hydrophobic side-chains pointing towards the lipid tails (“dyad-repeat” motif) (Wimley, 2002).

Owing to the helix-breaking property of proline, one might expect to rarely find this amino acid in TMHs. However, early studies suggested that proline is in fact more common in transmembrane \(\alpha\)-helices of IMPs than it is in the helices of soluble proteins (von Heijne, 1991). More recently it was shown that the kinks introduced to helices tend to have larger angles in soluble proteins (Senes et al., 2004), that proline mutations are more tolerable at the interface regions near the lipid headgroups (Yohannan et al., 2004a) and that many proteins seem to have adapted their global structure to the introduction of a kinked helix to a degree where the proline is not required anymore for the helix to remain kinked (Yohannan et al., 2004b).

The previously mentioned hydrophobicity motifs are strong enough to facilitate fairly simple predictions of the TMS of \(\alpha\)-helical and to a lesser degree also \(\beta\)-barrel IMPs (Von Heijne, 1992; Schulz, 2000; Wimley, 2002). While prediction methods for \(\beta\)-barrel IMPs are arguably troubled most by extremely small training sets (Hayat and Elofsson, 2012), prediction of TMHs has developed towards additional challenges over the last two decades. After the initial simple approaches, based solely on experimentally derived hydrophobicity scales, more sophisticated machine learning classifiers arose. PHDhtm was the earliest method to use artificial neural networks, and at the same time included homology information into the prediction (Rost et al., 1995, 1996). Another method still popular today is TMHMM which uses a Hidden Markov model (HMM) and a model architecture that intrinsically also predicts the topology of the respective membrane protein (Krogh et al., 2001). Phobius took a similar approach but introduced a significant improvement by combining the prediction of TMHs and signal peptides in one model (Käll et al., 2004). Previously, the occurrence of signal peptides often lead to false positives due to their hydrophobic regions that likens them to TMHs. The method was subsequently improved by the inclusion of homology information, leading to PolyPhobius which is still among the best performing methods today (Käll et al., 2005; Reeb, 2011). Since then, even more sophisticated methods such as MEMSAT-SVM arose, which predicts not only TMHs, topology and signal peptides but also reentrant loops (Nugent and Jones, 2009).

1.1.3 Transmembrane protein structure determination

There is generally a large divide between the number of known protein sequences and resolved structures, the latter being around four orders of magnitude smaller (Kanehisa Group, 2014; Cochrane et al., 2011). However, three-dimensional data for transmembrane proteins is particularly underrepresented at less than 2% of the structures in the Protein Data Bank (PDB) (Kloppmann et al., 2012; Kozma et al., 2013; Berman et al., 2000). Although these proteins are of especially high interest, e.g. for drug development, as previously mentioned, their nature makes them hard to handle.

For crystallization experiments the proteins have to be extracted from the membrane. However, placing them in a water environment often has a severe impact on structural stability since the residues found at the outer surface are typically highly hydrophobic and could easily aggregate (Kessel and Ben-Tal, 2011). This is usually accounted for
by the use of detergents, which ideally keep the protein in its native structure while at
the same time allowing the formation of crystal contacts. However, finding the exact
experimental parameters and correct handling is far from trivial (Kessel and Ben-Tal,
2011; Mancia and Love, 2011; Eshaghi, 2009). As an alternative, there were also studies
where the deliberate addition of lipids proved to support crystallization (Hunte and
Richers, 2008).

Solid-state NMR allows the study of membrane proteins in their native environment
without the need to form crystals. It has however severe limits on the size of the studied
protein with the maximum feasible length being around 100 residues (McDermott, 2009).
Furthermore, sufficient expression quantities for purification and structure determination
far exceed normal in vivo levels and poses a challenge to obtain in the majority of
cases (Mancia and Love, 2010). One reasons for this is, that the overexpression of, for
example a transporter, can easily prove toxic to the host cell because of the change in
concentration. However, a recent study found that toxicity and overexpression success of
IMPs in E. coli is mainly a function of the protein's biochemical properties and seldom
affected by overexpression as such (Gubellini et al., 2011). Given these challenges, it
does not come as a surprise that the determination of a novel transmembrane protein’s
structure is estimated to have a cost one magnitude higher than that of a soluble protein
(Stevens, 2003).

1.2 Structural genomics and the New York Consortium on Membrane
Protein Structure

With next-generation sequencing now being a reality and the sequences databases rapidly
growing (Martinez and Nelson, 2010; Cochrane et al., 2011), structural genomics (SG)
seeks to decrease the gap between sequences and available structures, often by not
solving a specific, but any structure possible (Rost, 1998). These will then, by the use of
homology modeling as well as easier follow-up experiments of closely related proteins,
increase structural coverage of the current sequence space. In this process, a structure is
in many ways a proxy for protein function. Although knowledge of a protein’s structure
does not equate understanding of its function, it is an important step towards it (Rost,
1998). Even if the study of a target does not result in a structure, all previous steps
that succeeded and the specific protocols are valuable information - also for a group
performing traditional structural biology.

SG efforts began around the turn of the millennium with initial programs focused on
soluble proteins. In Japan, the “RIKEN Structural Genomics/Proteomics Initiative” (RSGI)
started in 2000 with a five year goal of solving 3000 protein structures (Sussman and
Silman, 2008). Today, the RSGI remains the single largest structural genomics center,
that contributed most with 2739 structures deposited in the PDB (PDB Stats for SG, 2014).
The largest project in Europe, “Structural Proteomics IN Europe” (SPINE), deposited 119
structures in the PDB between 2002 and 2006 (Stuart et al., 2006; PDB Stats for SG,
2014). Since 2006, the follow-up project SPINE2 focuses on structures of complete pro-
tein complexes, a very useful but also significantly more challenging endeavor (Sussman
and Silman, 2008). In the USA, several large initiatives are funded by the National
Institutes of General Medical Sciences through the “Protein Structure Initiative” (PSI;
The first phase, PSI-1, funded nine structural genomics centers from 2000 to 2005 and resulted in over 1100 structures deposited in the PDB (Sussman and Silman, 2008). After pipelines and protocols had been developed and tested in the first phase, two more phases, PSI-2 and PSI:Biology followed, partly targeting more specialized challenges. These include the application of structural genomics on IMPs, which has been investigated by the New York Consortium on Membrane Protein Structure (NYCOMPS) and several smaller membrane-specialized centers, since 2005. With the third and last phase of the PSI about to end in 2015, the initiative is the largest contributor to the PDB with 6507 structures deposited so far (Structural Biology Knowledgebase, 2014).

1.2.1 Impact

Structural genomics has a significant impact on structural coverage of SwissProt and PFAM families (Grabowski et al., 2007; Kloppmann et al., 2012). Around 12% of the structures currently deposited in the PDB originate from SG initiatives (PDB Stats for SG, 2014). Additionally, half of the SG domains added to the SCOP database represent a new family. These are often followed by the release of a related structure determined by traditional structural biology. This reinforces the previous point of SG also helping this community with their work (Andreeva et al., 2008). Levitt (2007) suggests that the steady growth of the PDB could only be maintained because of the SG efforts which were started around a time where growth rates began to decline.

As part of the PSI, NYCOMPS specializes on structures of IMPs. Next to eight other centers working on membrane proteins, NYCOMPS deposited most (34) structures in the PDB (Kloppmann et al., 2012; Berman et al., 2000; Structural Biology Knowledgebase, 2014). These include TehA, an anion channel homolog resolved at a very high resolution of 1.2 Å, that gives new insight into the molecular mechanisms of stomata opening and closing in plants (cf. Figure 1.1(b), Chen et al., 2010). Furthermore, an urea transporter structure that reinforced the proposed channel-like mechanism of these proteins (Levin et al., 2012), a formate channel (Waight et al., 2010), and a leucine transporter that gives implications for antidepressant drug action in humans (Zhou et al., 2007), have been solved and deposited in the PDB. The dataset accumulated by NYCOMPS since 2005 is unique in both content and size with roughly 14,000 analyzed targets, all of which are IMPs (cf. Figure 1.2).

1.2.2 NYCOMPS protocols

The heart of NYCOMPS forms a medium-throughput structural genomics pipeline, and at its beginning stands the target selection (Love et al., 2010). Irrespective of changes to this constantly evolving protocol, the basic selection process is as follows (Punta et al., 2009): NYCOMPS uses a fixed set of currently 145 prokaryotic “reagent genomes”, i.e. fully sequence genomes from which DNA is cloned. From these genomes, valid targets are extracted by choosing all proteins predicted to be transmembrane proteins with at least two TMHs. This is to account for prediction errors that will invariably happen and ensure targets are actually IMPs. The prediction was initially performed with TMHMM (Krogh et al., 2001) and has now been transitioned to PolyPhobius (Käll et al., 2005). A redundancy reduction is then performed on these sequences, using CD-HIT with a cutoff
of 98% sequence identity (Li and Godzik, 2006). Next, sequences were excluded when there were more than 15 consecutive residues predicted to be disordered (using IUPRED, Dosztányi et al., 2005).

Next, the so called “seeds” are chosen - IMP sequences of high interest, novelty and diversity (Punta et al., 2009). These can be selected from either prior analysis suggesting them as feasible targets (“central seed selection”), or handpicked e.g. by affiliated laboratories (“nominated seed selection”). These seeds do not have to be in the set of valid targets, as long as their expansion gives at least one hit in this set. This seed expansion is performed using PSI-BLAST to query a database of all valid targets. For target selection an E-value of at least $10^{-3}$ is required, and additionally that a minimum of 50% of the residues predicted to be in a TMH in both seed and target overlap (Altschul et al., 1997). While the nominated seeds do not underlie additional limitations, expanded seeds from the central selection were ensured to have no more than 25% overlap in the predicted TMHs to structures already in the PDB, again defined using PSI-BLAST. Notably, this approach retains the initial seed as the point of reference, instead of a protein family, and can easily be re-iterated e.g. after the addition of additional reagent genomes.

Targets are cloned into expression vectors using the very high-throughput amenable ligation-independent cloning technique (Mancia and Love, 2010). The expression vectors used, are two standard vectors, one N-, one C-terminal, developed by NYCOMPS as well as two vectors from the Midwest Center for Structural Genomics which is also taking part in the PSI. *E. coli* strain DH10B cells are used to amplify the resulting plasmids, which is then purified and sequenced to validate successful cloning. Next, small-scale expression (0.6 ml cultures) test are performed in cells of *E. coli* strain BL21. If they are positive, the respective plasmid is transitioned to mid-scale expression (0.5 l cultures) in the same strain. Incubation temperature for both experiments are typically 37°C and positive expression is verified using SDS-PAGE gels. Recently, the temperature has been changed to 22°C. Finally, proteins are subjected to a detergent assay to determine their stability in a moderately high temperature, high detergent concentration environment. Proteins that are amenable to all or at least some of the screened detergents are passed onto adjoint laboratories for structural investigation.

### 1.2.3 Predicting success of structural investigation in SG pipelines

The structural study of proteins, and transmembrane proteins in particular, is a highly challenging and costly task, not just in terms of crystallization but also all steps leading up to this as shown in Figure 1.2. This is true for all membrane protein centers in the PSI, including NYCOMPS which is the largest with more than 10,000 targets analyzed so far. It is therefore of high interest to develop methods that predict the success of such experiments, ideally during the step of target selection. The first advances on soluble proteins, using data from the first iteration of the PSI, go back to 2004 and are based on Bayesian inference. However Rupp and Wang (2004) point out themselves that data is sparse, results until this point mainly reinforce known principles and prediction can only increase probability of success.

In the following years many more methods were published using exceedingly sophisticated machine learning models such as kernel-based classifiers (Overton et al., 2008;
Figure 1.2. Attrition rate in membrane-specialized structural genomics pipelines. Shown is the number of targets remaining after each major stage in the structural genomics pipelines of all nine centers in the PSI that specialize in membrane proteins, as well as NYCOMPS in particular. Data for all centers was extracted from TargetTrack, while NYCOMPS data was retrieved directly from the LIMS (see Section 2.1.2).

Kurgan et al., 2009; Babnigg and Joachimiak, 2010; Mizianty and Kurgan, 2011), random forests (Jahandideh and Mahdavi, 2012), artificial neural networks (Overton et al., 2011), meta-predictors (Mizianty and Kurgan, 2009), and combinations thereof as well as others (Overton and Barton, 2006; Smialowski et al., 2006; Slabinski et al., 2007; Charoenkwan et al., 2013). There also is a related class of methods that predict the solubility of proteins upon heterologous overexpression (Smialowski et al., 2007; Magnan et al., 2009; Smialowski et al., 2012). However, none of the previously mentioned predictors are designed for application on transmembrane proteins, in fact, several of them are specifically not suggested for the use on transmembrane proteins (Smialowski et al., 2006, 2007, 2012).

These methods all report increasing accuracies and prediction success for soluble proteins, however in 2009, given the lack of data for membrane proteins, Punta et al. conclude that “predictions as to which membrane proteins constitute structure-prone targets can at the moment rather be based on a misunderstanding of statistics than on sustainable science”.

As an additional complication, training sets are not available for all of these methods, making it hard to compare against them. Furthermore, in all but two cases there is only a webserver available for use, which is not a viable option for an analysis such as the one presented here, working in the range of 10000 sequences.
REFERENCES


cytoplasm. *Protein expression and purification*, 27(1), 134–42. 20
Java 7 documentation on Patterns (2014). docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html. Accessed 2014/2/17. 12, 16
membrane proteins. *Protein engineering, design & selection: PEDS*, 24(8), 579–88. 3


REFERENCES


Rost, B., Fariselli, P., and Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. Protein science : a publication of the Protein Society, 5(8), 1704–18. 4
Schrödinger LLC (2011). 3
The Uniprot Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic acids research, 40(Database issue), D71–5. 21