High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. *Proc Natl Acad Sci* 105:7439–7444.

385. Chang G (2001) Structure of MsbA from *E. coli*: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science* 293:1793–1800.

386. Doyle, DA, Cabral, JM, Pfuetzner, RA, Kuo, A, Gulbis, JM, Cohen, SL, Chait, BT, Mackinnon, R (1998) The structure of the potassium channel: Molecular basis of K^+ conduction and selectivity. *Science* 280:69–77.

387. Webb, B, Lasker, K, Velázquez-Muriel, J, Schneidman-Duhovny, D, Pellarin, R, Bonomi, M, Greenberg, C, Raveh, B, Tjioe, E, Russel, D, Sali, A (2013) Modeling of proteins and their assemblies with the integrative modeling platform. *Methods Mol Biol* 1071:277–295.

388. Bond RA, Ijzerman AP (2006) Recent developments in constitutive receptor activity and inverse agonism, and their potential for GPCR drug discovery. *Trends Pharmacol Sci* 27:92–96.

389. Welch SK, O'Hara BF, Kilduff TS, Heller HC (1995) Sequence and tissue distribution of a candidate G-coupled receptor cloned from rat hypothalamus. *Biochem Biophys Res Commun* 209:606–613.

390. Langmead, CJ, Szekeres, PG, Chambers, JK, Ratcliffe, SJ, Jones, DNC, Hirst, WD, Price, GW, Herdon, HJ (2000) Characterization of the binding of [125I]-human prolactin releasing peptide (PrRP) to GPR10, a novel G protein coupled receptor. *Brit J Pharmacol* 131:683–688.

391. Fukusumi S, Fujii R, Hinuma S (2006) Recent advances in mammalian RFamide peptides: The discovery and functional analyses of PrRP, RFRPs and QRFP. *Peptides* 27:1073–1086.

392. Lin SH (2008) Prolactin-releasing peptide. *Results Probl Cell Different* 46:57–88.

393. Conner AC, Barwell J, Poyner DR, Wheatley M (2011) The use of site-directed mutagenesis to study GPCRs. *Methods Mol Biol* 746:85–98.

394. Lindner, D, van Dieck, J, Merten, N, Mörl, K, Günther, R, Hofmann, H-J, Beck-Sickinger, AG (2008) GPC receptors and not ligands decide the binding mode in neuropeptide Y multireceptor/multiligand system. *Biochemistry* 47:5905–5914.

395. Merten, N, Lindner, D, Rabe, N, Rompler, H, Morl, K, Schoneberg, T, Beck-Sickinger, AG (2006) Receptor subtype-specific docking of Asp6.59 with C-terminal arginine residues in Y receptor ligands. *J Biol Chem* 282:7543–7551.

396. Lagerström, MC, Fredriksson, R, Bjarnadóttir, TK, Fridmanis, D, Holmquist, T, Andersson, J, Yan, Y-L, Raudsepp, T, Zoorob, R, Kukkonen, JP, Lundin, LG, Klovins, J, Chowdhary, BP, Postlethwait, JH, Schiöth, HB (2005) Origin of the prolactin-releasing hormone (PRLH) receptors: Evidence of coevolution between PRLH and a redundant neuropeptide Y receptor during vertebrate evolution. *Genomics* 85:688–703.

397. Carter PJ, Winter G, Wilkinson AJ, Fersht AR (1984) The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase *Bacillus stearothermophilus*. *Cell* 38:835–840.

398. Krylov D, Mikhailenko I, Vinson C (1994) A thermodynamic scale for leucine zipper stability and dimerization specificity: E and G interhelical interactions. *EMBO J* 13:2849–2861.

399. Lang, M, Bufe, B, De Pol, S, Reiser, O, Meyerhof, W, Beck-Sickinger, AG (2006) Structural properties of orexins for activation of their receptors. *J Pept Sci* 12:258–266.

400. Böhme I, Stichel J, Walther C, Mörl K, Beck-Sickinger AG (2008) Agonist

induced receptor internalization of neuropeptide Y receptor subtypes depends on third intracellular loop and C-terminus. *Cellular Signalling* 20:1740–1749.

401. Reynolds, CP, Biedler, JL, Spengler, BA, Reynolds, DA, Ross, RA, Frenkel, EP, Smith, RG (1986) Characterization of human neuroblastoma cell lines established before and after therapy. *J Natl Cancer Inst* 76:375–387.

402. Park JH, Scheerer P, Hofmann KP, Choe H-W, Ernst OP (2008) Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature* 454:183–187.

403. Scheerer, P, Park, JH, Hildebrand, PW, Kim, YJ, Krauss, N, Choe, H-W, Hofmann, KP, Ernst, OP (2008) Crystal structure of opsin in its G-protein-interacting conformation. *Nature* 455:497–502.

404. Warne, T, Serrano-Vega, MJ, Baker, JG, Moukhametzianov, R, Edwards, PC, Henderson, R, Leslie, AGW, Tate, CG, Schertler, GFX (2008) Structure of a β1-adrenergic G-protein-coupled receptor. *Nature* 454:486–491.

405. Jaakola, VP, Griffith, MT, Hanson, MA, Cherezov, V, Chien, EYT, Lane, JR, Ijzerman, AP, Stevens, RC (2008) The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* 322:1211–1217.

406. Mandell DJ, Coutsias EA, Kortemme T (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat Meth* 6:551–552.

407. Gray, JJ, Moughon, S, Wang, C, Schueler-Furman, O, Kuhlman, B, Rohl, CA, & Baker, D (2003) Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* 331:281–299.

408. Madsen KL, Thorsen TS, Rahbek-Clemmensen T, Eriksen J, Gether U (2012) Protein interacting with C kinase 1 (PICK1) reduces reinsertion rates of interaction partners sorted to Rab11-dependent slow recycling pathway. *J Biol Chem* 287:12293–12308.

409. Nygaard R, Frimurer TM, Holst B, Rosenkilde MM, Schwartz TW (2009) Ligand binding and micro-switches in 7TM receptor structures. *Trends Pharmacol Sci* 30:249–259.

410. Donohue, PJ, Sainz, E, Akeson, M, Kroog, GS, Mantey, SA, Battey, JF, Jensen, RT, Northup, JK (1999) An aspartate residue at the extracellular boundary of TMII and an arginine residue in TMVII of the gastrin-releasing peptide receptor interact to facilitate heterotrimeric G protein coupling. *Biochemistry* 38:9366–9372.

411. Govaerts, C, Lefort, A, Costagliola, S, Wodak, SJ, Ballesteros, JA, Van Sande, J, Pardo, L, Vassart, G (2001) A conserved Asn in transmembrane helix 7 is an

on/off switch in the activation of the thyrotropin receptor. *J Biol Chem* 276:22991–22999.

412. Arsenault, J, Cabana, J, Fillion, D, Leduc, R, Guillemette, G, Lavigne, P, Escher, E (2010) Temperature dependent photolabeling of the human angiotensin II type 1 receptor reveals insights into its conformational landscape and its activation mechanism. *Biochem Pharmacol* 80:990–999.

413. Scheer A, Fanelli F, Costa T, De Benedetti PG, Cotecchia S (1996) Constitutively active mutants of the alpha 1B-adrenergic receptor: Role of highly conserved polar amino acids in receptor activation. *EMBO J* 15:3566–3578.

414. Lefkowitz RJ, Cotecchia S, Samama P, Costa T (1993) Constitutive activity of receptors coupled to guanine nucleotide regulatory proteins. *Trends Pharmacol Sci* 14:303–307.

415. Okazaki, R, Chikatsu, N, Nakatsu, M, Takeuchi, Y, Ajima, M, Miki, J, Fujita, T, Arai, M, Totsuka, Y, Tanaka, K, Fukumoto, S (1999) A novel activating mutation in calcium-sensing receptor gene associated with a family of autosomal dominant hypocalcemia 1. *J Clin Endocrinol Metab* 84:363–366.

416. Samama P, Cotecchia S, Costa T, Lefkowitz RJ (1993) A mutation-induced activated state of the beta 2-adrenergic receptor. Extending the ternary complex model. *J Biol Chem* 268:4625–4636.

417. D'Antona, AM, Ahn, KH, Wang, L, Mierke, DF, Lucas-Lenard, J, Kendall, DA (2006) A cannabinoid receptor 1 mutation proximal to the DRY motif results in constitutive activity and reveals intramolecular interactions involved in receptor activation. *Brain Res* 1108:1–11.

418. Högger P, Shockley MS, Lameh J, Sadée W (1995) Activating and inactivating mutations in N- and C-terminal I3 loop junctions of muscarinic acetylcholine Hm1 receptors. *J Biol Chem* 270:7405–7410.

419. Spalding TA, Burstein ES, Brauner-Osborne H, Hill-Eubanks D, Brann MR (1995) Pharmacology of a constitutively active muscarinic receptor generated by random mutagenesis. *J Pharmacol Exp Therapeutics* 275:1274–1279.

420. Seifert R, Wenzel-Seifert K (2002) Constitutive activity of G-protein-coupled receptors: cause of disease and common property of wild-type receptors. *Naunyn-Schmiedeberg's Arch Pharmacol* 366:381–416.

421. Spiegel AM (1996) Defects in G protein-coupled signal transduction in human disease. *Annu Rev Physiol* 58:143–170.

422. Gether, U, Ballesteros, JA, Seifert, R, Sanders-Bush, E, Weinstein, H, Kobilka, BK (1997) Structural instability of a constitutively active G protein-coupled

receptor. Agonist-independent activation due to conformational flexibility. *J Biol Chem* 272:2587–2590.

423. Ford DJ, Essex A, Spalding TA, Burstein ES, Ellis J (2002) Homologous mutations near the junction of the sixth transmembrane domain and the third extracellular loop lead to constitutive activity and enhanced agonist affinity at all muscarinic receptor subtypes. *J Pharmacol Exp Therapeutics* 300:810–817.

424. Gromoll, J, Simoni, M, Nordhoff, V, Behre, HM, De Geyter, C, Nieschlag, E (1996) Functional and clinical consequences of mutations in the FSH receptor. *Mol Cell Endocrinol* 125:177–182.

425. Elling, CE, Frimurer, T.M, Gerlach, LO, Jorgensen, R, Holst, B, Schwartz, TW (2006) Metal ion site engineering indicates a global toggle switch model for seven-transmembrane receptor activation. *J Biol Chem* 281:17337–17346.

426. Kleinau, G, Jaeschke, H, Mueller, S, Worth, CL, Paschke, R, Krause, G (2008) Molecular and structural effects of inverse agonistic mutations on signaling of the thyrotropin receptor--A basally active GPCR. *Cell Mol Life Sci* 65:3664–3676.

427. Grüters, A, Schöneberg, T, Biebermann, H, Krude, H, Krohn, HP, Dralle, H, Gudermann, T (1998) Severe congenital hyperthyroidism caused by a germ-line neomutation in the extracellular portion of the thyrotropin receptor 1. *J Clin Endocrinol Metab* 83:1431–1436.

428. Cotecchia S, Fanelli F, Costa T (2003) Constitutively active G protein-coupled receptor mutants: Implications on receptor function and drug action. *Assay Drug Devel Tech* 1:311–316.

429. Domazet, I, Holleran, BJ, Martin, SS, Lavigne, P, Leduc, R, Escher, E, Guillemette, G (2009) The second transmembrane domain of the human type 1 angiotensin II receptor participates in the formation of the ligand binding pocket and undergoes integral pivoting movement during the process of receptor activation. *J Biol Chem* 284:11922–11929.

430. Flanagan, CA, Rodic, V, Konvicka, K, Yuen, T, Chi, L, Rivier, JE, Millar, RP, Weinstein, H, Sealfon, SC (2000) Multiple interactions of the Asp(2.61(98)) side chain of the gonadotropin-releasing hormone receptor contribute differentially to ligand interaction. *Biochemistry* 39:8133–8141.

431. McAllister, SD, Tao, Q, Barnett-Norris, J, Buehner, K, Hurst, DP, Guarnieri, F, Reggio, PH, Harmon, KWN, Cabral, GA, Abood, ME (2002) A critical role for a tyrosine residue in the cannabinoid receptors for ligand recognition. *Biochem Pharmacol* 63:2121–2136.

432. Ukena K, Vaudry H, Leprince J, Tsutsui K (2011) Molecular evolution and

functional characterization of the orexigenic peptide 26RFa and its receptor in vertebrates. *Cell Tissue Res* 343:475–481.

433. Osugi T, Ukena K, Sower SA, Kawauchi H, Tsutsui K (2006) Evolutionary origin and divergence of PQRFamide peptides and LPXRFamide peptides in the RFamide peptide family. Insights from novel lamprey RFamide peptides. *FEBS J* 273:1731–1743.

434. Findeisen M, Rathmann D, Beck-Sickinger AG (2011) RFamide peptides: structure, function, mechanisms and pharmaceutical potential. *Pharmaceuticals* 4:1248–1280.

435. Allen LF, Lefkowitz RJ, Caron MG, Cotecchia S (1991) G-protein-coupled receptor genes as protooncogenes: Constitutively activating mutation of the alpha 1B-adrenergic receptor enhances mitogenesis and tumorigenicity. *Proc Natl Acad Sci* 88:11354–11358.

436. Alexander N, Woetzel N, Meiler J (2011) bcl:: Cluster: A method for clustering biological molecules coupled with visualization in the Pymol molecular graphics system. *IEEE Comput Adv Bio Med Sci*:13–18.

437. Augustine, N, Lane, N, Andrews, NC, *et al*. (2014) *Restoring the foundation: The vital role of research in preserving the American dream*. (American Academy of Arts & Sciences, Cambridge, MA).

438. Stokes, DE (1997) *Pasteur's quadrant: Basic science and technological innovation*. (Brookings Institution Press, Washington, D.C.).