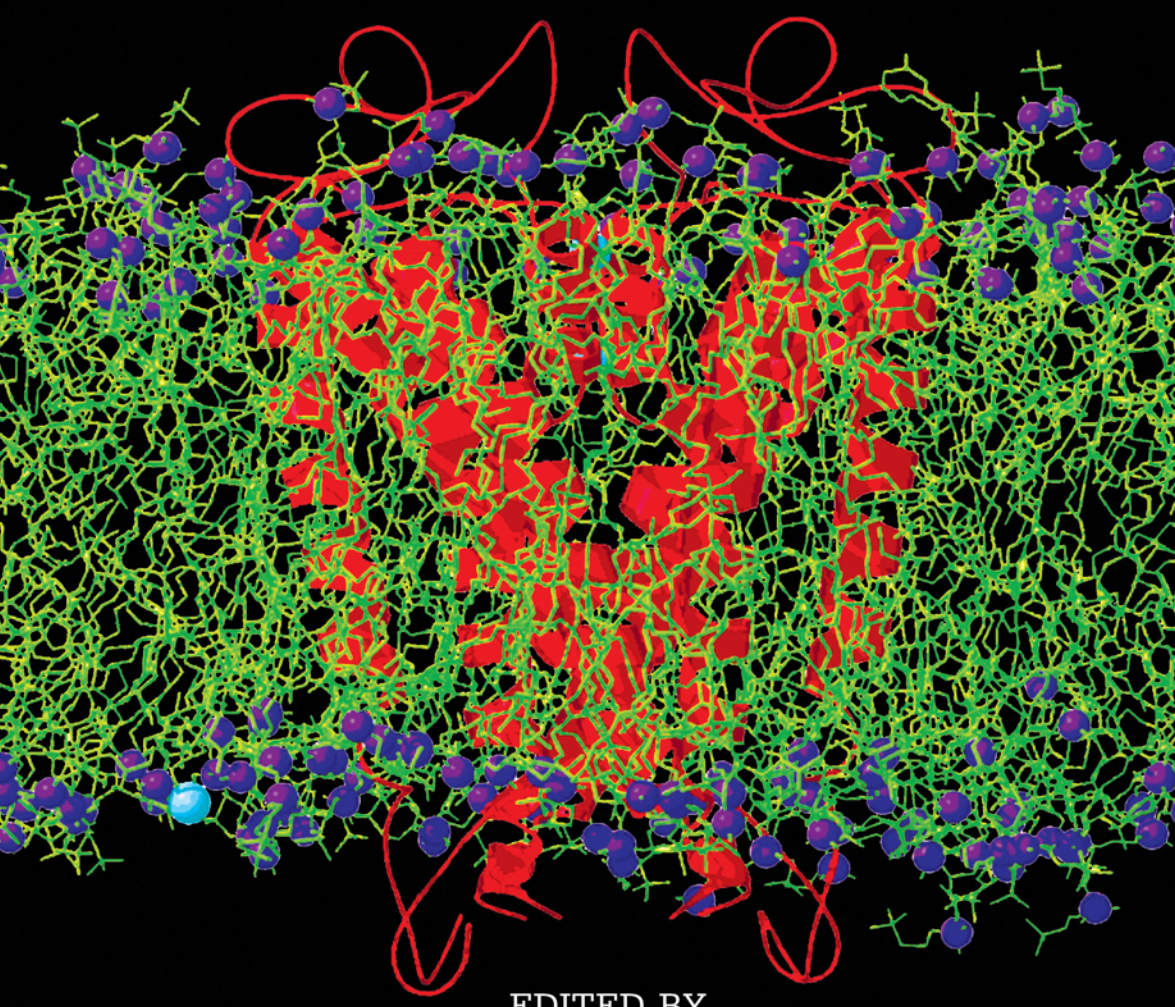
Structural Genomics on Membrane Proteins



EDITED BY

Kenneth H. Lundstrom

CRC Taylor & Francis
Taylor & Francis Group

Structural Genomics on Membrane Proteins

Structural Genomics on Membrane Proteins

EDITED BY

Kenneth H. Lundstrom

Taylor & Francis
Taylor & Francis Group
Boca Raton London New York

A CRC title, part of the Taylor & Francis imprint, a member of the
Taylor & Francis Group, the academic division of T&F Informa plc.

Published in 2006 by
CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-57444-526-X (Hardcover)
International Standard Book Number-13: 978-1-57444-526-8 (Hardcover)
Library of Congress Card Number 2005019887

Library of Congress Cataloging-in-Publication Data

Structural genomics on membrane proteins / [edited by] Kenneth H. Lundstrom.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-1-57444-526-8 (alk. paper)
ISBN-10: 1-57444-526-X (alk. paper)
1. Membrane proteins--Conformation. 2. Genomics. I. Lundstrom, Kenneth H.
[DNLM: 1. Membrane Proteins--genetics. 2. Membrane Proteins--ultrastructure. 3. Genomics.
QU 55 S928 2006]

QP552.M44S87 2006
572'.696--dc22 2005019887

T&F informa
Taylor & Francis Group
is the Academic Division of T&F Informa plc.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Foreword

In the near future, the unprecedented use of novel technologies for membrane proteins will hopefully aid in conquering the last frontier in structural biology. Although there are more than 30,000 protein structures deposited in the protein data bank (PDB), less than 1% of these represent membrane proteins. In light of the fact that membrane proteins constitute >20% of all proteins, this disparity in the number of membrane proteins and available structures is due to the inherent transmembrane nature of these proteins, which makes their expression, purification, stabilization, and crystallization substantially more difficult than for their soluble counterparts. The significance of structural biology has been recently demonstrated in the field of rational drug design. Unfortunately, from the drug discovery point of view, membrane proteins comprise 60 to 70% of current medicinal targets. In order to improve the success in obtaining high resolution structures on membrane proteins, development of appropriate technology is necessary for all areas including expression, purification, and crystallography. *Structural Genomics on Membrane Proteins* provides an excellent overview on novel research in bioinformatics and modeling on membranes as well as the latest technological developments in recombinant protein expression, refolding of membrane proteins from inclusion bodies, solubilization, and purification methods. Moreover, methods for the application of NMR and miniaturization in structural biology as well as electron and atomic force microscopy on membrane proteins are discussed. It is very helpful and a great read!

Krzysztof Palczewski, Ph.D.,
Chair, Department of Pharmacology
Case Western Reserve University
Cleveland, Ohio, U.S.A.

Preface

Genomics, proteomics, and other types of "omics" approaches have proven most fruitful in both current basic and applied research. There are numerous examples of studies on whole genomes of specific organisms, types of genes/proteins and gene/protein families. In genomics and proteomics strategies, the aims are to study the function of gene activities and to characterize proteins. Structural genomics approaches, again, have the goals set on structure determination of target proteins, which can facilitate rational drug design by speeding up the drug development process and improving the selectivity and efficacy of medicines. In this context, membrane proteins are of great significance as more than 60% of current drugs are based on this group of proteins.

As in many other modern approaches, a thorough knowledge in bioinformatics is a necessity for successful applications in structural genomics. Another cornerstone of any structure determination is the supply of proteins for purification and crystallization attempts. Very few proteins, especially those of therapeutic interest, are available in high quantities in native tissue, and even so, ethical issues prevent proceeding with this approach. The scientific community therefore relies, to a large extent, on recombinant expression technologies. As the requirements for membrane protein expression are more demanding, expression has been evaluated in bacterial, yeast, insect, and animal cells, applying a variety of specifically designed vectors containing various expression-enhancing sequences and tags to facilitate purification. The expression mode dictates the downstream processing requests as recombinant proteins in *Escherichia coli* inclusion bodies need to be subjected to refolding, as those expressed in cell membranes require solubilization procedures. Purification technologies are also of great importance as the presence of detergents makes the procedure more difficult and the target proteins less stable. Furthermore, crystallization of membrane proteins has been more difficult than for their soluble counterparts, clearly indicated by the submission of less than 100 structures in public databases compared with more than 30,000 entries for soluble proteins. However, x-ray crystallography is not the only approach to receive structural information on proteins. Alternative methods include NMR (nuclear magnetic resonance) and EM (electron microscopy) technologies.

I would like to acknowledge the authors of the chapters of this book. The enthusiasm encountered was overwhelming and made the project possible. I am also grateful to CRC Press and Jill Jurgensen, Project Coordinator, and Jay Margolis, Project Editor, for the efficient and professional contribution to get the book published. Sadly, during the preparation of the manuscript, Dr. Helmut Reiländer (Aventis Pharmaceuticals, Frankfurt, previously at Max Planck Institute for Biochemistry, Frankfurt) passed away after a long illness. Helmut, as a colleague and a good friend,

made a significant contribution to the field of recombinant membrane proteins, and his support for many structural genomics programs in Europe has been vital. I dedicate this book to his memory.

Kenneth H. Lundstrom, Ph.D.

Editor

Kenneth H. Lundstrom received his Ph.D. in "Overexpression of Viral Membrane Proteins in *Bacillus subtilis*" from the University of Helsinki, Finland. He conducted his postdoctoral research at Cetus Corporation in California on "Antisense Expression and PCR Technologies." Dr. Lundstrom then went back to his native Finland, where he was appointed Senior Scientist at Orion Pharmaceuticals and was involved in cloning, expression, and structural studies on catechol-o-methyltransferases. From 1992 to 1995, he developed Semliki Forest virus vectors for overexpression of receptors (GPCRs) and ion channels at the Glaxo Institute of Molecular Biology, in Geneva, Switzerland. Dr. Lundstrom worked as Principal Biologist at the Glaxo Medicines Research Centre in Stevenage, United Kingdom, from 1995 to 1996. From 1996 to 2001, he was responsible for receptor expression in the CNS Department of F. Hoffmann-La Roche, in Basel, Switzerland. In 2001, Dr. Lundstrom was appointed the Scientific Coordinator of the MePNet program; he became the Chief Scientific Officer of BioXtal in Lausanne, Switzerland, in 2002. He is also part of the Senior Management Team (Vice President, Science & Technology) of Regulon Inc., Mountain View, California, a biotech company involved in cancer therapy. Dr. Lundstrom has published more than 100 scientific papers and reviews in international journals; acts as editor for books in the fields of GPCRs, structural genomics, and gene therapy; and is a frequent speaker at international conferences.

Contributors

Dr. Enrique Abola
The Scripps Research Institute,
La Jolla, California, U.S.A.

Dr. Mark Bacon
University of Leeds
Leeds, United Kingdom

Dr. Monika Bährer
Roche Diagnostics GmbH
Penzberg, Germany

Emma Barksby
University of Leeds
Leeds, United Kingdom

Dr. Sabrina A. Berettta
Abbott Laboratories
Abbott Park, Illinois, U.S.A.

Kim Bettaney
University of Leeds
Leeds, United Kingdom

Dr. Roslyn Bill
Aston University
Birmingham, United Kingdom

Dr. Giel Bosman
Nijmegen Center for Molecular Life Sciences
Nijmegen, The Netherlands

Dr. J. Robert Bostwick
AstraZeneca Pharmaceuticals LP
Wilmington, Delaware, U.S.A.

Dr. Bernadette Byrne
Wolfson Laboratories
Imperial College
London, United Kingdom

Dr. Mark Chiu
Abbott Laboratories
Abbott Park, Illinois, U.S.A.

Dr. Joanne Clough
University of Leeds
Leeds, United Kingdom

Dr. Richard Cogdell
Institute of Biomedical & Life Sciences
University of Glasgow
Glasgow, United Kingdom

Dr. Willem de Grip
Nijmegen Center for Molecular Life Sciences
Department of Membrane Biochemistry
Nijmegen, The Netherlands

Dr. Andreas Engel
University of Basel
Basel, Switzerland

Dr. Said Eshaghi
Stockholm University
Stockholm, Sweden

Dr. Slawomir Filipek
International Institute of Molecular and Cell Biology
Warsaw, Poland

Dr. Alastair Gardiner
Wolfson Laboratories
Imperial College
London, United Kingdom

Dr. Marie Groves
University of Leeds
Leeds, United Kingdom

Dr. Emmanuel G. Guignet
Institute of Biomolecular Sciences
Swiss Federal Institute of Technology
Lausanne, Switzerland

Dr. Frank Gunn-Moore
Imperial College
London, United Kingdom

Dr. Deborah S. Hartman
AstraZeneca Pharmaceuticals LP
Wilmington, Delaware, U.S.A.

Dr. Peter J.F. Henderson
University of Leeds
Leeds, United Kingdom

Dr. Richard Herbert
University of Leeds
Leeds, United Kingdom

Dr. Ruud Hovius
Institute of Biomolecular Sciences
Swiss Federal Institute of Technology
Lausanne, Switzerland

Dr. Mika Jormakka
Wolfson Laboratories
Imperial College
London, United Kingdom

Dr. Hans Kiefer
m-Phasys GmbH
Tuebingen, Germany

Dr. Alla Korepanova
Abbott Laboratories
Abbott Park, Illinois, U.S.A.

Christoph Krettler
Max Planck Institute for Biophysics
Frankfurt, Germany

Dr. Peter Kuhn
The Scripps Research Institute
La Jolla, California, U.S.A.

Dr. Kenneth Lundstrom
BioXtal
Epalinges, Switzerland

Mulugeta Mamo
Abbott Laboratories
Abbott Park, Illinois, U.S.A.

Dr. Johan Meuller
University of Leeds
Leeds, United Kingdom

Dr. Bruno H. Meyer
Institute of Biomolecular Sciences
Swiss Federal Institute of Technology
Lausanne, Switzerland

Anna Modzelewska
International Institute of Molecular & Cell Biology
Warsaw, Poland

Dr. Pär Nordlund
Stockholm University
Stockholm, Sweden

Dr. Thomas Ostermann
m-Phasys GmbH
Tuebingen, Germany

Dr. Simon Patching
University of Leeds
Leeds, United Kingdom

Dr. Bengt Persson
Linköping University
Linköping, Sweden

Dr. Mary Phillips-Jones
University of Leeds
Leeds, United Kingdom

John O'Reilly
University of Leeds
Leeds United Kingdom

Dr. Christoph Reinhart
Max Planck Institute for Biophysics
Frankfurt, Germany

Dr. Nick Rutherford
University of Leeds
Leeds United Kingdom

Dr. Massoud Saidijam
University of Leeds
Leeds, United Kingdom

Dr. Keigo Shibayama
University of Leeds
Leeds, United Kingdom

June Southall
University of Glasgow
Glasgow, United Kingdom

Dr. Geoffrey F. Stamper
Abbott Laboratories
Abbott Park, Illinois, U.S.A.

Dr. Raymond C. Stevens
The Scripps Research Institute
La Jolla, California, U.S.A.

Shun'ichi Suzuki
University of Leeds
Leeds, United Kingdom

Dr. Gerda Szakonyi
University of Leeds
Leeds, United Kingdom

Dr. Horst Vogel
Institute of Biomolecular Sciences
Swiss Federal Institute of Technology
Lausanne, Switzerland

Dr. Alison Ward
University of Leeds
Leeds United Kingdom

Dr. Xiang-Qun Xie
University of Houston
Houston, Texas, U.S.A.

Table of Contents

1 Introduction

Kenneth H. Lundstrom

CONTENTS

1.1 SCOPE OF BOOK

The aim of this book is to provide the reader with an overview on structural biology research and recent technology development on integral membrane proteins (IMPs). IMPs represent the last frontier in structural biology that has not been conquered. Paradoxically, although IMPs are the most important drug targets, very few high-resolution structures are available. Among the more than 30,000 structures deposited in public databases, only some 60 are on IMPs.[1] A similar situation was encountered for soluble proteins in the 1970s when structural determination methods were less advanced. Technology development led to an almost exponential increase in the number of resolved structures. However, IMPs are more complicated to handle due to their topology, which affects the inefficient transport and insertion in cellular membranes, the toxic effects of recombinant IMPs on host cells, and the instability of IMPs. As the density of IMPs in native tissue — with the few exceptions of IMPs such as the bacterial rhodopsin in *Halobacterium salinarium,*[2] the bovine rhodopsin in cow retina,[3] and the nicotinic acetylcholine receptor in the electric organ of *Torpedo marmorata*[4] — is insufficient for purification attempts directly from the native tissue, recombinant expression of IMPs is a necessity. Furthermore, large-scale isolation and purification of IMPs from human tissues for structural studies would be ethically unacceptable.

For this reason, much emphasis has been put on the overexpression of IMPs from various expression vectors. Generally, the overexpression of prokaryotic IMPs has been more successful than that of their eukaryotic counterparts.[5] The main reason is that it has been possible to overexpress them in *Escherichia coli* or alternative bacterial organisms. The complexity of prokaryotic IMPs is also lower as most post-translational modifications do exist only in eukaryotes. Chapter 3 describes the **Expression of Bacterial Membrane Proteins**, and Chapter 4 is an overview of **Prokaryotic Membrane Protein Production Strategies** as a high throughput approach. One of the drawbacks of eukaryotic IMP expression in *E. coli* has been the toxicity the foreign IMP has caused when inserted in bacterial membranes. For

this reason, an alternative strategy has been to overexpress recombinant IMPs in bacterial inclusion bodies, which has significantly reduced host-cell toxicity and improved IMP yields substantially. However, in this case the drawback has been the requirement of refolding procedures to re-establish the functionality of the IMP, which has been frustratingly inefficient.[6] Recent method development has brought some improvements to the refolding technology as described in Chapter 5 on **Refolding and Purification Technologies**. Chapter 6 is an overview of **Crystallization of Membrane Proteins** with a special emphasis on bacterial IMPs.

Eukaryotic IMPs can be divided into GPCRs, ion channels, transporters and single-transmembrane proteins, which play such important roles in cellular signaling events.[7] A variety of factors such as small molecule ligands, light, odors, ions, changes in cell membrane potential, pressure, and pH can trigger the activation of IMPs.[8] Activation of IMPs results in signal transduction cascades including changes in intracellular calcium levels, protein phosphorylation, transcriptional regulation, proliferation, and cell death. IMPs are therefore directly involved in cardiovascular, metabolic, neurodegenerative, neurological, psychiatric, and viral diseases.[9] Additionally, certain GPCRs and single-transmembrane receptors play a role in cancer development. Chapter 7 has therefore been dedicated to **Signaling through Membrane Proteins** to provide an overview of how IMPs function. Special attention is given to the **Expression of Eukaryotic Membrane Proteins** in various host systems. Chapter 8 deals with **Yeast Expression Vectors**, Chapter 9 with IMP **Expression in Insect Cells**, and Chapter 10 with the application of **Mammalian Cells** for recombinant IMP production. The downstream processing of expressed recombinant IMPs is described in Chapter 11, where methods for **Solubilization and Purification of Membrane Proteins** are outlined. Chapter 12 is dedicated to **Fluorescence Technologies**, which are applied as quality control measurements for the functionality of the overexpressed recombinant IMP. These methods are extremely important and allow fast and reliable optimization of the conditions required for expression and purification.

Chapter 14 describes the **Miniaturization of Structural Biology Technologies**, which provides a means for high-throughput screening of crystallization parameters and conditions and substantially reduces the requirement of precious purified protein. Alternative methods to x-ray crystallography are described in Chapter 13, **Membrane Proteins and NMR**, and in Chapter 16, **Applications of Electron Microscopy Technologies on Membrane Proteins**. As most efforts in structural biology today — especially structural genomics initiatives,[10] where whole gene or protein families or whole genomes are studied — require a major input of expertise in a broad range of areas such as expression, purification and crystallization, large national and international networks have been established as described in Chapter 16.

Chapter 2 provides a description of **Bioinformatics on Membrane Proteins** and *in silico* methods for drug screening. Homology studies on sequences and topologies on membrane proteins are important for a better understanding of their various functions. Chapter 17 discusses in detail the application of **Molecular Modeling** as a prerequisite for drug development programs today. In Chapter 18, **Structure-Based Drug Design**, demonstrates through examples how the drug

discovery process can be accelerated and improved with the aim of generating better and safer medicines.

1.2 SUMMARY

The structural biology of IMPs is reaching a critical stage in development. During the past 20 years less than 1% of the accumulated high-resolution structures have been represented by IMPs. This situation is really unsatisfactory as the majority of drug targets today are based on IMPs. Major technology development has taken place in the areas of recombinant protein expression, purification, and crystallization. The advances in molecular and cell biology have also substantially enhanced our understanding of the function of IMPs. However, further technology improvement is necessary. Although tens to hundreds of milligrams of recombinant proteins can be produced from bacterial, yeast, insect, and mammalian cells today to give further insight into the structural characterization of IMPs, obtaining high-resolution structures is a far more commplex process. Although efforts on certain IMPs in individual research teams have generated some success, the trend is now to establish large networks where a large number of targets, whole gene families, or genomes, can be studied in parallel. These efforts seem to be the most efficient way forward to achieve the much awaited breakthrough in structural genomics of IMPs in the near future.

REFERENCES

1. Michel, H. Membrane proteins of known structure. Max Planck Institute of Biophysics, Frankfurt, Germany. http://www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct.html, accessed 2005.
2. Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckman, E., and Downing, K.H. Model for the structure of bacteriorhodopsin based on high-resolution electron cryomicroscopy. *J. Mol. Biol.,* 213, 899–929, 1990.
3. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M., Miyano, M. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*, 289, 739–745, 2000.
4. Unwin, N. Structure and action of the nicotinic acetylcholine receptor explored by electron microscopy. *FEBS Lett,* 555, 91–95, 2003.
5. Drew, D., Fröderberg, L., Baars, L., and De Gier, J.W.L. Assembly and overexpression of membrane proteins in *Escherichia coli*. *Biochim. Biophys. Acta.,* 1610, 3–10, 2003.
6. Kiefer, H. *In vitro* folding of alpha-helical membrane proteins. *Biochim. Biophys. Acta.,* 1610, 57–62, 2003.
7. Pierce, K.L., Premont, R.T., and Lefkowitz, R.J. Seven-transmembrane receptors. *Nat. Rev.* 3, 639–650, 2002.
8. Lundstrom, K. Structural genomics of GPCRs. *Trends Biotechnol.,* 23, 103–108, 2005.
9. Drews, J. Drug discovery. *Science*, 287, 1960–1964, 2000.
10. Lundstrom, K. Structural genomics on membrane proteins: Mini review. *Comb. Chem. High Throughput Screen.,* 7, 431–439, 2004.

2 Bioinformatics in Membrane Protein Analysis

Bengt Persson

CONTENTS

2.1 INTRODUCTION

Membrane proteins are of critical importance for a wide variety of biological processes. They constitute ion channels, transport proteins, receptors for hormones, light, and odorants, just to mention a few examples. Over half of prescription drugs act on G protein-coupled receptors.[1] In the completely sequenced genomes, the proportion of genes coding for membrane proteins is estimated to be about 25%.[2–4] In spite of the biological importance of membrane proteins, there are only a few proteins for which the three-dimensional structures have been solved experimentally due to difficulties in crystallizing these proteins. Currently, only about 1 to 2% of all structures in the PDB[5] are membrane proteins.[6] Thus, there is a large gap to bridge. Bioinformatics can be of great value when it comes to identifying membrane-

spanning proteins from the amino acid sequence alone and predicting their topology, that is, delineating the transmembrane segments and the orientation of the protein in the membrane.

This chapter presents a survey of various prediction methods, most of which are available as Web services, allowing for easy and user-friendly access. Several consensus methodologies have also appeared. In addition, the chapter will review evaluations of prediction performance and availability of data sets of experimentally verified membrane proteins.

2.2 PREDICTION OF MEMBRANE-SPANNING REGIONS OF PROTEINS

Among the presently known three-dimensional structures of membrane proteins,[7] most consist of one or several membrane-spanning alpha helices with intervening short or long loops on each side of the membrane (Figure 2.1). There is also an alternative architecture with beta sheets forming a barrel inserted in the membrane,

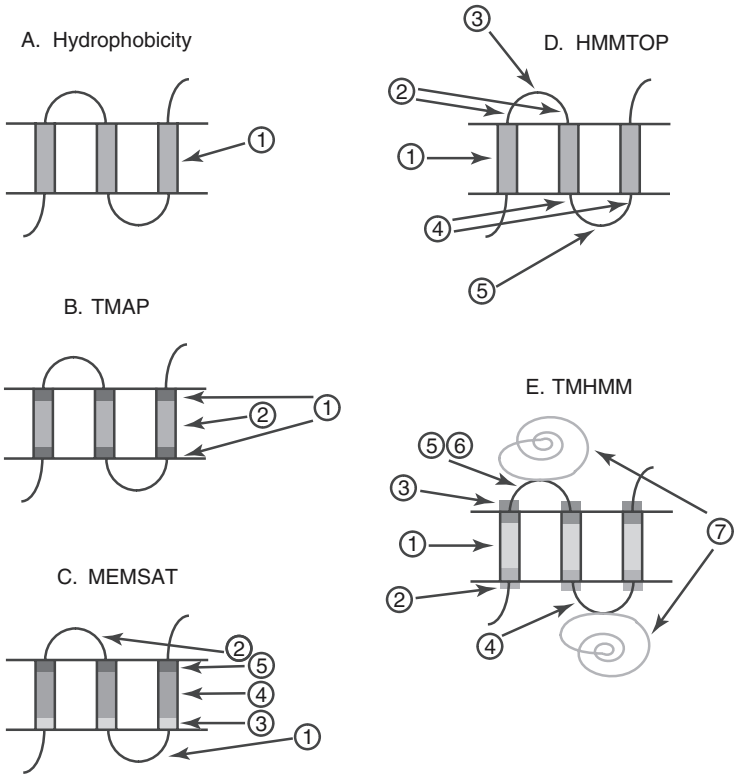


FIGURE 2.1 Schematic view of a transmembrane protein illustrating five different prediction methods. The thick vertical boxes illustrate three transmembrane helices, while the thin lines represent the intervening loops and the N- and C-terminal tails. Encircled numbers with arrows point to the segments used in the predictions (see text for the details).

but only a few proteins of this type are known so far. Most prediction algorithms are developed for alpha-helical membrane proteins, but there are algorithms for beta-barrel proteins as well. The topology prediction methods aim at identifying the membrane-spanning segments and the orientation of the protein in the membrane, that is, if the N terminus is cytosolic or noncytosolic. In the subsequent sections, the different types of prediction algorithms will be described. Most methods are available as Web servers and are listed in Table 2.1.

2.2.1 Hydrophobicity Analysis

One of the first and most basic methods to identify the membrane-spanning regions is to find hydrophobic segments in the protein sequence that are long enough to traverse the membrane (Figure 2.1A). Since the hydrophobic region corresponding to the apolar phospholipid tails of the lipid bilayer is about 30 Å,[8] these segments

TABLE 2.1
Membrane Protein Prediction Methods Available as Web Servers

Method	Web Server	Ref.
	A. Alpha-Helical Proteins	
BPROMPT	http://www.jenner.ac.uk/BPROMPT	44
ConPred	http://bioinfo.si.hirosaki-u.ac.jp/~ConPred2/	39
DAS	http://www.sbc.su.se/~miklos/DAS/	17
HMMTOP	http://www.enzim.hu/hmmtop	31
MEMSAT	http://www.psipred.net	28,29
PRED-TMR2	http://o2.db.uoa.gr/PRED-TMR/	46
PRODIV-TMHMM	http://www.sbc.su.se/PRODIV-TMHMM/	34
SOSUI	http://sosui.proteome.bio.tuat.ac.jp/sosui_submit.html	43
SVMtm	http://ccb.imb.uq.edu.au/svmtm/SVMtm_Predictor.shtml	35
THUMBUP	http://phyyz4.med.buffalo.edu/service.htm	15
TMAP	http://www.ifm.liu.se/bioinfo/services	24,25
TMFinder	http://www.bioinformatics-canada.org/TM/	12
TMHMM	http://www.cbs.dtu.dk/services/TMHMM	3
TMpred	http://www.ch.embnet.org/software/TMPRED_form.html	42
TMMOD	http://liao.cis.udel.edu/website/servers/TMMOD/	33
TopPred	http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html	23
	B. Beta-Barrel Proteins	
B2TMPRED	http://gpcr.biocomp.unibo.it/	49
Omp_topo_predict	http://strucbio.biologie.uni-konstanz.de/~kay/om_topo_predict.html	48
PRED-TMBB	http://bioinformatics.biol.uoa.gr/PRED-TMBB/	53
PROFtmb	http://cubic.bioc.columbia.edu/services/proftmb	54
TBBPred	http://www.imtech.res.in/raghava/tbbpred/	51
TMBeta-Net	http://psfs.cbrc.jp/tmbeta-net/	50

should be about 20 residues if they form an alpha helix. To identify these segments, a hydropathy plot is made by plotting the hydrophobic character against the residue number. Normally, the values are averaged over a sliding window to smooth the curve.

In 1982, Kyte and Doolittle developed a hydropathy scale for amino acid residues, which they used to calculate a moving average of segment hydrophobicity along the protein chain.[9] They showed that such hydropathy plots could be used to identify membrane-spanning regions and interior regions of globular proteins. Still, the Kyte–Doolittle method is widely used and is considered to be the standard technique for hydrophobicity analysis.

Since then, a number of hydrophobicity scales have been developed.[10] In 1990, Degli-Esposti and co-workers evaluated different hydropathy scales for prediction of membrane proteins.[11] From their comparisons, it can be seen that most hydrophobicity scales correlate quite well, even if some clearly deviate. A new average scale, denoted AMP07, was created based on seven scales.[11]

Optimal window sizes vary among methods and have also been investigated. For the Kyte–Doolittle method, it was found to be five to nine residues,[11] while in the original paper, a 19-residue window was suggested.[9] The drawbacks of long windows are the loss of local information and the risk of an apparent fusion of closely spaced transmembrane segments.[11] It should be remembered that when changing the window size, the threshold values might also have to be changed from those of the original method descriptions.

Prediction accuracy could be improved by selecting a scale developed especially for membrane proteins, which is used in several prediction algorithms (see below), where propensity values are calculated from known membrane proteins.

The method that TMFinder uses is a combination of hydrophobicity and helicity scales for transmembrane protein prediction to reliably distinguish between membrane proteins and globular proteins.[12] It uses a hydrophobicity scale that was derived from experimental properties of transmembrane-mimetic model peptides[13] and a helicity scale derived from the structural properties of these peptides.[14] The principle of TMFinder is that a transmembrane segment must show both hydrophobicity and helicity.

Recently, a scale reflecting burial propensity has been developed from a set of 200 structurally known proteins.[15] The prediction method THUMBUP is based on this scale of burial propensity and the positive inside rule, and the method has been shown to reach similar levels of accuracy as the parametrically more complicated Hidden Markow Model (HMM)-based methods.[15]

Amphiphilicity can be measured using a method developed by Eisenberg and co-workers.[16] The so-called helical hydrophobic moment can be used as a means to distinguish transmembrane helices from those of globular soluble proteins.

A variant of hydrophobicity analysis is the method known as dense alignment surface (DAS).[17] In this method, a number of low-stringency dot plot analyses are performed between the sequence to be predicted and a set of nonhomologous membrane proteins.[17,18] The putative membrane-spanning regions are seen as diagonals reflecting distant similarity.

2.2.2 The Positive Inside Rule

A major breakthrough in the prediction of membrane protein topology was the discovery by von Heijne in 1986 that there is a preponderance of positively charged residues on the cytosolic side of prokaryotic membrane proteins.[19] The rule was later found to be valid also for eukaryotic membrane proteins,[20] even if not to the same extent. Recently, a genome-wide investigation showed that the positive inside rule is detectable in all completely sequenced genomes.[21]

TopPred,[22,23] the first membrane protein topology prediction method, combines hydrophobicity analysis with the positive inside rule. The hydrophobicity analysis was performed using a trapezoid sliding window, giving more weight to the central residues than to the flanking residues. After the identification of putative transmembrane segments, the positive inside rule was applied to predict the topology. In cases where hydrophobic segments were on the borderline to be judged as transmembranous or not, the positive inside rule was used to distinguish between the alternatives.

2.2.3 Use of Multiple Sequence Alignments

The first method to use the information available from multiple sequence alignments of homologous proteins in membrane protein prediction was TMAP.[24] The idea is that by including information about the amino acid residue variation at each position, the membrane-spanning regions can be identified with higher accuracy than by using only a single sequence, similarly to a strategy applied for secondary structure predictions. The TMAP method also takes into consideration the differences in residue distributions between the central, hydrophobic part of the membrane-spanning helix and the ends, corresponding to the regions interacting with the polar head groups of the lipid bilayer (see Figure 2.1B). Two sets of propensity values are used in the prediction, derived from the statistical analysis of known membrane proteins. The use of two sets of propensity values together with the use of multiple sequences increases the prediction accuracy[24]

The TMAP method has been extended to predict the topology by analyzing the residue distributions of the loops on each side of the membrane.[25] Apart from the prominent Lys/Arg preponderance for the cytosolic side, the distributions of further residue types were also analyzed, which has led to improvements over considering the positive inside rule only.

Another method that also includes knowledge from related sequences is the PHD-htm by Rost and co-workers.[26,27] The sequence to be predicted is first used to search sequence databases for homologues, which subsequently are included in a multiple sequence alignment. From this alignment, a profile reflecting the residue distribution for each alignment position is calculated . The profile is analyzed by a neural network algorithm that estimates, for each residue, the preference to be in a transmembrane helix. After refinements of the results from the neural network, the positive inside rule is applied to predict the orientation of the protein in the membrane.

2.2.4 Model-Recognition Approaches

In 1994, Jones and co-workers developed a prediction method, MEMSAT, using five sets of propensity values, derived from proteins with known membrane topologies.[28] These values represent the statistical likelihood of each amino acid residue type to be in the structural states: inside loop, outside loop, inside helix end, helix middle, and outside helix end (Figure 2.1C). The helix end regions were set to four residues. For the protein sequence of length n to be predicted, these five scores are assigned to each residue, giving an $n \times 5$ matrix. A dynamic programming algorithm recognizes the optimal membrane topology model, revealed by the path with the highest score through this matrix. In difficult cases, the method might suggest several alternative models. The MEMSAT algorithm has been refined using information from multiple sequence alignments[29]

The modeling concept has been further developed using hidden Markov models (HMMs) in the two prediction methods HMMTOP and TMHMM. HMM is a machine learning technique that can be used for a number of modeling purposes in various fields, and HMMs have become very popular in bioinformatics.[30] When used for membrane protein predictions, an HMM is first trained on a number of known cases, during which phase all parameters of the model are optimized. Subsequently, the HMM is tested and evaluated on a test set that is different from the training set.

HMMTOP, devised by Tusnády and Simon, uses an architecture that models the membrane proteins in five different states: membrane helix, outside tail, outside loop, inside tail, and inside loop[31] (Figure 2.1D). The idea is that the residue distributions in each of these five structural parts better reflects reality than does the mere consideration of the hydrophobic character of the residues. The HMM searches possible topologies and finds the one with the maximum likelihood.

TMHMM, devised by Krogh and co-workers,[3] uses seven different states: helix core, cytosolic cap, noncytosolic cap, cytosolic loop, noncytosolic short loop, noncytosolic long loop, and globular region (see Figure 2.1E). Loops of lengths up to 20 residues are classified as "loop" regions, while longer loops are classified as "globular" regions. The cap regions correspond to the five residues at each end of the transmembrane segments.

One difficulty with transmembrane protein prediction is distinguishing a signal sequence from a transmembrane segment, since both are similar in length and hydrophobic in nature. Käll and co-workers have developed Phobius, an HMM-based predictor that combines transmembrane segment prediction with signal peptide prediction.[32] The method is based on the models in TMHMM and SignalP-HMM. Phobius makes fewer misclassifications between transmembrane segments and signal peptides compared with the TMHMM and SignalP used individually. However, Phobius is less sensitive in signal peptide detection and less accurate when predicting the cleavage site of the signal peptide.

In 2005, a third HMM-based method, TMMOD, was described.[33] This method is based on the ideas of TMHMM but differs in the modeling of loops on both sides of the membrane and in the training procedure of the HMM. The authors report a small improvement compared with the TMHMM method.

Evolutionary information from multiple sequence alignments has long since been shown to improve membrane protein predictions.[24,26] Recently, Viklund and Elofsson showed that the HMM-based predictors could also be significantly improved by inclusion of information from evolutionary related sequences.[34] They have developed a method, PRODIV-TMHMM, which can correctly predict the topology for about two-thirds of all membrane proteins. Sequence profiles are created from the multiple sequence alignments and used as input to the HMMs. The performance is increased by approximately 10 percentage units when multiple sequence information is included; this is in the same range as reported for other methods, including the multiple sequence version of HMMTOP.[34]

2.2.5 Support Vector Machines

A support vector machine (SVM) is a machine learning technique suitable for classification purposes. SVMs have also recently been used in membrane protein predictions, as is shown with the method SVMtm.[35] The authors claim a sensitivity of 93.4% and specificity of 92% for the transmembrane helix predictions. The method can distinguish transmembrane proteins from soluble proteins with 99% accuracy. In addition, the method calculates a reliability measure for each transmembrane segment.

2.2.6 Consensus Techniques

One strategy to improve the accuracy of the predictions is to combine several different methods using the consensus as the result. Such a combination was shown early on to increase the accuracy of secondary structure predictions.[36,37] To determine the consensus, it has to be decided how large the deviations can be between the segments predicted by the different methods. One consensus method uses a minimum overlap of five residues and has shown that there are small differences in the outcome using values between 1 and 10.[38] Another consensus method, in contrast, limits the differences in the distance between the mid-positions of the transmembrane segments to 11 to 15 residues.[39]

In 2000, Nilsson and co-workers reported that by using the consensus of five different prediction methods, the fraction of correctly predicted topologies could be considerably increased.[40] The five methods used were TMHMM, HMMTOP, MEMSAT, TopPred, and PHD. It was shown that when all five methods agreed, the fraction correctly predicted was 100% and that when four methods agreed, the accuracy was over 80%. These numbers can be compared with accuracies of 48 to 73% for each method used individually. However, the number of predictions decreases when applying the stringent criterion of 5:5 or 4:1 majority consensus. For *E. coli* membrane proteins, it was shown that using a threshold of at least four agreeing methods, close to half of the proteins could be predicted with high accuracy.

ConPred is another consensus prediction method[39] that uses nine different methods: KKD,[41] TMpred,[42] TopPred, DAS, TMAP, MEMSAT, SOSUI,[43] TMHMM, and HMMTOP. For topology predictions, only TMpred, TMAP, MEMSAT, TMHMM, and HMMTOP are used. The prediction accuracy reported is almost 100%. However,

the prediction coverage is only 20 to 30%. The consensus methodology is calculated to increase accuracy by up to 11 percentage units over the individual methods. The ConPred server reports the predicted model accompanied by graphical representations showing topology, hydropathy plot, and helical wheel diagram.

A third consensus prediction method is BPROMPT, which uses a Bayesian belief network to integrate the results from several Web-based predictors.[44] The methods included are HMMTOP, DAS, SOSUI, TMpred, and TopPred. The BPROMPT method is reported to arrive at a topology prediction accuracy of 70% for prokaryotes and 53% for eukaryotes.

An additional method is CoPreTHi, which combines the output from seven methods (DAS, ISREC-SAPS,[45] PHD, SOSUI, TMpred, TopPred, and PRED-TMR[46]) and considers an amino acid residue to be transmembranous if it is predicted by at least three methods.[47]

2.2.7 Prediction of Beta-Barrel Proteins

The beta-barrel membrane proteins have hitherto been characterized in bacteria, where they mediate transport of ions and small molecules. They have also been found in the organelles mitochondria and chloroplasts. The hydrophobic properties of the membrane-spanning beta strands are similar to those of soluble proteins, making predictions difficult. Furthermore, only few proteins are structurally characterized, and therefore only a small training set is available for the development of prediction methods. Consequently, there is a risk of over-training of the algorithms. So far, neural networks and hidden Markov models have been popular in prediction of beta-barrel membrane proteins.

One of the first reports on neural network methods used in topology predictions of beta-strand membrane proteins was published in 1998 by Diederichs and co-workers.[48] This method predicts the residue locations along the z-axis perpendicular to the membrane plane, where low values indicate periplasimc turns, medium values transmembrane beta-strands, and high values extracellular loops. The method was developed based on seven known structures and shown to be able to correctly predict two structures not related to the training set.[48]

Neural networks were combined with dynamic programming in a method developed by Jacoboni and coworkers.[49] They also included evolutionary information in the input to the network and achieved an accuracy of 78%. The topology prediction is based on the observation that the longest loops are at the extracellular side of the membrane.

A third example of a neural-network-based method is reported by Gromiha and co-workers.[50] Their method reports probabilities for each beta strand allowing for further interpretations after analysis. Trained on 13 known structures, this method achieved 73% prediction accuracy.

Natt and co-workers have used both neural networks and support vector machines (SVM) to predict the transmembrane regions of beta-barrel proteins.[51] They used a feed-forward neural network with a standard back-propagation training algorithm. By including information from multiple sequence alignments, they could

increase the accuracy to 80%, an improvement of the same range as for alpha-helical proteins.[34] They also developed an SVM-based method based on the amino acid sequence together with 36 physicochemical parameters. The accuracy is reported to be 77%. However, by combining the two techniques, the accuracy could be increased to 82%.

In analogy to prediction of alpha-helical membrane proteins, HMMs have also been used for beta-barrel membrane proteins. Martelli and coworkers[52] constructed a predictor using six states — beta-strand transmembrane core (two states), beta strand caps at each side of the membrane, inner loop, outer loop, and globular domain in the middle of each loop. They used input from multiple sequence alignments in the HMM to increase the accuracy, which is reported to be 83%. The information from the multiple sequence alignment is entered into the HMM as vectors, representing the sequence profile. The discriminatory ability between beta-membrane proteins and globular proteins is about 90%.

Another HMM-based method is PRED-TMBB, developed by Bagos and co-workers.[53] This HMM is a cyclic 61-state model, consisting of three submodels, representing the transmembrane strand and the inner and outer loops. The lengths of the transmembrane strands are between 7 and 17 residues. The method that was trained on 14 known proteins and tested using a jack-knife procedure shows 84% accuracy. Furthermore, PRED-TMBB discriminates beta-barrel proteins correctly from water-soluble proteins in 89% of the cases.

Bigelow and co-workers[54] have invented PROFtmb, a profile-based HMM for beta-barrel membrane proteins, with an accuracy of 86%. They have included a new definition of beta-hairpin motifs. This HMM includes 91 states representing the transmembrane beta strand in each direction, beta hairpins, inner loop, and outer loop. The discrimination between membrane proteins and soluble beta proteins is reported to be 100% at 45% coverage. The authors have applied this method on completed genomes from Gram-negative bacteria.

Finally, an alternative approach is used in the beta-barrel finder (BBF) program, which is based on analysis of the secondary structure, hydropathy, and amphipathicity of six outer membrane structures.[55] The authors have used BBF to estimate the proportion of beta-barrel membrane proteins in *E. coli* to be 2.8%. The program is available from the authors.

2.3 PREDICTION CONFIDENCE

When predicting membrane proteins, some regions are correctly predicted, while other regions are wrongly predicted. It would be of great value if the accuracy of the prediction could be estimated using a type of quality measurement, for example, that presented by PHD_htm.[26] Melén and co-workers[56] have developed reliability measures for the transmembrane prediction methods TMHMM, HMMTOP, MEMSAT, PHD, and TopPred. For TMHMM and MEMSAT, the reliability scores have been shown to correlate with prediction accuracy and will therefore add valuable information to the predictions.

2.3.1 Partial Predictions with High Accuracy

In cases when it is difficult to get correct predictions of the complete protein, it would still be valuable to get at least a partial prediction, especially if there is additional information available from elsewhere. Many times, it would also be important to know if these partial predictions are of high confidence. It has been shown that using a consensus technique with the criterion that at least four of five methods should agree gives predictions with high accuracy.[40] However, with this strict criterion, only a small number of membrane proteins will be predicted. Thus, in order to increase the number of predicted proteins, a method for prediction of partial topologies was developed using the strict criterion of the consensus methodology.[38] Partial consensus topologies could then be predicted for 60 to 70% of all proteins, on average covering 58% of the sequence length.

2.3.2 Combination of Predictions and Experimental Determination

Partial predictions can be used in combination with experimental analyses of membrane topology. For instance, the experimental determinations can be directed to those regions for which the predictions are ambiguous. Thus, a limited number of experiments combined with reliable predictions can give the complete picture of a membrane protein topology. It has been shown that a combination of experimental determination of the C-terminal location and consensus predictions can be used to give reliable topology models for *E. coli*.[57,58]

Another example of a successful combination is cases where there are two alternative predicted topologies and one experiment thereby could be sufficient to distinguish between these two models.[57] An experimentally determined C-terminal location can be used as a constraint for TMHMM to improve the outcome of the predictions.[59]

2.4 EVALUATION OF METHODS

It is of importance to try to estimate the accuracy of the available prediction methods, and therefore several evaluations have been reported.

An evaluation of different alpha-helical membrane protein prediction methods has been made by Möller and co-workers.[60] For evaluation, they used a test set of 188 membrane proteins with experimentally verified topology.[61] They measured the reliability of both transmembrane segment predictions and sidedness predictions. Overall, they found the methods TMHMM and MEMSAT to be generally the best performing. HMMTOP was best for sidedness predictions. Interestingly, Kyte–Doolittle-based analyses (KKD)[41] and analysis of hydrophobic moment [16] were quite reliable in identifying membrane-spanning regions, even though the methods lacked specificity for membrane proteins.

TMHMM and SOSUI are most reliable in not making false predictions, that is, predicting transmembrane helices in proteins not bound to the membrane. For signal peptides that often are mispredicted as transmembranous, the methods ALOM,[62] PHD, and TopPred were most successful.

The evaluation also shows that these methods have problems when predicting proteins with four or more transmembrane segments. In proteins with many membrane passages, all transmembrane segments do not need to be hydrophobic, since not all are in direct contact with the lipid bilayer. If the segments are amphiphilic, it is difficult to distinguish them from a helix at the exterior of a globular cytosolic protein. These difficulties are also seen when trying to predict the topology of the G protein-coupled receptors.[63]

In 2002, Ikeda et al. compared 10 transmembrane prediction methods on a test set of 122 experimentally characterized transmembrane topologies.[64] They also reported that methods based on HMMs and other model-based approaches were most successful. Furthermore, they noticed that generally the prediction performance is better for prokaryotic sequences than for eukaryotic ones.

In general, the methods fared less well in these evaluations than in the original reports. One major reason could be that the methods might have been "over-trained" on the proteins available at the time for development. The training set might not have been representative enough due to only a small number of proteins with known topologies being available. Thus, more recently developed techniques would be better, since they are trained on a much larger test set. However, more importantly, significant advances have been made in the recent algorithms. The early techniques only judged single properties, such as hydrophobicity, while recent algorithms have subdivided the membrane protein into several parts, using multiple parameters for the different parts of the protein (see Figure 2.1).

2.5 TEST SETS AND DATABASES OF MEMBRANE PROTEINS

A collection of experimentally characterized membrane proteins has been assembled and made publicly available by Möller and co-workers at ftp://ftp.ebi.ac.uk/databases/testsets/transmembrane.[60] The entries are human curated and annotated, depending on experimental reliability. The top level (A) consists of proteins with known three-dimensional structure, level B of proteins characterized biochemically with at least two complementary methods, followed by level C with proteins for which only basic biochemical characterization has been reported. The database will be continuously updated and is provided in Swissprot format, making it easy to use for development and evaluation of new membrane protein prediction algorithms.

TMPDB is another database of experimentally characterized membrane protein topologies. The release of 2003 contained over 300 proteins, of which the vast majority was of the alpha-helical type.[65] TMPDB is based on information from examination of scientific articles and sequence and structure databases. The data are valuable for all scientists developing and optimizing new methods for transmembrane protein predictions. The database is available at http://bioinfo.si.hirosakiu.ac.jp/~TMPDB/.

PDB_TM is a database of transmembrane proteins with known structures, extracted from PDB. In PDB_TM, the membrane-spanning segments are determined

using the TMDET algorithm[66] for calculation of the position of the protein in the lipid bilayer. The PDB_TM is updated weekly and is available at http://www.enzim.hu/PDB_TM.

2.6 CONCLUSION

Even if today's methods for membrane protein prediction are quite accurate and these structural predictions are among the most successful in bioinformatics, there is still much room for improvement. For development and training of the algorithms, the number of experimentally determined structures is still far too low, which might lead to methods that are biased and lack generality. Hopefully, the ongoing structural genomics initiatives worldwide will contribute to a considerable increase in the number of available structures. Also, large-scale experimental topology mappings will add important knowledge regarding membrane protein properties that can be used in new methods.

Hitherto, most methods have been based on neural networks and HMMs, but now methods based on support vector machines have started to appear. It can be anticipated that further sophisticated machine learning techniques will be used in membrane protein predictions. It is also likely that various combinations of these techniques will increase reliability. Thus, more training data together with improved prediction algorithms will hopefully help approach 100% accuracy.

REFERENCES

1. Attwood, T.K. A compendium of specific motifs for diagnosing GPCR subtypes. *Trends Pharmacol. Sci.,* 22, 162–165, 2001.
2. Wallin, E. and von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.,* 7, 1029–1038, 1998.
3. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580,, 2001.
4. Stevens, T.J. and Arkin, I.T. Do more complex organisms have a greater proportion of membrane proteins in their genomes? *Proteins*, 39, 417–420, 2000.
5. Berman, H.M., Westbrook, J., Feng, Z., et al. The Protein Data Bank. *Nucleic Acids Res.,* 28, 235–242, 2000.
6. Chen, C.P. and Rost, B. State-of-the-art in membrane protein prediction. *Appl. Bioinformatics,* 1, 21–35, 2002.
7. Tusnady, G.E., Dosztanyi, Z., and Simon, I. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucl. Acids Res.,* 33, D275–D278, 2005.
8. Lewis, B.A. and Engelman, D.M. Lipid bilayer thickness varies linearly with acyl chain length in fluid phosphatidylcholine vesicles. *J. Mol. Biol.,* 166, 211–217, 1983.
9. Kyte, J. and Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.,* 157, 105–132, 1982.

10. Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A., and DeLisi, C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.,* 195, 659–685, 1987.
11. Degli-Esposti, M., Crimi, M., and Venturoli, G. A critical evaluation of the hydropathy profile of membrane proteins. *Eur. J. Biochem.,* 190, 207–219, 1990.
12. Deber, C.M., Wang, C., Liu, L.P., et al. TM Finder: A prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Sci.,* 10, 212–219, 2001.
13. Liu, L.P. and Deber, C.M. Guidelines for membrane protein engineering derived from *de novo* designed model peptides. *Biopolymers*, 47, 41–62, 1998.
14. Liu, L.P. and Deber, C.M.. Uncoupling hydrophobicity and helicity in transmembrane segments. Alpha-helical propensities of the amino acids in non-polar environments. *J. Biol. Chem.,* 273, 23645–23648, 1998.
15. Zhou, H. and Zhou, Y. Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. *Protein Sci.,* 12, 1547–1555, 2003.
16. Eisenberg, D., Weiss, R.M., Terwilliger, T.C. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, 299, 371–374, 1982.
17. Cserzo, M., Wallin, E., Simon, I., et al. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. New alignment strategy for transmembrane proteins. *Protein Eng.,* 10, 673–676, 1997.
18. Cserzo, M., Bernassau, J.M., Simon, I., and Maigret, B. New alignment strategy for transmembrane proteins. *J. Mol. Biol.,* 243, 388–396, 1994.
19. von Heijne, G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.,* 5, 3021–3027, 1986.
20. Sipos, L. and von Heijne, G. Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.,* 213, 1333–1340, 1993.
21. Nilsson, J., Persson, B., and von Heijne, G. Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes. *Proteins*, 60, 606–616, 2005.
22. von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.,* 225, 487–494, 1992.
23. Claros, M.G. and von Heijne, G. TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.,* 10, 685–686, 1994.
24. Persson, B. and Argos, P. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.,* 237, 182–192, 1994.
25. Persson, B. and Argos, P. Topology prediction of membrane proteins. *Prot. Sci.,* 5, 363–371, 1996.
26. Rost, B., Casadio, R., Fariselli, P., and Sander, C. Transmembrane helices predicted at 95% accuracy. *Prot. Sci.,* 4, 521–533, 1995.
27. Rost, B., Fariselli, P., and Casadio, R. Topology prediction for helical transmembrane proteins at 86% accuracy. *Prot. Sci.,* 5, 1704–1718, 1996.
28. Jones, D.T., Taylor, W.R., and Thornton, J.M. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 33, 3038–3049, 1994.
29. McGuffin, L.J., Bryson, K., and Jones, D.T. The PSIPRED protein structure prediction server. *Bioinformatics*, 16, 404–405, 2000.

30. Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.,* 235, 1501–1531, 1994.
31. Tusnady, G.E. and Simon, I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.,* 283, 489–506, 1998.
32. Kall, L., Krogh, A., and Sonnhammer, E.L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.,* 338, 1027–1036, 2004.
33. Kahsay, R.Y., Gao, G., and Liao, L. An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. *Bioinformatics,* 21, 1853–1858, 2005.
34. Viklund, H. and Elofsson, A. Best (alpha)-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Prot. Sci.,* 13, 1908–1917, 2004.
35. Yuan, Z., Mattick, J.S., and Teasdale, R.D. SVMtm: support vector machines to predict transmembrane segments. *J. Comput. Chem.,* 25, 632–636, 2004.
36. Schulz, G.E., Barry, C.D., Friedman, J., Chou, P.Y., Fasman, G.D., Finkelstein, A.V., Lim, V.I., Ptitsyn, O.B., Kabat, E.A., Wu, T.T., Levitt, M., Robson, B., and Nogano, K. Comparison of predicted and experimentally determined secondary structure of adenyl kinase. *Nature*, 250, 140–142, 1974.
37. Argos, P. and Schwarz, J. An assessment of protein secondary structure prediction methods based on amino acid sequence. *Biochim. Biophys. Acta.,* 439, 261–273, 1976.
38. Nilsson, J., Persson, B., and von Heijne, G. Prediction of partial membrane topologies using a consensus approach. *Prot. Sci.,* 11, 2974–2980, 2002.
39. Arai, M., Mitsuke, H., Ikeda, M., Xia, J.X., Kikuchi, T., Satake, M., and Shimizu, T. ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucl. Acids Res.,* 32, W390–W393, 2004.
40. Nilsson, J., Persson, B., and von Heijne, G. Consensus predictions of membrane protein topology. *FEBS Lett.,* 486, 267–269, 2004.
41. Klein, P., Kanehisa, M., and DeLisi, C. The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta.,* 815, 468–-476, 1985.
42. Hofmann, K. and Stoffel, W. TMbase — a database of membrane spanning proteins segments. *Biol. Chem. Hoppe-Seyler,* 347, 166, 1993.
43. Hirokawa, T., Boon-Chieng, S., and Mitaku, S. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14, 378–379, 1998.
44. Taylor, P.D., Attwood, T.K., and Flower, D.R. BPROMPT: a consensus server for membrane protein prediction. *Nucl. Acids Res.,* 31, 3698–3700, 2003.
45. Brendel, V., Bucher, P., Nourbakhsh, I.R., Blaisdell, B.E., and Karlin, S. Methods and algorithms for statistical analysis of protein sequences. *Proc. Natl. Acad. Sci. USA,* 89, 2002–2006, 1992.
46. Pasquier, C., Promponas, V.J., Palaios, G.A., Hamodrakas, J.S., and Hamodrakas, S.J. A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. *Prot. Eng.,* 12, 381–385, 1999.
47. Promponas, V.J., Palaios, G.A., Pasquier, C.M., Hamodrakas, J.S., and Hamodrakas, S.J. CoPreTHi: a Web tool which combines transmembrane protein segment prediction methods. *In Silico. Biol.,* 1, 159–162, 1999.

48. Diederichs, K., Freigang, J., Umhau, S., Zeth, K., and Breed, J. Prediction by a neural network of outer membrane (beta)-strand protein topology. *Prot. Sci.,* 7, 2413–2420, 1998.
49. Jacoboni, I., Martelli, P.L., Fariselli, P., De Pinto, V., and Casadio, R. Prediction of the transmembrane regions of (beta)-barrel membrane proteins with a neural network-based predictor. *Prot. Sci.,* 10, 779–787, 2001.
50. Gromiha, M.M., Ahmad, S., and Suwa, M. Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins. *J. Comput. Chem.*, 25, 762–767, 2004.
51. Natt, N.K., Kaur, H., and Raghava, G.P. Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins*, 56, 11–18, 2004.
52. Martelli, P.L., Fariselli, P., Krogh, A., and Casadio, R. A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, 18 Suppl 1, S46–S53, 2002.
53. Bagos, P.G., Liakopoulos, T.D., Spyropoulos, I.C., and Hamodrakas, S.J. A Hidden Markov Model method, capable of predicting and discriminating beta-barrel outer membrane proteins. *BMC Bioinformatics*, 5, 29, 2004.
54. Bigelow, H.R., Petrey, D.S., Liu, J., Przybylski, D., and Rost, B. Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res.,* 32, 2566–2577, 2004.
55. Zhai, Y. and Saier, M.H., Jr. The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. *Prot. Sci.* 11, 2196–2207, 2002.
56. Melen, K., Krogh, A., and von Heijne, G. Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.,* 327, 735–744, 2003.
57. Drew, D., Sjostrand, D., Nilsson, J., Urbig, T., Chin, C.N., de Gier, J.W., and von Heijne, G. Rapid topology mapping of *Escherichia coli* inner-membrane proteins by prediction and PhoA/GFP fusion analysis. *Proc. Natl. Acad. Sci. U.S.A.,* 99, 2690–2695, 2002.
58. Daley, D.O., Rapp, M., Granseth, E., Melen, K., Drew, D., and von Heijne, G. Global topology analysis of the *Escherichia coli* inner membrane proteome. *Science*, 308, 1321–1323, 2005.
59. Kim, H., Melen, K., and von Heijne, G. Topology models for 37 *Saccharomyces cerevisiae* membrane proteins based on C-terminal reporter fusions and predictions. *J. Biol. Chem.,* 278, 10208–10213, 2003.
60. Möller, S., Croning, M.D., and Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, 17, 646–653, 2001.
61. Möller, S., Kriventseva, E.V., and Apweiler, R. A collection of well characterised integral membrane proteins. *Bioinformatics*, 16, 1159–1160, 2000.
62. Nakai, K. and Kanehisa, M. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, 14, 897–911, 1992.
63. Ji, T.H., Grossmann, M., and Ji, I. G protein-coupled receptors. I. Diversity of receptor-ligand interactions. *J. Biol. Chem.,* 273, 17299–17302, 1998.
64. Ikeda, M., Arai, M., Lao, D.M., and Shimizu, T. Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. *In Silico. Biol.,* 2, 19–33, 2002.
65. Ikeda, M., Arai, M., Okuno, T., and Shimizu, T. TMPDB: a database of experimentally-characterized transmembrane topologies. *Nucl. Acids Res.,* 31, 406–409, 2003.
66. Tusnady, G.E., Dosztanyi, Z., and Simon, I. TMDET: Web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, 21, 1276–1277, 2005.

3 Prokaryotic Membrane Transport Proteins: Amplified Expression and Purification

Joanne Clough, Massoud Saidijam, Kim Bettaney, Gerda Szakonyi, Simon Patching, Johan Meuller, Shun'ichi Suzuki, Keigo Shibayama, Mark Bacon, Emma Barksby, Marie Groves, Richard Herbert, Mary Phillips-Jones, Alison Ward, Frank Gunn-Moore, John O'Reilly, Nick Rutherford, Roslyn Bill, and Peter Henderson

CONTENTS

3.1 INTRODUCTION

Membrane transport proteins are involved in nutrient capture, antibiotic efflux, protein secretion, toxin production, photosynthesis, oxidative phosphorylation, envi-

ronmental sensing, and other vital functions in bacteria (Figure 3.1). Already there is commercial interest in inhibiting the activities of some membrane transport proteins, optimizing the activities of others, employing them as transducers of electrical/chemical/mechanical energy for nanotechnology, and so on. However, membrane proteins are notoriously difficult to study. Owing to their extreme hydrophobicity, they are refractory to direct manipulation and can only be removed from the membrane, and their solubility maintained, in the presence of a detergent.[1] In addition, transport proteins are usually only expressed at low levels and constitute less than 0.1% of total cell protein. Such difficulties help explain why fewer than 100 unique membrane protein structures have been resolved (see relevant examples in References[2,3]), although the structures of over 8000 unique soluble proteins (from almost 30,000 total structures, many not unique) have been solved. In fact, less than 1% of unique structures in the Protein Structures Database are membrane proteins, whereas they account for about 30% of all proteins in the cell.[2,3]

Prokaryote membrane transport proteins fall predominantly into two classes.[4,5] One of these uses adenosine triphosphate (ATP) to energize the transport of substrates across the membrane — the "ATP-Binding Cassette" (ABC) superfamily[6,7]

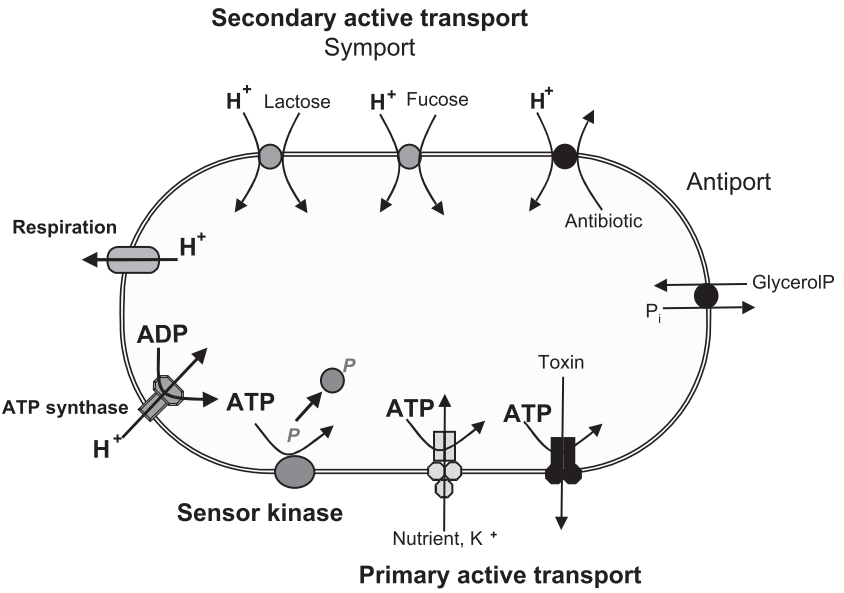


FIGURE 3.1 Active transport systems in bacteria. The large oval represents the cytoplasmic membrane of the microorganism. A transmembrane electrochemical gradient of protons is generated by respiration or ATP hydrolysis, shown on the left. The proton gradient may be used to drive ATP synthesis and the proton–nutrient symport and proton–substrate antiport secondary active transport systems shown along the top; alternatively, sodium (not shown) or phosphate (right side) may be the accompanying ions. Each is generally a single protein, usually of the 12-helix type. Along the bottom are illustrated primary active transport systems energized directly by ATP and a sensor kinase system.

— and the second, the "Major Facilitator Superfamily" (MFS)[8,9] is usually energized by the electrochemical gradient of protons (Figure 3.1) or sometimes sodium or phosphate ions (Figure 3.1). Members of the ABC superfamily and other types of transport systems[5] are not considered further in this chapter, which focuses mainly on the MFS transport proteins. These are found in nearly all organisms, from cyanobacteria to humans.[10,11]

In bacteria, individual MFS proteins may accomplish: the active accumulation of nutrients by a cation–substrate symport mechanism (Figure 3.1); the active efflux of compounds such as antibiotics, antibacterials, or toxins by a cation–substrate antiport mechanism (Figure 3.1); or substrate/substrate antiport reactions (Figure 3.1). They are thought to be single polypeptides comprising 10 to 14 (usually 12) transmembrane alpha helices,[12] illustrated for the *Escherichia coli* "FucP" protein in Figure 3.2. This conclusion is usually based on analysis of the hydropathic profile of the amino acid sequence of each protein predicted from its DNA sequence. In a few cases, the prediction is reinforced by genetic, immuno-chemical or other types of topological experiments.[13] There is structural information consistent with the 12-helix composition from electron diffraction analyses of two-dimensional protein crystals.[14–16] A spectacular confirmation came from x-ray diffraction analysis of three-dimensional crystals of three MFS proteins from *E. coli:* the lactose-H^+ symporter LacY[17], the glycerol-P/Pi antiporter GlpT[18] (Figure 3.1), and the Na^+/H^+ antiporter NhaA.[19]

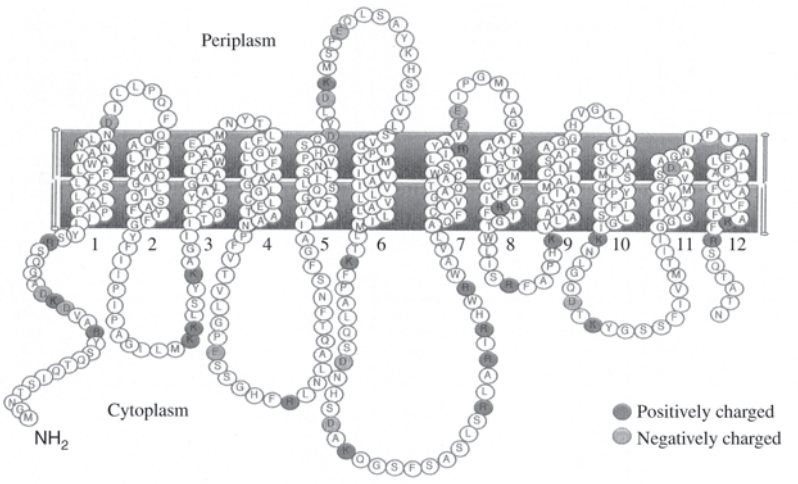


FIGURE 3.2 (**See color insert following page 272.**) A two-dimensional model for the folding of the FucP protein of *Escherichia coli* in the cell membrane based on hydropathy plot, positive inside rule, and beta-lactamase fusions. Positive and negative residues are highlighted in gray and black, respectively. From Gunn, F., Tate, C.G., Sansom, C.E., Henderson, P.J.F. *Molec. Microbiol.*, 15, 771–783, 1995; and Clough, J.L. Ph.D. Thesis, University of Leeds, 2001. With permission.

To enable the determination of structures of membrane transport proteins, a continuing supply of milligram quantities of protein is required. As native expression levels are usually less than 0.1% of total cell protein, heterologous gene expression is a necessity. Even if this approach is successful, a suitable detergent must be found for purification. The protein may also require "conformation locking" to overcome the probable flexibility of transport proteins, which is invoked to account for their ability to bind substrate in or on one side of the membrane, and effect its translocation. The strategy needed to purify sufficient active protein is therefore complex.

A general approach has been devised for the amplified expression, purification, and characterization of bacterial membrane transport proteins (Table 3.1) in *E. coli*. The strategy is described in this chapter, using the L-fucose-H^+ symport protein, FucP[20] (Figure 3.2) as an example to facilitate future examination of the large number of transport proteins arising from genome analyses; these proteins have potential for development of novel antibacterials and perhaps applications in biotechnology. L-Fucose — 6-deoxy-L-galactose — is reasonably abundant in nature as the breakdown product of plant cell wall polysaccharides and is used as a carbon source by free-living bacteria.[20] So far, the strategy has been successful for over 30 prokaryotic transporters, including MFS transport proteins from *E. coli, Brucella abortus, B. melitensis, Helicobacter pylori*, *Microbacterium liquefaciens, Enterococcus faecalis, Bacillus subtilis, Staphylococcus aureus, Campylobacter jejuni,* and *Neisseria meningitides* (Table 3.1). Proteins produced in this way have been tested in crystallization trials, and so far, two have yielded diffracting crystals.

Other organisms that appear to be successful for the propagation of vectors and expression of heterologous prokaryotic membrane proteins are *Lactococcus lactis*, *Streptococcus thermophilus,*[21] and *Halobacterium salinarum*.[22]

3.2 *E. coli* EXPRESSION SYSTEMS

To facilitate the study of prokaryotic membrane proteins, numerous *E. coli* expression systems have been used,[23,24] with levels of expression as high as 50% of inner membrane protein[24] and 80% of outer membrane protein.[25]

The pET system (Novagen), which is widely used for the expression of soluble proteins, has also led to the amplified production of membrane proteins.[26–30] In pET vectors, the strong bacteriophage T7 promoter is recognized by T7 but not by *E. coli* RNA polymerase. Expression from pET vectors is achieved by transforming the recombinant plasmid into a host strain that carries a chromosomal copy of T7 RNA polymerase. For the usual *E. coli* (DE3) lysogen host strains, the T7 RNA polymerase is under the control of the lacUV5 promoter,[30] thereby allowing some expression of T7 RNA polymerase even in the absence of an inducer. In the case of the *fucP* gene, a construct exploiting the T7 system (Figure 3.3A) did yield expression but not amplification of the protein.[31] However, constructs exploiting transcription from the lambda leftward promoter (plasmid AD5827)[31–33] (Figure 3.3A) or the *tac* promoter (plasmid pTTQ18)[34] (Figure 3.3A) — which in the absence of isopropyl-β-D-thiogalactoside (IPTG) is repressed by the plasmid-encoded *lac* repressor (Figures 3A,B),[34–35] — were successful.[31,36] We have now successfully overexpressed more than 30 prokaryotic membrane transport proteins in *E. coli* using the plasmid pTTQ18