1.4 Computational methods to study membrane protein structure

In conjunction with experimental methods to study membrane protein folding, computational methods have been invented by evaluating previously solved MP structures. These methods look to further understand MP folding by establishing energetic terms that estimate the thermodynamics of MP folding. In this section, I review computational methods used to predict MP structures, highlighting unique features of each tool.

1.4.1 Rosetta

Rosetta houses a variety of energy functions and prediction tools for soluble environments, including the ability to dock or design proteins *de* novo (Chaudhury et al., 2011; Koehler Leman et al., 2017; Leman et al., 2020; Weitzner et al., 2017; Yarov-Yarovoy, Schonbrun, et al., 2006). These energetics includes but are not limited to: terms parameterized by CHARMM (van der Waals and electrostatics), a hydrogen bond and disulfide function curated from polar contacts found in ~8000 high-resolution crystal structures, and a side chain conformation energy based on the probability of occurrence from the Dunbrack rotamer database (Alford et al., 2017; RICHARDSON et al.). These Rosetta energy functions have been adapted to accommodate predicting helical TMs within the membrane environment. The updated functions include an energy term that models the membrane into layers classified defining atoms as water-exposed, polar, interface, and hydrophobic (White & Wimley, 1999; Yarov-Yarovoy, Schonbrun, et al., 2006). Successful predictions helped to discover structural details in MPs associated with voltage sensing and gating mechanisms (Vargas et al., 2012; Yarov-Yarovoy, Baker, et al., 2006).

Recently, RosettaMP was devised to enhance the functionality of MP prediction which includes modeling TM helices *de novo* from sequence, a representation of the membrane bilayer, and the ability to transform a protein into the membrane (Koehler Leman et al., 2017). These tools increase the accessibility to MP structure, improving the ability to visualize and predict structures of MPs that have not yet been solved, as well as for membrane protein design (Duran & Meiler, 2018). Simultaneously, energetic predictions enhance the knowledge in the field, permitting researchers to analyze structural mutations *in silico* prior to testing with *in vitro* or *in vivo* experiments (Thieker et al., 2022).

1.4.2 Molecular Software Library

Another computational tool for modeling protein structures is the Molecular Software Library, or MSL (Kulp et al., 2012). Similar to Rosetta, MSL contains a variety of tools to perform MP structure prediction. These tools include the ability to transform proteins in space, mutate specific residues, extract geometric information from known structures, and predict the structure and energetics of an input sequence. Additionally, it was used to develop the CATM algorithm. Briefly, the CATM algorithm predicts structures of known and unknown TM dimers that associate by the GASright motif, and experimental studies have shown that it accurately estimates the energetics of association (Anderson et al., 2017; Díaz Vázquez et al., 2023; Mueller et al., 2014). It uses the Energy-Based conformer library applied at the 95% level for side chain mobility (Subramaniam & Senes, 2012), and energetics of predicted proteins are determined using the CHARMM 22 van der Waals function (MacKerell et al., 1998), the IMM1 membrane implicit solvation model (Lazaridis, 2003), and the hydrogen bonding function of SCWRL4 (Krivov et al., 2009). Each of these energy terms is applied to optimize the dimer geometry by Monte Carlo (MC) backbone perturbation cycles where all parameters (xShift, zShift, axialRotation, and crossingAngle) are locally varied. The association energy is calculated as the energy of the dimer minus the energy of two monomers. My research adapts the CATM algorithm to design structures with strong van der Waals packing in the absence of hydrogen bonding, allowing me to assess the extent at which packing can drive MP association. I further detail how I used MSL to design TM homodimers in chapter 3.

1.4.3 Topology Prediction and Docking Algorithms

Other methods used to determine interactions between MPs focus on predicting the topology or docking of individual TM helices. OCTOPUS predicts TM topology using a combination of Markov models and neural networks (Viklund & Elofsson, 2008). HADDOCK can apply experimental knowledge of the interface region between proteins to refine docking (de Vries et al., 2010; Dominguez et al., 2003). PREDDIMER utilizes a novel surface-based modeling approach to predict and screen TM dimers for conformation heterogeneity (Polyansky et al., 2012). EVFold employs evolutionary based structural restraints to refine their docked structures (Braun T et al., 2015). TMDOCK applies an all-atom model for helices, inserting them in the membrane and outputting a structure alongside a predicted ΔG of insertion and ΔG of association (Lomize & Pogozheva, 2017). Each of these methods was adapted into a webserver for online usage, where users can input the sequence and additional information to guide the process.

1.4.4 AlphaFold and RoseTTaFold

Most notably, at the 2020 Critical Assessment of Structure Prediction (CASP) conference, Google’s Deepmind introduced the machine learning model AlphaFold. Unlike previously mentioned prediction algorithms, AlphaFold predicts structures without energetics. AlphaFold utilizes a combination of neural networks and training on multiple sequence alignments (MSAs) and solved protein structures to predict unknown structures to near atomic precision with a 95% confidence interval (Jumper et al., 2021). Shortly afterward, David Baker’s group introduced RoseTTAFold, improving on the Rosetta prediction by incorporating a similar architecture to AlphaFold, with the inclusion of a third track network that connects sequence, residue-residue distances, and atomic coordinates (Baek et al., 2021). Each of these methods drastically improved the ability to predict unknown protein structures using information from previously studied and solved proteins. With increasing interest in using these technologies, multiple free webservers have been established to enhance access to these advanced protein prediction algorithms (Mirdita et al., 2022; Roberts et al., 2024). However, these machine learning algorithms are limited by the amount of information available. AlphaFold struggles to predict proteins with <30 homologs in their MSAs, and accuracy decreases for protein structures dependent on other protein-protein interactions, while RoseTTAFold has difficulty predicting higher-order oligomers (Agard et al., 2022). These limitations are amplified in MPs due to the lack of MP structures, making small or complex TM proteins difficult to predict. To better understand the dynamics of association and folding in MPs, it is necessary to advance our knowledge of the forces involved in folding.