UPTEC X 06 047
DEC 2006

ISSN 1401-2138

LINN FAGERBERG

Bioinformatic analysis of human membrane proteins for antibody-based proteomics

Master's degree project

UPPSALA UNIVERSITET

Bioinformatics Program

Uppsala University School of Engineering

UPTEC X 06 047	Date of issue 2006-12

Author

Linn Fagerberg

Title (English)

Bioinformatic analysis of human membrane proteins for antibody-based proteomics

Title (Swedish)

Abstract

Membrane proteins are important targets for the pharmaceutical industry and therefore in focus for antibody-based proteomics efforts such as the HPA program. In this project, prediction methods for membrane protein topology have been assessed, and six methods were selected for implementation into an antigen selection software. A pilot study for validation of the selected methods was performed by flow cytometry using HPA antibodies.

Keywords

proteomics, antibody, membrane protein, topology prediction methods, flow cytometry

Supervisors

Mathias Uhlén
School of Biotechnology, Royal Institute of Technology

Scientific reviewer

Erik Sonnhammer
Stockholm Bioinformatics Center

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **38**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Bioinformatic analysis of human membrane proteins for antibody-based proteomics

Linn Fagerberg

Sammanfattning

År 2001 blev sekvensen för det humana genomet tillgänglig och sedan dess har målet varit att analysera de proteiner som generna kodar för. I det stora svenska projektet Human Protein Atlas (HPA) försöker detta uppnås genom tillverkning av antikroppar som binder specifikt till ett protein. Varje antikropp färgas sedan in för att undersöka var proteinet den binder till är uttryckt i olika vävnader, cancertyper och cellinjer. HPA har producerat antikroppar från olika sorters proteinfamiljer men ska framöver fokusera främst på membranproteiner som sitter i cellmembranet.

Dessa membranproteiner består av intracellulära och extracellulära delar. Deras funktion kan bland annat vara att fungera som transportkanaler genom cellmembranet, eller som receptorer som tar emot signaler från andra celler och signalmolekyler, vilket gör dem till intressanta läkemedelskandidater. För den här typen av proteiner finns mycket få experimentella strukturer och därför används prediktionsmetoder för att predicera topologin, det vill säga vilka delar av proteinsekvensen som sitter i membranet, respektive intra- och extracellulärt. I detta projekt undersöktes flera prediktionsmetoder för att hitta den mest passande kombinationen att använda i HPA projektet. En validering av resultatet från de valda metoderna gjordes genom att analysera antikropparnas bindning till celler med hjälp av flödescytometri.

Examensarbete 20 p i Bioinformatikprogrammet
Uppsala universitet December 2006

Table of contents

1. Introduction

The sequence of the human genome became available in 2001[8] and created a new world of possibilities for research in biomedical fields. One of the new challenges in the post-genome era is to perform systematic analysis of the proteins encoded by the genes in the genome, an approach called proteomics[9]. One way to analyze the proteome is by genome-based proteomics, where the approach is based on a gene-by-gene analysis. The goal is to get a "catalogue" of relevant characteristics, such as structure, function, interaction, localization and expression, for each protein encoded by the human genome.[10]

The Human Protein Atlas (HPA) program[11-13] has been set-up to allow for the systematic exploration of the human proteome with antibody-based proteomics, which involves the generation of protein-specific polyclonal antibodies for functional exploration of the human proteome. The program combines high-throughput generation of affinity-purified antibodies with protein profiling using tissue arrays. The aim is to obtain a protein expression and subcellular localization profile for one representative protein from every gene locus in the Ensembl human genome database.[14] There are numerous areas where the generated antibodies can be applied, e.g. pull-out experiments, *in vitro* and *in vivo* protein profiling, and protein assays such as ELISA or protein arrays.[12]

The strategy of the HPA program is based on the generation of Protein Epitope Signature Tags (PrESTs)[15]. A PrEST is a fragment of a protein suitable for protein expression and with low sequence identity to other human proteins. The PrESTs are used both as antigens for generation of polyclonal antibodies, and as affinity ligands in the purification of antibodies to obtain specificity.[11] Information generated from the project is available in a public database (*http://www.proteinatlas.org*) [13, 16]. The current version 2.0, released 2006-10-30, contains 1514 antibodies representing all major types of protein families, such as transcription factors, protein receptors, nuclear receptors, kinases, and phosphatases.[11] For the next phase, the HPA program will focus on the large group of human membrane proteins.

Membrane proteins are crucial for many biological functions.[17] They constitute around one fourth of the human proteins and are involved in signalling, energy-transfer, ion-transport, cell-cell interactions, nerve impulses and more. Membrane proteins are targets for more than 45% of all pharmaceutical drugs [18] and are thus of utter importance for the pharmacological industry.[19, 17] Generation of polyclonal antibodies against plasmamembrane-spanning proteins for pharmaceutical and other *in vivo* purposes requires intelligent selection of antigen to ensure targeting of exposed extracellular domains of the proteins. The structure of most membrane proteins remains unknown due to the technical difficulties to experimentally analyze them, and membrane proteins only account for 1% of the proteins with known structures.[17, 20] The use of bioinformatical methods is crucial for obtaining more information, and an essential characteristic for membrane protein structure prediction is the topology, which can be defined as the identification of transmembrane helices and their overall in/out orientation relative to the membrane.

Today there exist numerous prediction methods for membrane protein topology, based on hydrophobicity analysis, statistical models and other techniques such as hidden Markov models. In this project, the weaknesses and strengths of various prediction methods have been assessed in order to find a reliable approach to predict the topology of membrane proteins and discriminate between soluble proteins and membrane proteins for the purposes of the HPA program. To be able to select the most suitable prediction method for selection of PrESTs on membrane proteins, it is essential to learn more about the biology of membrane proteins, such as how they are inserted into the membrane and how they can be categorized. An understanding of how the sequence of a membrane protein can be distinguished from non-membrane proteins is important to correctly evaluate the different prediction methods. An appropriate PrEST design strategy is needed for generation of antibodies against membrane proteins and the selected prediction method must be incorporated into the PrEST selection software used in the HPA pipeline.

The HPA program has already been able to generate antibodies towards a number of proteins predicted to be located in the membrane. The knowledge of the exact PrEST position in the target protein can be exploited for experimental validation of membrane topology predictions by, for example, flow cytometry in Fluorescent Activated Cell Sorting (FACS). The results can provide information about the intracellular/extracellular location of the specific fragment of the protein towards which the HPA antibody was generated and can be used to validate the output from various prediction methods, but also adds important information to the limited knowledge about the structure of membrane proteins.

1.1 Aim of project

The objective of this master thesis is to gain more knowledge about the human membrane proteins and develop bioinformatics tools to generate optimal affinity reagents for proteomics research. The four major goals are:

1. Assessment of methods for prediction of membrane protein topology
2. Preparation of PrEST design criteria for membrane proteins
3. Implementation of prediction methods and design criteria into a PrEST design software
4. Validation of bioinformatic methods for membrane protein topology prediction through analysis of FACS results from selected HPA antibodies

2. The Human Protein Atlas program

The Swedish Human Protein Atlas (HPA) program is funded by the Knut and Alice Wallenberg Foundation. The program is run by the Human Proteome Resource center (HPR) and has two major sites. The Stockholm site, located at the AlbaNova University Center at the Royal Institute of Technology, is responsible for the generation of high-quality monospecific antibodies, involving methods such as high-throughput cloning, expression of the protein fragments, affinity purification and quality assurance of the antibodies. The large-scale profiling of proteins in cells and tissues using immunohistochemical methods, and the annotation and generation of digital images, take place at the Rudbeck Laboratory, Uppsala University, in Uppsala.[16]

The resource centre has two main objectives: i) to produce specific antibodies to human target proteins, and ii) to produce a public Protein Atlas with histological images to obtain the location of each human protein in various tissues. The HPA program has generated a large set of affinity ligands, in the form of monospecific antibodies, using a combination of bioinformatics, recombinant protein expression, and cost-effective antibody production.[15] These antibodies have been shown to be valuable tools to explore protein expression profiles using human tissue arrays, and allow for the systematic approach used in the HPA program to generate and use antibodies as affinity reagents.[15]

2.1 Antibody-based proteomics

Affinity proteomics can be defined as "the systematic generation and use of protein-specific affinity reagents to functionally explore the genome".[10] Affinity reagents can for example be used *in vivo* for histochemistry analysis and *in vitro* for various pull-out experiments of certain proteins. They need to be specific, sensitive and quantitative, and the three major types are monoclonal antibodies (mAbs), polyclonal antibodies (pAbs), and monoclonal binding reagents such as affibody molecules or Fabs.[10]

Monoclonal antibodies are identical, homogenous and produced by a single clone of B-cells from antibody-producing animal cells.[21] One limitation is that they only recognize a single epitope, which makes them difficult to use in some platforms, e.g. assays where proteins are denatured. Generation of mAbs is also time consuming, making them unsuitable for large-scale purposes.[11] However, for diagnostic applications, mAbs are currently the most commonly used type of antibody.[10] Polyclonal antibodies are a mixture of antibodies that recognize different epitopes of the same antigen and are generated by immunization in animals.[22] The weakness of pAbs is that polyclonal serum from an animal is irreproducible and unique. A significant advantage of pAbs is that multiple antibodies recognize a single target[10] which makes them suitable for cross-platform assays that involve proteins both in a native and denatured form.[11]

The choice to use polyclonal instead of monoclonal antibodies in the HPA program has been made for two reasons. First, the generation is relatively cost-effective compared to the generation of monoclonal antibodies. Second, the probability of specific recognition of the target protein during various denaturing conditions is increased by the use of polyclonal antibodies.[15] The possibility of using polyclonal antibodies as reagents in the HPA program instead of monoclonal antibodies is dependent on sufficient purification to achieve specificity. This development has lead to a new type of antibodies called monospecific antibodies (msAbs), generated from pAbs using antigen-specific purification.[12]

2.2 PrEST design

The HPR center consists of eight different modules that perform the different steps of the project. The first module in the pipeline is the bioinformatic **PrEST design module**. Since this master thesis project is closely related to the PrEST design module, it will be reviewed in more detail than the other modules.

The main aim of the PrEST design module is to perform computer-based analysis of protein sequences to be able to select a suitable PrEST. Data from the Ensembl database[14] is used to obtain sequences and other information about the genes and proteins. Different protein families are selected by the HPA program to be used for the generation of antibodies, and additional lists of interesting genes are obtained through collaborations with external researchers from all around the world.

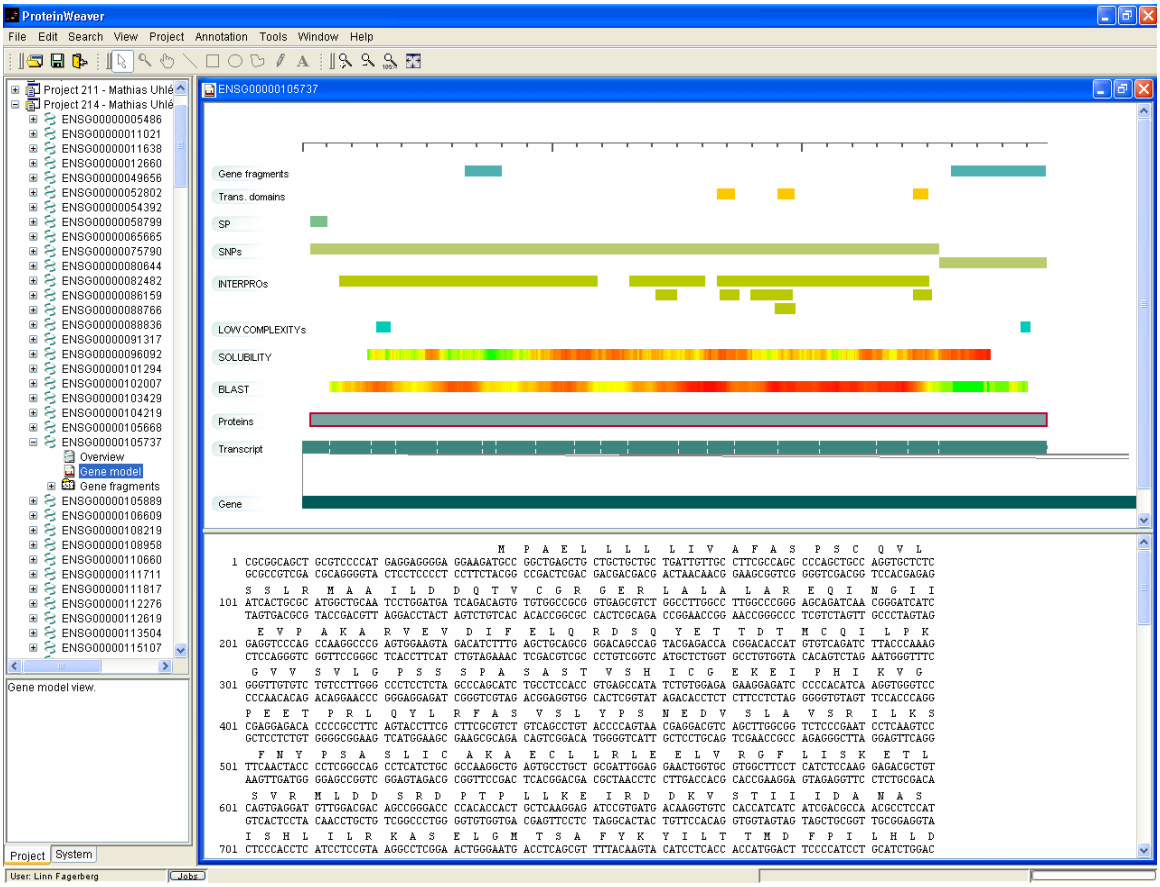

Figure 1. Screenshot from the ProteinWeaver software.

One representative protein for each gene is selected for PrEST design, and the goal is to select two PrESTs per gene. The average length of the protein fragments is ~115 amino acids. This size is small enough to be easily handled in RT-PCR and cloning, and yet large enough to contain multiple epitopes.[23] Several protein features have to be considered when trying to find the most suitable PrEST for generation of optimal protein production and purification and to ensure a good immune response. Examples of such features are sequence identity to other human proteins, transmembrane regions, signal peptides, restriction enzyme sites, certain Interpro domains, and low complexity sequence regions. When genes with alternative splicing are designed, regions that are common or unique to the splice variants also must be considered. The software currently used for PrEST design is a visualization tool named ProteinWeaver (Affibody AB, Stockholm, Sweden). In Figure 1, an example of the graphical view of most of these features is shown.

In the attempt to select the most suitable sequence region for a PrEST, transmembrane regions of proteins are omitted since they are not accessible for the antibodies in *in vivo* studies, but also because they might be difficult to express and purify.[15] Similarly, signal peptides are omitted because they are cleaved off and will not be part of the mature protein.[23] Two restriction enzymes, Asc1 and Not1, are used in the cloning procedure, and thus the restriction sites for these enzymes have to be avoided. For the protein to be expressed in *E. coli*, the selected transcript sequence must be in the correct frame.

The selection of protein fragments with low identity to other proteins of the human proteome is important for the mono-specificity of the generated antibodies.[12] The PrESTs may not always fold like the native protein, but most of the subsequent protein profiling in the HPA program is based on denaturated proteins. The use of polyclonal antibodies, and the fact the PrEST often is long enough to contain multiple epitopes, increase the probability that some epitopes are present during the various conditions for the procedures that the PrESTs are used in.

When a PrEST has been selected, a PrEST-specific oligonucleotide primer pair is designed and ordered, and data for the position and sequence of the PrEST and the corresponding primers is stored in the HPA database. The primers are used for the RT-PCR (Reverse Transcription Polymerase Chain Reaction) amplification performed in the next step of the HPA pipeline. The PrEST design module continuously analyzes results from all subsequent modules in order to improve the success rates and further develop the design strategy. Today more than 17000 PrESTs have been designed on ~10000 genes. The current Ensembl version 41.36 contains 23224 genes coding for 48403 proteins. Near 50% of the Ensembl genes have been analyzed in the PrEST design module, with a PrEST selection success rate of 90%.

2.3 HPR pipeline

The pipeline of the HPA program generates several products; the most important being antigens, monospecific antibodies (msAbs) and TMA images.[13] This section contains a short summary of the different modules, and more information about the materials and methods can be found on the project's webpage (*http://www.proteinatlas.org*) and in the listed publications.[16] Figure 2 shows a schematic illustration of the pipeline.

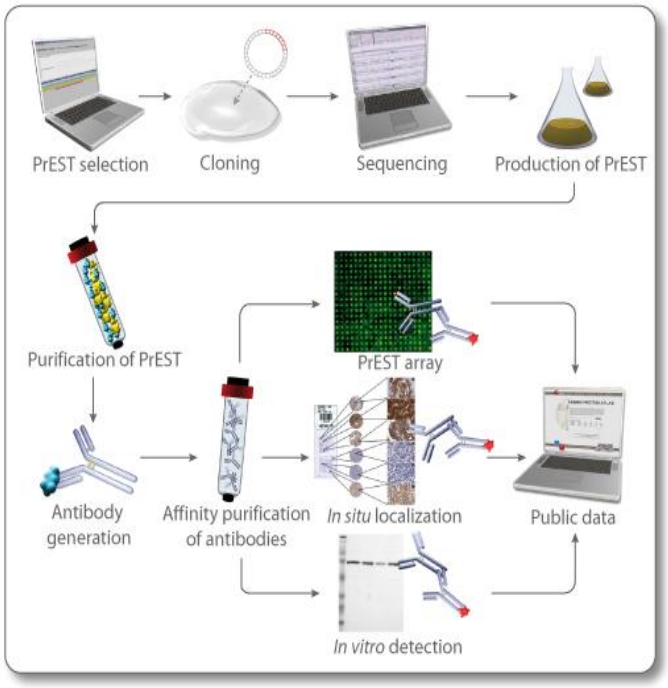

Figure 2. Schematic illustration of the HPR pipeline. Image used with permission from Mats Lindskog. [1]

The final product from the PrEST design module is the primer pair. The primers are used for PrEST amplification by RT-PCR from a total RNA template pool in the **molecular biology module**. The amplified PrEST is sequenced for quality control, and is then cloned into expression vectors. The expression vector used is pAff8C[15] and the PrEST fragment is fused to a histidine tag for efficient purification of the resulting PrEST protein and to an albumine binding protein (ABP) for induction of immune response.

In the **protein factory module**, the recombinant PrESTs are expressed in *Escherichia coli* and purified by metal affinity chromatography, enabled by the histidine tag. After quality control with mass spectrometry, the purified proteins are used for both antigen preparation and for production of the PrEST-ligand affinity columns used for antibody purification.

The purified PrEST antigens are sent to animal farms for immunization in rabbits to generate polyclonal sera. [15] The retrieved antisera are purified in the **immunotechnology module**, by a two-step immunoaffinity based protocol on the ÄKTAexpress chromatography system. This allows for high-throughput generation of monospecific antibodies.[16] The **array-technology module** determines the quality and binding specificity of the purified antibodies on PrEST-arrays; a protein microarray chip. Antibodies are also validated by western blot analysis.

In the **immunohistochemistry (IH) module**, a protocol is established for immunostaining of the antibody. Results from PrEST arrays and western blots are considered as well as information available from public gene and protein databases and literature. In addition to the internally produced HPA antibodies, commercial antibodies are also analyzed and evaluated. Tissue microarrays are produced in the **tissue microarray (TMA) module**. The tissues are biobank material and the TMA:s include 48 normal tissues, 20 cancer types and 66 cell lines. TMA sections are stained in the IH-module with the established staining protocol, and the stained TMA slides are scanned to generate digital images in an automated scanning procedure. The result of an antibody

binding to its corresponding antigen is a brown-black staining, whereas the cells and extracellular material are stained blue. All images are processed and stored in a database to be annotated by pathologists in an annotation software. The results for each antibody are analyzed in an approval stage before they become released for the public web.

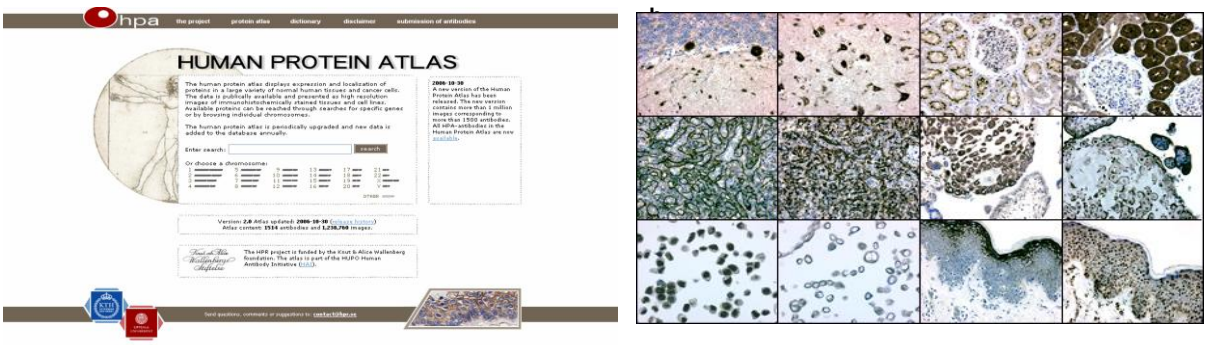

Figure 3. a) The Protein Atlas webpage (http://www.proteinatlas.org) **b)** examples of TMA images.

An essential component of the HPA program is the **informatics and LIMS-module**. This group delivers custom made software solutions for each of the modules in the pipeline and stores all production data in the HPR-LIMS (Laboratory Information Management System) database. They are also responsible for the development and maintenance of the public database, the Human Protein Atlas (Figure 3a)[16], which displays localization and expression patterns of proteins in various human tissues and cells. The Protein Atlas database is today the only open access database that contains information about the localization of proteins in a wide range of human tissues.[13] The database was first released in August 2005 and recently a new updated version with 1514 antibodies was released. Open access of the database allows everyone to retrieve data about specific proteins and download the annotated images which display the localization and expression patterns, as displayed in Figure 3b.[13] Currently there are 1238760 images in the protein atlas and new data is added to the database annually.

3. Membrane protein biology

'Membrane protein' is a term widely used for a protein that is either inserted into the membrane or have regions permanently attached to the membrane. The first type is often referred to as 'integral membrane protein' and spans the lipid bilayer of the membrane. The second type is called 'peripheral membrane protein' or 'membrane-associated protein' and is indirectly attached to the membrane by binding to an integral membrane protein or by interactions with the lipids in the membrane (lipid-linked proteins).

Integral membrane proteins can be further divided into two big subgroups: helical bundle proteins, where the transmembrane (TM) region consists of α-helices, and β-barrel proteins, where multiple transmembrane strands are arranged as β-sheets. Membrane proteins with β-barrel structures have been found in the outer membranes of bacteria, mitochondria and chloroplasts.[24] The focus of this work is on human membrane proteins, in particular proteins in the plasma membrane, and the prediction methods to be analyzed in this project are only applicable on helical membrane proteins. Moreover, the membrane protein prediction methods are trained to find membrane-spanning regions, so they cannot be used to find peripheral membrane proteins. From now on when the term membrane protein is used, this refers to integral α-helical membrane proteins.

3.1 Groups of membrane proteins

The two major types of membrane proteins in a cell are receptors and transporters. Membrane proteins that have a large region on the outside of the plasma membrane are often involved in interactions and signaling between cells, whereas proteins with regions in the cytoplasm can be involved in intracellular signaling pathways and anchoring of proteins. Other proteins with domains buried within the membrane can form channels to allow molecules to be transported across the membrane. Some examples of families of membrane proteins are ion channels, motor proteins, G protein-coupled receptors (GPCRs) and bioenergetically-related proteins that transport electrons.[17]

Many of the drug-targets of the pharmaceutical industry are membrane-bound receptors. An analysis of the pharmaceutical industry showed that cell membrane receptors account for at least 45% of the drug targets and

constitute the largest subgroup.[18] The GPCRs form the largest known membrane protein family, and it includes receptors for visual sense (rhodopsin), sense of smell (olfactory) , hormones, neutrotransmitters (serotonin, dopamine) and regulation of the immune system (chemokine and histamine receptors).[17, 25] Since GPCRs are involved in various normal biological processes, they are consequently involved in many pathological conditions. Due to their ability to present novel targets, a large fraction of prescription drugs act on GPCRs.[26]

3.2 What defines a membrane protein?

The α-helices in membrane proteins are either single-spanning or multi-spanning. The reason for the helical structure is that the hydrogen bonding between peptide bonds is maximized if the polypeptide chain forms a regular α-helix as it crosses the membrane. Since water is absent in the membrane and the peptide bonds are polar, all peptide bonds are driven to form hydrogen bonds with one another. The α-helix is a challenging structure since the hydrophobic core is insoluble in aqueous phase, such as the cytoplasm inside the cell, and therefore has a strong tendency to aggregate.[7]

In general, an α-helix has an amino acid composition with a hydrophobic center and short border regions followed by polar caps. The hydrophobic core is rich in aliphatic residues such as Glycine (Gly), Alanine (Ala), and Leucine (Leu). The border regions are often enriched in the aromatic residues Tryptophan (Trp) and Tyrosine (Tyr). The polar caps contain helix-capping residues such as Asparagine (Asn) and Gly.[27] Helix-capping is defined as motifs found at the ends of helices, containing specific patterns of hydrogen bonding and hydrophobic interactions. Since the first four N—H groups and the last four C=O groups in an α-helix lack intra-helical hydrogen bonds, they need to be capped by alternative hydrogen bond partners.[28] Most of the loops that connect α-helices in membrane proteins are short. However, sometimes large globular domains can be found between two consecutive helices.[27]

During the last years, the number of high-resolution structures of membrane proteins has increased and today the Protein Data Bank (PDB, *http://www.pdb.org*) contains more than 100 high-resolution structures of α-helical membrane proteins.[29] The information obtained from the new structures has changed the view of the complexity of integral membrane proteins, and it is now clear that not all helices are long, hydrophobic and oriented perpendicularly to the membrane. For example, an α-helix can form a re-entrant loop by spanning a part of the membrane and then return, the variation in size can be much bigger than the previously expected 20-30 residues, it can be kinked in the middle and even lie flat on the surface.[29]

3.3 Translocation and insertion of membrane proteins

In this section, the mechanisms that exist to translocate a protein into the cell membrane are explored. There are other pathways by which membrane proteins are inserted into membranes, such as the nuclear or mitochondrial membrane, but these mechanisms are less understood and not the focus of this project. The membrane proteins discussed here partly follow the secretory pathway used by proteins that are secreted by the cell.

3.3.1 The secretory pathway and protein translocation

To enter the secretory pathway, a protein needs to have some sort of signal sequence that guides the protein to the endoplasmic reticulum (ER). Proteins with ER signal sequences are either secretory proteins that will leave the cell, or membrane proteins that will be inserted in membranes such as the ER membrane, Golgi membrane or the plasma membrane of the cell. A hydrophobic signal sequence emerging from a translating ribosome can be recognized by a GTPase called the signal recognition particle (SRP). Next, the whole ribosome-peptide chain complex is targeted to the ER membrane where it binds to the SRP receptor. The bound complex interacts with another protein complex named the translocon, where translocation of the sequence can be

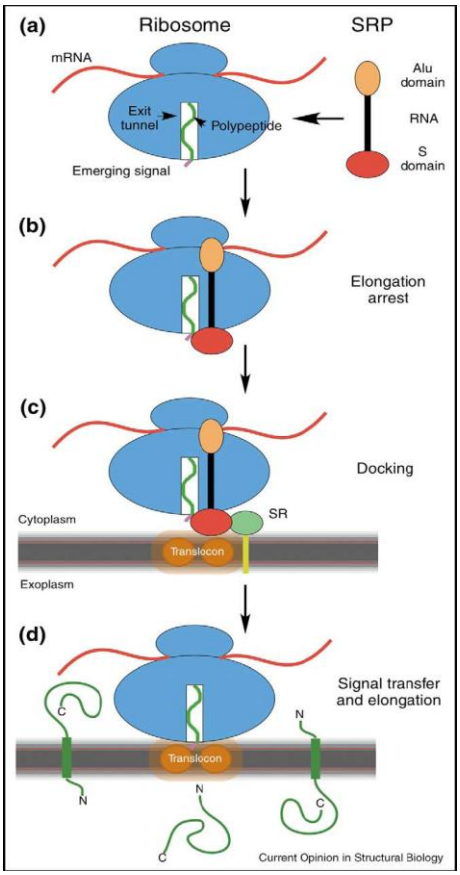

Figure 5. Membrane protein assembly by the ribosome-translocon complex. **a)** SRP can bind to a translating ribosome with an ER signal sequence. **b)** After binding of the SRP and ribosome, elongation is interrupted. **c)** The ribosome-SRP complex binds to the SRP receptor (SR, shown in green), and associates with the translocon (shown in orange). **d)** Secretory proteins are transported via the translocon into the ER lumen, whereas membrane proteins are transferred to the membrane bilayer. Reprinted from [4] with permission from Elsevier.

initiated.[30] This is illustrated in Figure 5. The mammalian ER translocon Sec61 is one of the most studied translocons.

3.3.2 The translocon

The translocon is a membrane-embedded protein complex in the ER, used both by secretory proteins to enter the secretory pathway and by membrane proteins. The translocon functions as a switching station by receiving a peptide sequence from a translating ribosome and then directing the sequence either into the membrane bilayer or across the membrane into the ER lumen.[7] Transmembrane helices are presumed to be pushed into the lipid bilayer and hence not fully translocated all the way across the membrane.[29]

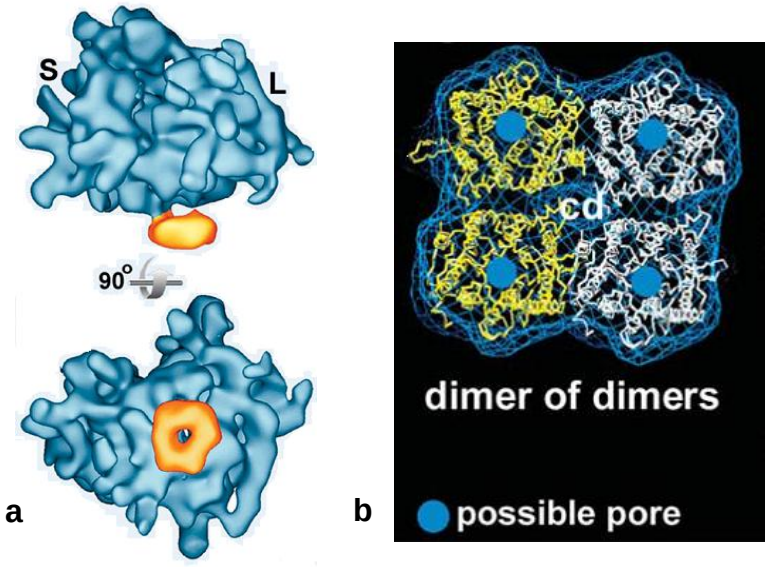

Figure 6. a) Cryo-EM images of front and bottom view of the canine ribosome–translocon (Sec61) complex without TRAP. The translocon (embedded in the membrane) is shown in yellow and the ribosome in blue. **b)** The tetrameric structure of SecY as a dimer of dimers, and the possible pores used to translocate a polypeptide (marked in blue). Modified from [6] with permission from Elsevier.

The translocon complex consists of heterotrimeric proteins that are called Sec61 in eukaryotes and SecY in bacteria. Since this work focuses on human membrane proteins, only Sec61 will be considered here, but the two mechanisms are very similar. Sec61 is composed of three subunits named α, β and γ. Recent research with cryo-EM images of the ribosome-translocon complex[6] suggest that every assembly is composed of four Sec61 copies and two copies of a complex named translocon-associated protein complex, or TRAP.[7] Figure 6a displays the ribosome-translocon complex from two different angles.

Each of the four Sec61 heterotrimers has a nascent pore believed to be the passageway for the proteins emerging from the ribosome. Figure 6b shows an example of the SecY as a dimer of dimers and the four possible pores. The channel is hour-glass shaped and has a central ring of hydrophobic amino acids that may form a seal around the peptide chains coming out from the translating ribosome.[4] Different models for the translocation have been proposed, but here only the most common will be discussed. The connections between the ribosome and the translocon complex suggest that at any particular time, only one of the Sec61 heterotrimers is used for protein export and membrane insertion. It is therefore believed to act as a monomer, and it appears like the tetrameric use of Sec61provides an assembly platform for the ribosome.[7]

So how is a protein with a transmembrane helix inserted into the membrane? Results from molecular dynamics modeling of one Sec61 heterotrimer has proposed that two of the helices, TM2b and TM7, form a gate that provides a passageway from the translocon into the lipid bilayer.[4] This gate is called the 'lateral gate' and can be viewed in Figure 7 for SecY which is very similar to Sec61. A combination of helices has also been proposed to provide a binding site for signal sequences.[31] The TM2a helix has been found to serve as a 'plug' that seals the translocon when there is no elongating polypeptide, and is visualized in Figure 7 with the translocon in a closed state.[7] The plug prevents ions from moving across the membrane, since it is necessary to keep the permeability of the membrane tightly regulated.

It has been suggested that the integration of TM helices into the membrane is a result of helices partitioning between the membrane and the translocon. Helices that are hydrophobic and have other necessary properties would prefer to leave the channel and be inserted into the lipid bilayer. More polar helices would prefer to stay in the translocon channel for subsequent transfer to the aqueous phase of the ER lumen or cytoplasm.[4]

The Sec61 channel has been further analyzed to find an explanation of how and why some proteins are inserted into the membrane whereas others pass through the pore to be secreted. There must be some sort of code in the sequence that makes it possible for the translocon to recognize TM-regions. Recent work implies that direct protein-lipid interactions are involved in the recognition of TM helices and that estimates of the free energy of membrane insertion for each amino acid located in the center of the TM segment can be used.[7] One of the questions not yet solved is how the elongating peptide sequence is captured from the ribosome and whether the peptide is able to fold in the ribosome exit tunnel. It has been proposed that polypeptides can form α-helices both inside the Sec61 channel and in the exit tunnel of the ribosome which is 100 Å long.[32]

3.3.3 Signal sequences

There are different types of signal sequences that can target a protein to the ER; cleavable signals, signal-anchors and reverse signal-anchors.[30] Cleavable signals are often referred to as signal peptides (SPs) and are present in all secreted proteins. Some membrane proteins have a combination of an N-terminal signal peptide and other signal sequences. Single-spanning membrane proteins only have one transmembrane region and hence two options for the final topology: cytoplasmic N- and exoplasmic C-terminal (N_{cyt}-C_{exo}) or opposite direction (N_{exo}-C_{cyt}). However, single spanning membrane proteins can be divided into four different classes:

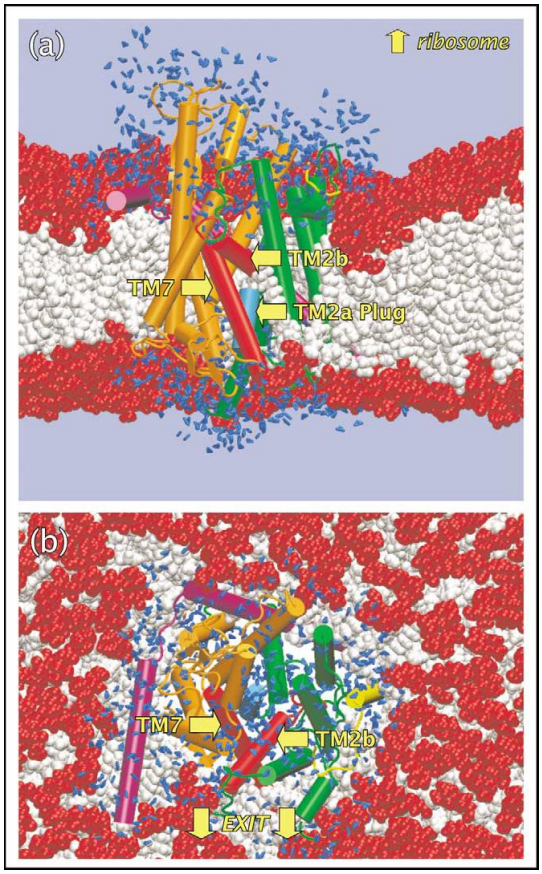

Figure 7. The structure of a single SecY in the lipid bilayer, obtained by molecular dynamics methods. Blue triangles show water molecules, acyl chains are white and phospholipid headgroups are red. **a)** The 'gate' formed by helices TM2b and TM7, and the TM2a plug helix with the translocon in its closed state, viewed along the bilayer plane. **b)** A top view looking from the ribosome indicates the presumed exit for membrane helices. Reprinted from [7] with permission from Elsevier.

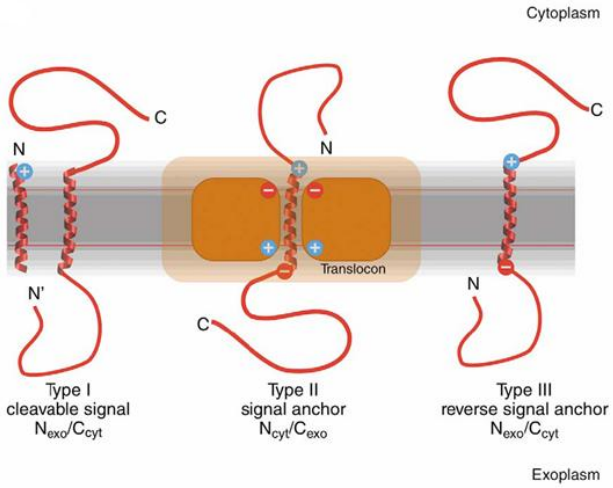

Figure 8. Three types of single-spanning membrane proteins. Reprinted from [4] with permission from Elsevier.

Type I membrane proteins (Figure 8) share the same cleavable signal sequence as secretory proteins. The signal sequence targets the protein to the ER and consists of a N-terminal hydrophobic sequence between 7-15 residues long. Type I membrane proteins also include a stop-transfer sequence typically made up of ~20 hydrophobic amino acids.[30] The stop-transfer sequence functions as a membrane-anchor which means that it has an α-helical structure and remains in the translocon, where it stops all further translocation of the sequence. This is performed by disrupting the ribosome-translocon association, and the rest of the synthesis is completed with the ribosome in the cytosol. The C-terminal of the protein sequence is never translocated.[30] The orientation of this type of protein is N_{exo}-C_{cyt} (exoplasmic N-terminal and cytoplasmic C-terminal).

Type II membrane proteins (Figure 8) have a signal-anchor sequence that functions both as a target to the translocon and for anchoring in the membrane.[30] They lack a cleavable signal sequence and the signal-anchor sequence is positioned internally within the protein sequence. Type II membrane proteins enter the translocon with N_{exo}-C_{cyt} orientation and then invert during translocation according to the positive-inside rule that will be described in section 3.4. The final orientation is N_{cyt}-C_{exo}.

Type III membrane proteins (Figure 8) contain reverse signal-anchors and translocate their N-terminal end across the membrane. A reverse signal-anchor functions as a stop-sequence by preventing further extrusion of the sequence, and it also functions as a membrane-anchor after synthesis is complete.[30] Since the N-terminal is fully synthesized before the signal-sequence is translated and ready to target the sequence to the translocon, the N-terminal sequence can start to fold in the cytoplasm. A folded sequence will have trouble entering the translocon and therefore type III proteins only have short N-terminal domains. The orientation is N_{exo}-$C_{cyt.}$

The fourth class is sometimes referred to as **tail-anchored proteins** as they are anchored to the membrane by a C-terminal sequence. The main part of the protein is consequently exposed to the cytosol. The insertion of these proteins is post-translational as opposed to the first three groups, since the signal sequence only emerges from the ribosome when it reaches the stop codon and translation already is finished. It is debated whether these proteins require assistance for membrane integration, but it is clear that insertion is independent of SRP and the translocon. [30]

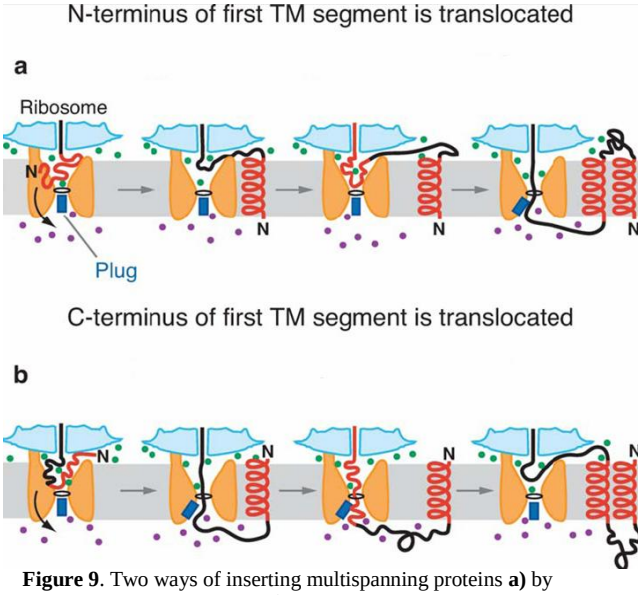

Figure 9. Two ways of inserting multispanning proteins **a)** by translocating the N-terminus first **b)** by translocating the C-terminus first. Reprinted, with permission, from the Annual Review of Cell and Developmental Biology, Volume 21© 2005 by Annual Reviews www.annualreviews.org [5].

Multi-spanning proteins span the membrane multiple times and therefore contain several hydrophobic α-helix regions. The first TM segment is believed to be responsible for targeting the protein to the ER and initiate translocation,[30] and is the most critical since the subsequent TM segments often are less hydrophobic. Each membrane-spanning α-helix acts as a topogenic sequence but SRP and the SRP-receptor only participate in the insertion of the first segment. There are at least two different insertion models for multi-spanning proteins. The simplest model, the linear insertion model, [30] proposes that the helices are inserted subsequently, so that odd numbered helices function as signal-anchor sequences and even-numbered helices as stop-transfer membrane-anchor sequences. However, other studies show that internal transmembrane segments also follow a charge rule and hence multi-spanning proteins may contain topogenic information all through their sequence.[29, 30] Figure 9 shows two ways of inserting multi-spanning proteins in a sequential procedure.

3.4 Topology of membrane proteins

The topology of an integral membrane protein describes both the overall orientation of the protein in a membrane and the number and positions of the transmembrane helices in the sequence. In most cases, the topology of a membrane protein is determined during insertion into the membrane.[29] The topology of a membrane protein in general follows the 'positive-inside rule', established by von Heijne in 1986.[33] The positive-inside rule for the topology of a membrane protein is that the flanking segment with a greater positive charge generally is on the cytoplasmic side of the membrane. There is also an opposite correlation, although weaker, for acidic amino acids.[30]

The topology of membrane proteins can generally be said to be dependent of the distribution of amino acids throughout the sequence and particularly in the TM segment. The hydrophobic residues are more abundant in the core of the helix, while the aromatic residues are common in the lipid-water interface regions. Polar and charged residues are rare in the interior of the membrane.[29] The aromatic residues Trp and Tyr, besides having a preference for the ends of helices, also affect the orientation of the helix. Trp promote a C_{cyt} orientation when placed in any of the ends, whereas Tyr has the same effect only when placed in the C-terminal end of the helix.[7] Positively and negatively charged residues have different consequences on a membrane helix, explained by the so called 'snorkel' effect. The positively charged residues Arg and Lys have very long side-chains and can therefore reach up to allow the charged end to reside in the less hydrophobic region of the lipid headgroup.[34] Among other sequence motifs found is the GxxxG motif, which enables close packing of helices.[7]

There are other factors affecting the topology of multi-spanning proteins, such as rapid folding of globular N-terminal domains, N-linked glycosylation of loops exposed to the ER lumen during the assembly of the protein, and the length of N-terminal signal anchors (longer segments have been shown to favor N_{exo}-C_{cyt} orientation).[29] A search for homologous proteins in datasets with membrane proteins in *E. coli* and *S. cerevisiae* [29] resulted in a few interesting cases of homologous proteins with opposite C-terminals. This can either be a result of the addition of an extra helix in one of the homologs, or of two proteins being oriented in opposite ways in the membrane. There have also been findings of 'dual-topology' proteins, which can insert in two opposite directions. [29]

A typical genome is predicted to contain 20-30% membrane proteins.[27, 35] Topologies where both the N-terminus and C-terminus of the protein are in the cytoplasm are the most abundant. These proteins have an even number of TM segments, which indicate a preference for inserting pairs of TM helices during the assembly, a so called helical-hairpin.[27] All measures performed so far, however, are based on membrane protein prediction methods which ignore the complicated newly discovered issues with breaks in helices, re-entrant loops and helices that lie flat on the surface of the membrane.[29]

4. Prediction methods for membrane protein topology

In this section, prediction methods for membrane protein topology will be discussed. The prediction of the full topology of a protein can be defined as the combined prediction of the total number of TM regions and their orientation in or out relative to the membrane.[35] Because of the difficulties in using methods such as crystallography and NMR spectroscopy on membrane proteins, there are few high-resolution structures available.[17] To be able to obtain more information about the structures and understand their functions, it is necessary to develop new and improve existing bioinformatical prediction methods for membrane protein topology.

Some concepts of membrane proteins have been used, with modifications, by all prediction methods (1) TM helices are between 12-35 residues long.[17] (2) Globular loops are usually shorter than 60 residues if they are placed in between two membrane helices.[36] (3) Globular loops longer than 60 residues have different composition than the shorter ones when it comes to the positive-inside rule.[17] (4) The positively charged amino acids Arg and Lys have a particular distribution within the TM protein (the positive-inside rule) and this provides important information for the topology prediction.[17, 33] Although it is now clear that many membrane proteins do not fulfill all of these concepts[29], most prediction methods were developed before this complexity was discovered.

Identifying a well characterized TM, a stretch of hydrophobic residues with a distinct length, could seem like an easy task. However, it often gets complicated. For instance, other types of regions, such as globular proteins and signal peptides, also contain long hydrophobic parts. When it comes to multispanning TM proteins, some helices may be shielded by the other TM helices and thus not entirely exposed to the lipid bilayer.[37] These helices sometimes contain hydrophilic residues, which give the helix amphiphatic properties.

All methods can generally be divided into two different classes of predictors. One class focuses primarily on the propensity of each amino acid to be in a certain region to get a residue-based evaluation of the protein sequence. Examples of methods belonging to this class are TopPred, PHDhtm, Thumbup and Split. The second class of predictors is the knowledge-based methods that use a membrane protein model to align the sequence to, such as MEMSAT and all HMM-predictors.[38] A timeline for the publishing year of different methods, assessed in this project, is given in Figure 10.

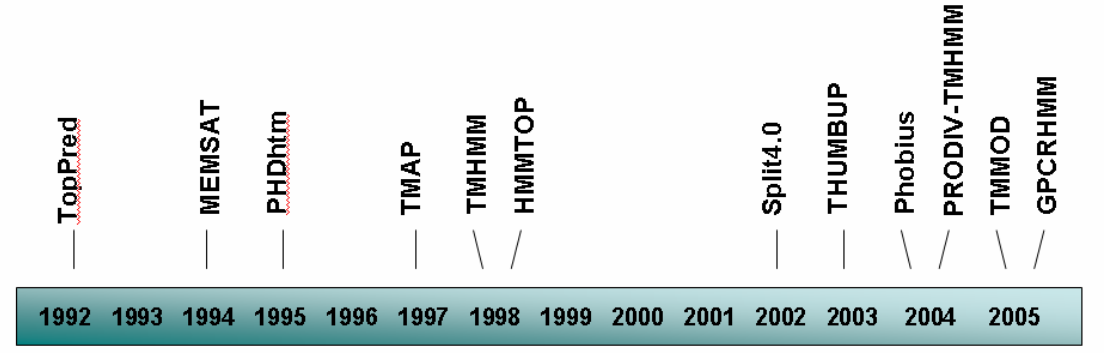

Figure 10. Timeline showing the year of publishing for a set of assessed prediction methods.

4.1 First generation's prediction methods

The first simple criteria to predict membrane-spanning helices were based on hydrophobicity scales, where distinctive patterns of hydrophobic and polar region within the protein sequence are used.[17] The first hydrophobicity scale, Kyte and Doolittle, was introduced more than 20 years ago[39] and associated a hydropathy value to each amino acid. Other scales, such as Eisenberg[40], followed and could be used to identify membrane regions. One of the drawbacks of the first hydropathy-based methods was that they often failed to discriminate between globular segments that were hydrophobic and hydrophobic membrane regions.[17]

Gunnar von Heijne introduced the positive-inside rule in 1986[33], and combining this property with hydrophobicity improved the predictions. The predictor TopPred was published in 1992[41] and implemented a more complex processing of the hydrophobicity scales. TopPred uses hydrophobicity analysis with a sliding trapezoid window and automatic generation of possible topologies for the protein to predict the complete topology by ranking the possible topologies by the positive-inside rule.[17]

In 1994, the model-based method MEMSAT[42] combined statistical tables with log likelihood values and a dynamic programming algorithm to predict membrane protein topology. The model used in MEMSAT is based on expectation maximization and five states with separate propensity scales for the residues. A constrained dynamic programming algorithm finds the optimal score and the best prediction.

In 1995, PHDhtm was one of the first methods to use information from alignments with protein families to improve the prediction accuracy.[43, 44] Topology and location of the TM regions are predicted using a system of neural networks and a second post-processing step to maximize the positive charge on the cytoplasmic side. To process the output from the neural network, a dynamic-programming algorithm similar to the one in MEMSAT is used.[44] This combination of information from algorithms and multiple alignments makes PHDhtm one of the most accurate prediction methods.[17]

4.2 Hidden Markov models

In a hidden Markov model (HMM), a series of observations are described by a stochastic hidden Markov process.[45] A first order Markov chain consists of a sequence of random values that has the Markov property, absence of memory so that the probability at a certain time *t* only depends on the value of the previous time step *t-1*, and a finite number of states. In an HMM, the current state is not observable and hence is called 'hidden' and only observable as a probabilistic function of the state.[46] In each state a symbol is emitted, and the model is based on transition probabilities and emission probabilities that have to be properly determined in order to have a good model.[38] Emission probabilities are the probabilities of emitting a certain symbol in a certain state of the model. Transition probabilities are the conditional probabilities of moving to a new state given the current.

Hidden Markov models have been used for a long time in computational biology.[2] The aim when using HMMs is to build a model that resembles the biological system being modelled as closely as possible. The states in an HMM for membrane protein prediction are connected to each other in a way that is reasonable in a biological way. For instance, a loop state is connected to itself to allow the loop to be longer than 1, and it is also connected to a helix state.[2] Each transition, i.e. to move from one state to another, is associated with a transition probability.

The membrane proteins can be said to have a "grammar" in their structure that constrains the possible topologies, and this can be incorporated into a model for prediction. A loop has to be followed by a helix, and cytoplasmic/non-cytoplasmic loops have to alternate. If a model such as an HMM uses this kind of information, better predictions can be obtained.[35] Another advantage of HMMs is that it is possible to set upper and lower limits for the length of the TM regions. An HMM for transmembrane protein prediction can include helix length, hydrophobicity, charge bias (positive-inside rule) and grammatical constraints in one single model.[35]

4.2.1 TMHMM

TMHMM (Transmembrane HMM) was the pioneer predictor using a Hidden Markov model to predict membrane protein topology. TMHMM was published in 1998 by Sonnhammer *et al.*[2] The layout of the HMM is cyclic and consists of seven types of states: globular domain, cytoplasmic loop, cytoplasmic/non-cytoplasmic helix cap, helix core, short and long loop on non-cytoplasmic side (Figure 11). The short loops are up to 20 residues long. Each sub-model contains several HMM states that models the length of the specific region.[35] The cap sub-model contains the five first or last residues of the helix. The helix core has five to 25 states, which means that the possible total length of the helix is 15-35 residues including caps.

Each state has a probability distribution over the 20 amino acids, estimated from known membrane proteins, which is supposed to characterise the variability of the residues in the modelled region.[2] For TMHMM, the probabilities for the HMM parameters were estimated using a set of 160 proteins with known locations of transmembrane helices, 108 of which were multi-spanning and 52 of which were single-spanning.[35] All emission probabilities of the same type of state are estimated collectively.[2] The prediction is performed by finding the most probably topology according to the results of the HMM. The output is a labelled sequence of three classes: i for inside or cytoplasmic, o for outside or non-cytoplasmic and h for helix.

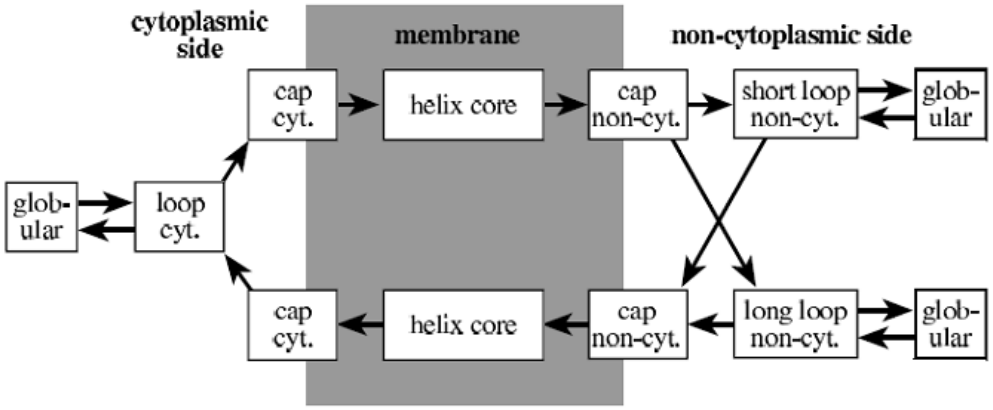

Figure 11. The overall layout of TMHMM. Reproduced from [2] by permission from the American Association for Artificial Intelligence ©1998.

4.2.2 HMMTOP

HMMTOP (Hidden Markov Model for Topology Prediction) was developed independently of TMHMM and published in 1998.[47] HMMTOP is built on a similar structure as TMHMM but uses another method for structure prediction. Both methods are reported to have similar prediction accuracy[48], however HMMTOP often confuses signal peptides with TM regions. The model is based on the principle that the maximum divergence of amino acid composition of sequence regions, and thus the differences between the amino acid distributions in the structural parts of the protein, determines the topology of TM proteins.[49] The HMM has been developed to find the topology that corresponds to the maximum likelihood for all possible topologies given the query sequence.

The HMMTOP model has five structural states: outside loop, outside helix tail, membrane helix, inside loop and inside helix tail. Two joined tails can form a short loop directly connected to the membrane, or they can be followed by a loop.[47] Three steps are used to obtain a prediction. First, the HMM parameters such as initial state and state transition probabilities are set, either by random or predetermined values. Next, these parameters are optimized for the given sequence. The last step is to use an algorithm to find the best path of states given the parameters and the model.[47]

4.2.3 Phobius

One of the main problems in the prediction of membrane protein topology is that TM regions often are confused with signal peptides. Phobius, a combined signal peptide and transmembrane protein topology predictor, was published in 2004.[37] Hydrophobic regions of TM helices have a high similarity to a hydrophobic signal peptide but there are ways to discriminate between them. The Phobius HMM models the sequence regions of a signal peptide and a membrane protein with states that are interconnected.[37] It can be looked upon as a combination of the model used in SignalP-HMM, a widely used predictor for SPs [50], and the model used in TMHMM with modifications.

It is estimated that 16% to 20% of the human proteins contain signal peptides. The structure that Phobius uses to model SPs has three distinct regions. The n-region is slightly positively charged and consists of 1-12 residues near the N-terminal. The h-region is a hydrophobic α-helical region that is usually shorter than TM helices (7-15 residues). The c-region is rather polar and uncharged and consists of three to eight amino acids, positioned between the h-region and the cleavage site.[37, 50]

If a signal peptide can be successfully predicted, this gives valuable topology information since it states that the N-terminus of the protein is on the cytoplasmic side of the membrane. Hence, the orientation of the protein is given by the prediction of a signal peptide.[37] Phobius has been observed to be more sensitive but less specific than TMHMM, which means that it has a higher false positive rate but lower false negative rate.[37] Compared to SignalP[50], it is more conservative, i.e. it has a higher rate of false negatives but lower false positive rate.

4.2.4 GPCRHMM

GPCRs constitute a large superfamily and are involved in various important signal transduction pathways. All GPCRs span the plasma membrane seven times and have a N_{exo}-C_{cyt} orientation, but are still so diverse that there is a lack of common sequence motifs within the superfamily. However, when analyzed more closely[25], certain common features can be found, such as differences in the amino acid composition between membrane regions, extracellular- and cytoplasmic loops, and distinct patterns in loop length. These features were incorporated into an HMM that was trained on a dataset that represented the GPCR superfamily. GPCRHMM was published in 2005 [25] and is based on the TM topology features mentioned to specifically recognize GPCRs. It is therefore not a general TM predictor and always predicts seven helices in the proteins predicted by GPCRHMM as GPCRs. The model was reported to have a sensitivity of about 15% higher than the best TM predictors on GPCRs.[25]

4.2.5 PRODIV-TMHMM

PRODIV-TMHMM, published in 2004[38], is a profile-based hidden Markov Model, which means that it uses sequence profiles based on evolutionary information in the form of multiple sequence alignments. The sequence profiles is combined with an HMM that is proposed to include the best features from the models of TMHMM and HMMTOP.[38] The multiple sequence alignments are based the query sequence and its homologs and the model differs from standard HMMs in how emission probabilities are calculated. The profiles can be use both for estimating the model parameters and for predictions. PRODIV-TMHMM is not optimal for distinguishing between membrane proteins and non-membrane proteins. When the method was run on a set of 1087 globular proteins without membrane regions, 79% were predicted to contain at least one TM segment.[38]

4.3 Methods based on amino acid property

There are newer methods based on other approaches than HMMs, for example methods that evaluate the protein sequences using algorithms based on amino acid properties.

4.3.1 THUMBUP

THUMBUP is an abbreviation for 'the topology predictor of transmembrane helical proteins using mean burial propensity'. The method is based on a simple scale of burial propensity and uses the fact that transmembrane helices are packed more tightly than non-membrane helices.[51] Burial propensity is the tendency of a amino acid to be buried by other residues. In THUMBUP, published in 2003, a sliding-window approach is used for the profile of burial propensity for the residues and another algorithm is used for identifying TM segments. To determine the orientation of the segment, the positive-inside rule is applied. It is claimed [51] that a method based on physiochemical property is able to provide topology predictions as accurate as the predictors based on more advanced algorithms with more parameters, such as TMHMM and MEMSAT. For instance, THUMBUP has 24 parameters compared to more than 100 parameters in HMMTOP. When tested for its ability to discriminate between TM proteins and soluble proteins, it was observed that THUMBUP had a higher rate of false positives but no false negatives.[51]

4.3.2 Split 4.0

Split 4.0 was published in 2002[52] and uses basic charge clusters for topology predictions. Basic charge clusters are clusters of the basic residues (Arg and Lys) that are predominantly found in cytoplasmic loops and therefore can be applied as topology determinants. Some common motifs are BB, BXB, BBB, BBXXB, BXXBB, BXBXB where B is a basic residue and X is any other residue.[52] The frequencies of common charge motifs were calculated from known proteins to find the distribution of basic amino acids among other amino acids. 15 different scales for amino acid attributes, including the Kyte-Doolittle hydropathy scale, are used to find potential TM helices. Bias in basic charge clusters is used in combination with the standard charge bias (positive-inside rule) and the charge difference across the first TM segment for determination of topology.[52]

4.4 Accuracy of prediction methods

One of the major problems when it comes to estimating the accuracy of membrane protein topology prediction methods is the lack of experimentally validated transmembrane annotations available. Less than 1% of all available protein 3D structures are membrane proteins.[20] Since all methods have been developed and trained using basically the same small set of known membrane proteins, accuracy is hard to estimate. [17]

A consequence of the limited amount of high-resolution experimental data is that low-resolution experimental data has been included in training and testing set.[17] A typical training set may consist of ~200 protein structures.[19] The test sets used for training and evaluation of prediction methods consist of datasets of well studied membrane proteins that probably are easier to predict correctly than data sets from complete genomes.[53] This has immense consequences on the result of the prediction methods and therefore the expectations on the accuracy must be lowered when genomic data is analyzed. [53] It is believed that when analyzing entire proteomes, a 55% to 60% overall topology prediction accuracy is possible with the methods available today.[20]

Most prediction methods use a constraint that transmembrane regions generally span between 17 to 25 residues and that loops between helices often are longer than 15 residues. However, it has been found that many loops are in fact shorter than ten residues and therefore are difficult to detect for the methods, and that half of the helices do not fall into the expected interval. Many membrane helices are actually longer than 32 residues.[54] The structure of membrane proteins also shows a higher diversity in eukaryotes than bacteria.[53] Also, membrane proteins do not seem to be entirely conserved across species and thus methods based on evolutionary information do not perform as well as expected.[17]

Another difficulty in analyzing the methods is that levels of prediction accuracy, as evaluated in comparative studies and in the corresponding publications for each method, can not be compared to one another. [17] The reason is that such comparisons are based on different measures for prediction accuracy and that they use different data sets. Using a data set that a method was trained on to test and validate the same method will automatically and incorrectly give great accuracy results. [48] It has also been realized that the test sets consisting of the available proteins with known topologies are biased and not representative to the set of membrane proteins in a complete genome.

In an evaluation by Chen *et al* from 2002[48], no prediction method was able to distinguish itself as remarkably better than the others in all tests performed. However, the best hydrophobicity-scale based methods were significantly less accurate than the best advanced methods. Most methods confused membrane helices and signal peptides and the advanced methods had a tendency to underpredict helices. The hydrophobicity-based methods, although able to identify many membrane-spanning helices, also predict membrane regions in a number of globular proteins. [17]

In a study in 2001 by Möller *et al*[19], TMHMM was the overall best performing method; especially at distinguishing between transmembrane and soluble proteins but with a tendency to underpredict helices. They also state that topology predictions should be performed in combination with signal peptide prediction methods. Another study by Melén *et al* in 2003[20] also ranked TMHMM and as the best performing prediction method together with MEMSAT. Most evaluation studies were performed before 2002, and hence the newer methods such as Phobius, TMMOD and THUMBUP have not been included in these analyses. However, a comparative evaluation performed by Cuthbertson et al in 2004[3], found that Split4.0, HMMTOP and TMHMM were among the methods that consistently performed well.

Table 1 shows results from the evaluation of 13 methods as described by Cuthberson[3]. The two datasets were a redundant dataset containing 434 TM helices and 112 proteins, and a non-redundant dataset with 268 TM helices and 73 proteins, the second obtained by removing proteins with sequence identitity ≥ 30 % to another protein in the dataset. The non-redundant dataset was created since redundancy in datasets has been proposed to bias accuracy estimations.[48] No single-spanning proteins were included in the redundant dataset. Although not all methods assessed in this master thesis are in these tables, it serves as an example of a comparison table between prediction methods.

a

Prediction method	N_P	N_C	Q_P (%)	Q_{3TM} (%)	Q_{3NTM} (%)	Q_3 (%)	N_{TM}
ALOM2	303	294	81.1	43.5	97.3	73.9	49
DAS	486	419	91.2	62.6	94.6	80.7	60
HMMTOP2	435	409	94.1	71.9	91.1	82.7	85
MEMSAT 1.5	423	402	93.8	70.7	92.3	82.9	83
MEMSAT2	413	388	91.6	68.5	92.5	82.1	62
MPEX	421	391	91.5	68.6	87.1	79.1	75
PHD	415	388	91.4	63.1	92.8	79.9	59
SPLIT4	410	398	94.4	78.3	89.7	84.7	88
TMAP	411	393	93.1	77.4	85.3	81.9	73
TM-FINDER	384	367	89.9	68.7	91.0	81.3	62
TMHMM2	405	389	92.8	72.3	91.8	83.3	71
TMPRED	423	400	93.4	69.0	92.0	82.0	79
TOPPRED2	447	409	92.9	71.6	89.1	81.5	74

[a] N_P is the number of predicted TM helices and N_C is the number of correctly predicted TM helices, out of a total of 434 observed TM helices. The measures of prediction accuracy are the per segment prediction power Q_P and the per residue accuracies for all residues (Q_3), for TM residues (Q_{3TM}) and for non-TM residues (Q_{3NTM}). N_{TM} is the number of chains (out of a total of 112) for which all TM segments are predicted correctly.

b

Prediction method	N_P	N_C	Q_P (%)	Q_{3TM} (%)	Q_{3NTM} (%)	Q_3 (%)	N_{TM}
ALOM2	178	175	80.1	41.8	97.8	73.8	35
DAS	297	260	92.2	61.5	95.2	80.7	43
HMMTOP2	270	254	94.4	70.4	91.1	82.2	58
MEMSAT 1.5	256	246	93.9	69.2	92.8	82.6	55
MEMSAT 2	250	237	91.6	67.2	93.2	82.1	41
MPEX	261	246	93.0	69.4	87.2	79.5	50
PHD	249	238	92.1	62.1	94.1	80.3	41
SPLIT4	254	250	95.8	78.3	90.3	85.2	61
TMAP	249	240	92.9	76.1	86.2	81.9	48
TM-FINDER	235	226	90.1	67.5	91.7	81.3	44
TMHMM2	246	239	93.1	71.1	92.5	83.3	49
TMPRED	259	247	93.8	68.5	92.1	82.0	53
TOPPRED2	273	253	93.5	70.7	89.5	81.4	52

[a] N_P is the number of predicted TM helices and N_C is the number of correctly predicted TM helices, out of a total of 268 observed TM helices. The measures of prediction accuracy are the per segment prediction power Q_P and the per residue accuracies for all residues (Q_3), for TM residues (Q_{3TM}) and for non-TM residues (Q_{3NTM}). N_{TM} is the number of chains (out of a total of 73) for which all TM segments are predicted correctly.

Table 1. The predictions of the number and location of TM helices by 13 methods in **a)** a redundant dataset **b)** a non-redundant dataset. Reprinted from [3] by permission of Oxford University Press © 2005.

5. Development of tools for antigen selection

The main subject for this master thesis project was to find the best solution for discriminating between soluble proteins and membrane proteins in the PrEST design tool used by the HPA program. Since the accuracy of membrane protein topology prediction methods in general is expected to be no better than 55% to 60 % for whole genome purposes[20], and no transmembrane topology prediction method has distinguished itself as being better than the others[48], the choice was to collect predictions from several methods.

5.1 Selection of prediction methods

The goal for the selection of prediction methods was to find reliable approaches that would be suitable for high-throughput purposes and also would complement each other. The methods that were interesting due to accuracy results were mostly the HMM-based methods TMHMM, Phobius, TMMOD[55], and HMMTOP. The ability of Phobius to discriminate between signal peptides and transmembrane regions was also desired. Split 4.0 showed good results[48, 52] and would, in combination with THUMBUP, be able to provide additional information to the HMMs due to the different underlying methodology. PHD_htm and Memsat, although older than most methods, seemed to be able to compete with the newer methods in accuracy.[48] GPCRHMM, although not a general TM prediction method, was appealing for its ability to predict the interesting group of GPCRs.

Although there are consensus prediction methods that use a number of prediction methods and a weight system to get a consensus results [56, 57], none of these includ any of the newer methods and were not suitable for the high-throughput analysis required by the HPA system. To provide information similar to a consensus prediction metod, five topology prediction methods were chosen out of all possible choices. GPCRHMM was added to the selection. The final selection was restricted by some practical issues. A number of the programs were easily accessed and available for download on web servers, although several required academic licenses. Other programs, such as TMMOD, had not yet issued stand-alone versions and therefore had to be left out from the selection. MEMSAT and PHD_htm were installed on the server, but those prediction methods included additional steps, such as using separate programs for BLAST[58] and alignments, making them inconvenient for whole-proteome analysis due to the effort and time required. TMHMM, being the program used by Ensembl and with data available in the HPA database, had already been applied for previous PrEST design of membrane proteins. It was therefore decided to keep TMHMM as one of the methods to use in the new PrEST design tool.

In summary, the final selection of membrane protein topology prediction methods was Phobius, TMHMM, HMMTOP, Split 4, Thumbup and GPCRHMM.

5.2 Comparison between selected prediction methods

A basic evaluation was performed to get an estimation of how well the methods agree with each other. A data set of 281 proteins, predicted by TMHMM to have six TM regions, was used as input to the methods. Since the programs were not implemented on the HPA servers yet, a file with the protein sequences in FASTA format was uploaded to the web server of each method. Split 4.0 unfortunately could not take multiple sequences and had to be left out of this evaluation. The result is displayed in Figure 12 and shows how the number of predicted TM regions varies between the different methods. Both Phobius and HMMTOP have more 7TM proteins predicted than 6TM. Although THUMBUP has the most proteins predicted as 6TM, it also shows the widest distribution ranging from three to more than 10 predicted TM regions. GPCRHMM predicts 76 of the 281 proteins as GPCRs, and hence 7TM proteins.

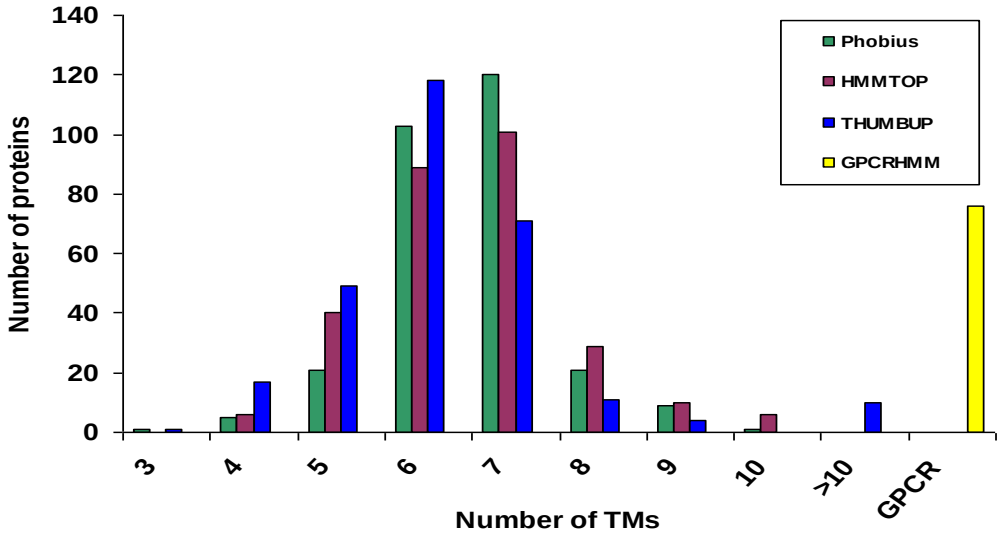

Figure 12. The results of four prediction methods for a dataset of 281 proteins predicted as 6TM by TMHMM.

5.3 Implementation of prediction methods and database design

The human proteome data to be used as input for the prediction methods was obtained from Ensembl[14], and the protein sequences were saved in FASTA format. The Ensembl version used was 41.36, which contained 48403 proteins corresponding to 23224 genes. However, 8 proteins had invalid characters in their sequences, such as * and X, and had to be removed. The final set consisted of 48395 proteins and 23218 genes.

Implementation of the membrane protein topology prediction methods had to be performed in such a way that it would be easy to update the information at each new release, approximately every two months. The programming language chosen was Perl and it was decided to run and parse the results of all methods in a single script that could be incorporated into the main Ensembl updating tool. The selected prediction methods were installed on the Linux-server used for the HPA program LIMS system, except for GPCRHMM, which turned out to only be available on a remote webserver. The techniques used to implement each of the methods will be described in more detail in the following sections. For examples of the output files generated from each program, see Appendix 1.

5.3.1 Database design

The first step was to design a database where all prediction data could be stored and easily accessed. For this purpose, three tables were added to a MySQL database schema, which contains all data needed for the PrEST design tool. Figure 13 shows the Entity Relationship (ER) diagram with the three tables and their attributes.

The **tm_prediction** table contains all information that is obtained from the run of prediction methods. The feature attribute is either ‘i’ (inside) ‘o’ (outside), ‘t’ (transmembrane) or ‘s’ (signal peptide). Each entry has a start and stop in amino acid position, and ‘ensp_id’ is the Ensembl protein id unique for each protein. ‘Type’ defines which method that was used for this prediction, and can be a number between 1 and 6. The **tm_prediction_type** table contains the number and corresponding name of the prediction methods, e.g. type_name = ‘Phobius’ has type_id=1. The reason for using id numbers instead of the full names of the

prediction method in the tm_prediction table is that it makes querying the database much faster. If the type_id of the method is known, only the tm_prediction table needs to be used. The **tm_prediction_feature** table contains the full names of the features, to be used if anyone is unsure what the abbreviated features stand for. An example of the entries in the tm_prediction table for a protein predicted to have a signal peptide, one TM region and cytoplasmic C-terminal by Phobius is:

ensp_id	aa_start	aa_stop	feature	type
ENSP00000343380	1	30	s	1
ENSP00000343380	31	53	o	1
ENSP00000343380	54	76	t	1
ENSP00000343380	77	166	i	1

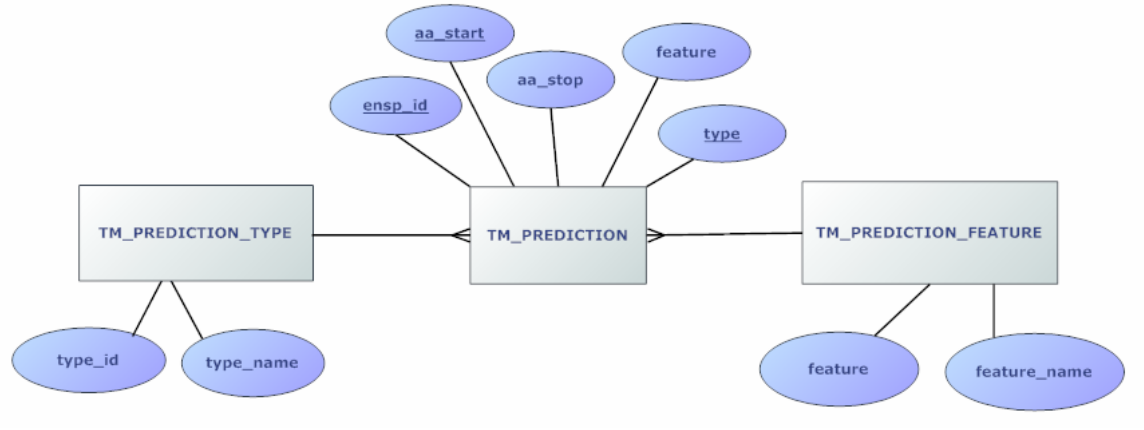

Figure 13. Entity Relationship diagram of the membrane protein information in the MySQL database.

5.3.2 Implementation of TMHMM

TMHMM is the method used by Ensembl and hence the positions of transmembrane regions were already available in our databases via the local Ensembl core database. However, Ensembl does not save topology information, so it was necessary to install a stand-alone version. A license for academic use was obtained and the program was easily installed. The version was TMHMM2.0 and it was run with the choice of short output for easy parsing. The input needed was a standard FASTA-formatted file so no reformatting was needed. The run time of the program for the 48395 proteins was about three hours. The output file was parsed and the result was written to a file *tm_prediction.txt* that could be directly loaded into the tm_prediction table and with the same format as the example in section 5.3.1.

5.3.3 Implementation of HMMTOP

A license for HMMTOP was applied for through a form on the HMMTOP webpage (*http://hmmtop.enzim.hu*). HMMTOP was run with the standard settings and the FASTA-formatted file as infile. The result of parsing the output was added to the tm_prediction file that already contained the TMHMM results. The run time of HMMTOP was about five hours.

When the result of the HMMTOP predictions were analyzed, it was found that this program predicted almost twice as many proteins as membranes as TMHMM did. After correspondence with Gabor Tusnady, the creator of HMMTOP, it became clear that HMMTOP was not developed to distinguish between membrane and non-membrane proteins, but only to predict transmembrane regions of transmembrane regions. It therefore over-predicts proteins and predicts too many proteins as membrane proteins. However, since the program already was up and running, it was decided to keep HMMTOP in the selection of methods but only show its results where at least one of the other programs also predicts a protein to have TM regions.

5.3.4 Implementation of Phobius

Phobius was selected for its ability to predict signal peptides combined with TM predictions, and for showing good results in evaluations.[37, 55] It is built on a structure very similar to TMHMM and is also run with the same settings and input format. However, parsing the output was more complicated since some proteins had signal peptides predicted, some had TM regions, and some had both a signal peptide and TM regions. With the signal

peptide feature included in the tm_prediction table this did not cause a problem. The run time of the program with the 48395 proteins as input was about six hours.

5.3.5 Implementation of THUMBUP

THUMBUP was obtained from its creators and was easy to install. However, some alterations were needed before it could be used for whole-genome purposes. The maximal length of the input sequences was 1950 amino acids, and the human proteome (Ensembl version 41.36) contains 949 proteins longer than 1950 amino acids. The longest protein sequence in this version was actually more than 24000 amino acids long. The creator, Yaoqi Zhou, was helpful and provided a new copy of the program where the maximum limit was changed to 39500 amino acids. The next alteration was to change the FASTA-formatted file format of the input. THUMBUP can only handle 70 characters of each line of the sequence in the input file. After this was done, the program was run and parsed. The run time was 24 seconds.

5.3.6 Implementation of Split 4.0

The fifth program, Split 4.0, posed a great challenge. It was written in Fortran and turned out to be difficult to install and run. First, the FASTA-formatted file used for the other programs could not be used as input to the program. Secondly, the program requires the user to actively input information about the name of the input file and the number of proteins to be used at start-up. This is not applicable when running an automated script over night. To solve this problem, a copy of the source code was obtained. Changing the Fortran code to always open the same file name instead of asking the user for it, and to set the number of proteins to 1 worked out this part.

The reason for running Split one protein at a time was that the required input format is a file with a list of names of text files, where each text file provides the name and sequence of the protein in a very specific format. Instead of creating 49395 different files in the Split directory, it was decided to run and parse the program for one protein at a time, looping through the same Fasta file that was used as input to the other programs, and use the same filename each time. After figuring out exactly how to create the protein sequence file what format to use, another problem was found. The input format could only handle proteins with 999 amino acids. After a lot of correspondence with Davor Juretic, the creator of Split 4.0, a collaboration to try to make the program more suitable for whole-genome purposes has started. First of all, alterations have to be made to the program to make sequences longer than 999 amino acids possible as input. In the mean time, Split 4.0 has been run with the 43525 proteins that are of suitable length (<1000 amino acids), and the run time was about 13.5 hours.

5.3.7 Implementation of GPCRHMM

The sixth program was chosen since it is specifically built to predict the important class of GPCRs, and thus is able to provide complementary information to the other methods. The more TM segments a protein has, the more often different prediction methods disagree.[53] A license for GPCRHMM was not yet available, so instead of running a stand-alone version on our servers, the FASTA-formatted file with the human proteome sequences was uploaded and run on the GPCRHMM webserver (*http://gpcrhmm.cgb.ki.se*). However, as soon as a separate version is accessible, it will be incorporated with the automated Perl script.

GPCRHMM had to be run and parsed in two steps. First, the GPCR detection method was run for all human proteins to find the proteins predicted to be GPCRs. The on-screen output, generated from the program at the web server, was copied to a text file and that could be parsed by a Perl script to create a list of the Ensembl protein ids predicted as GPCRs. Second, the GPCR transmembrane segment localization prediction method was run to get the positions of the seven predicted TM regions. The results from the second run was also copied to a text file, parsed and inserted into the database. Two separate perl-scripts were needed to perform the GPCRHMM analysis.

5.4 Whole-genome scan results

As mentioned earlier, the whole-genome scan was performed on 23218 protein coding human genes in Ensembl version 41.36. The reason for doing the scan on genome-level, instead of analyzing the results in number of proteins, is that for the HPA program only one protein per gene is used for PrEST design. Hence, the important information for the HPA program is the number of membrane protein coding genes according to the various prediction methods.

The result of the membrane protein topology prediction methods is presented in this section. The design of the database that stores the data from the predictions methods allows easy retrieval of the results by querying the

database using SQL (Structured Query Language). Figure 14 presents the number of genes coding for membrane proteins as predicted by the prediction methods, and the number of GPCRs predicted by GPCRHMM. The total number of membrane protein-coding genes ranges from 5634 to 10518 (Figure 14a). The result of 864 GPCRs predicted by GPCRHMM indicates that ~10%-15% of all membrane proteins are GPCRs. The high number of genes predicted as membrane protein-coding by HMMTOP is due to the fact that the method was developed to predict membrane helices only in membrane proteins, and not to discriminate non-membrane proteins from membrane proteins.

Split4.0 (Figure 14b) cannot be directly compared to the other methods since it uses a smaller dataset, but it is still obvious that the number of membrane proteins is overpredicted in the same manner as HMMTOP. After the analysis was performed, it was also found that similarly to HMMTOP, Split 4.0 was not optimized to distinguish membrane proteins from soluble proteins and ~10 % of soluble proteins are incorrectly predicted as membrane proteins.[52]

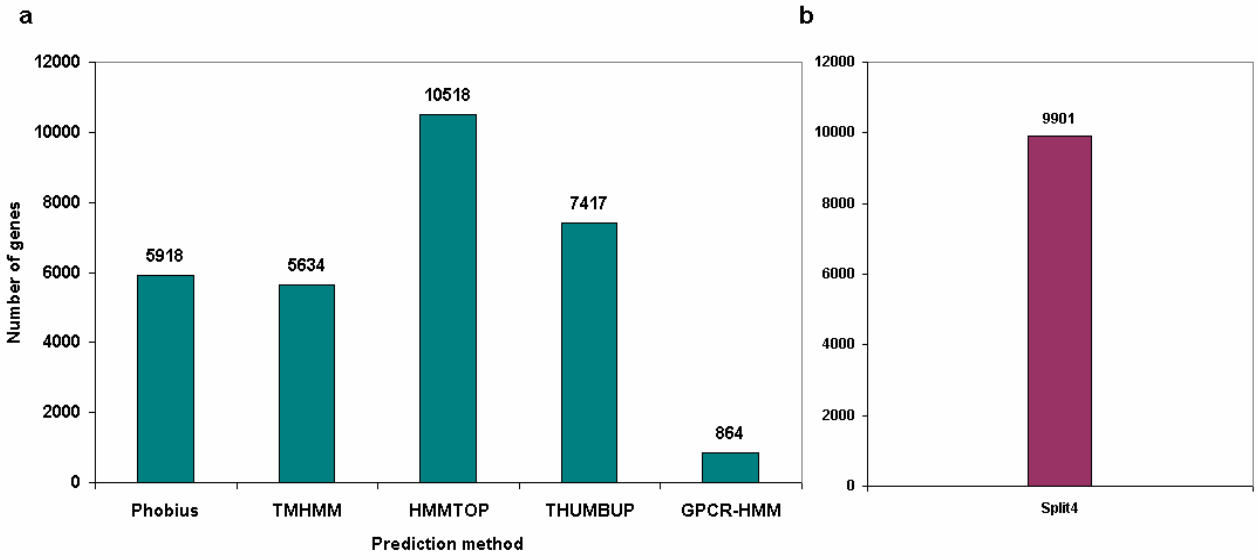

Figure 14. The number of membrane protein coding genes predicted by the prediction methods **a)** Phobius, TMHMM, HMMTOP, and THUMBUP using a set of 48395 proteins. It also shows the number of genes predicted as GPCRs by GPCRHMM **b)** Split4.0 using a smaller dataset of 43525 proteins less than 1000 residues long.

Figure 15 shows a Venn diagram with the overlap of the predictions by TMHMM, Phobius and THUMBUP. The numbers explain the overlap between all three methods (middle) and the remaining overlap between the three pairs (as indicated by arrows). The reason for comparing these is that HMMTOP and Split4.0 are not trained to discriminate soluble proteins from membrane proteins, and therefore a comparison between the other three gives a more accurate result. The total number of membrane protein-coding genes predicted by at least one of the methods is 8012.

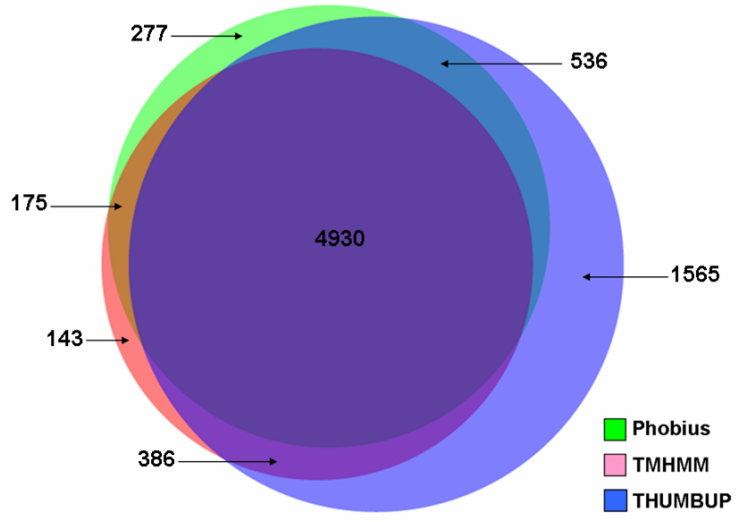

Figure 15. Venn diagram showing the overlap between the number of membrane protein coding genes predicted by Phobius, TMHMM and THUMBUP. Total number of predicted protein-coding genes by each method: Phobius 5918, TMHMM 5634, THUMBUP 7417.

5.5 Implementation and testing of PrEST design criteria and software tools

5.5.1 From ProteinWeaver to PrEST design tool

The tool currently used for PrEST design is the ProteinWeaver, as demonstrated in Figure 1. The drawback of this program is that the source code is not available and it is therefore not possible to change the program and add features that are desired for PrEST design, such as restriction sites and the results of transmembrane topology prediction methods other than TMHMM. Hence, a new PrEST design tool is under development by the information technology group and the PrEST design group of the HPA program. This tool will eventually replace ProteinWeaver in the design of PrESTs.

5.5.2 Membrane proteins in the PrEST design tool

The implementation of prediction methods as described in the previous sections allows for new features in the PrEST design tool. The most essential change for PrEST design on membrane proteins is the additions of the results from multiple topology prediction methods. The database design described in section 5.3.1 provides easy access to these results, and the membrane protein specific information has been incorporated into the design tool.

The topology information stored in the database can be displayed with the transmembrane regions as boxes similar to the visualization in the ProteinWeaver view, and different coloured lines represent inside and outside loops. Figures 16-19 show four examples of the PrEST design tool for membrane proteins. The graphics for the membrane proteins in the PrEST design tool has been implemented by the information technology group of the HPA program. Starting from the top, the features shown are PrEST position (if one of more PrESTs have been designed on the protein), identity to other human proteins in percent, predicted transmembrane segments and signal peptides, common/unique regions to transcripts of the same gene, low complexity sequence regions on the protein sequence level, Interpro domains, and Asc1/Not1 restriction sites. The features are only displayed if they are applicable to the proteins.

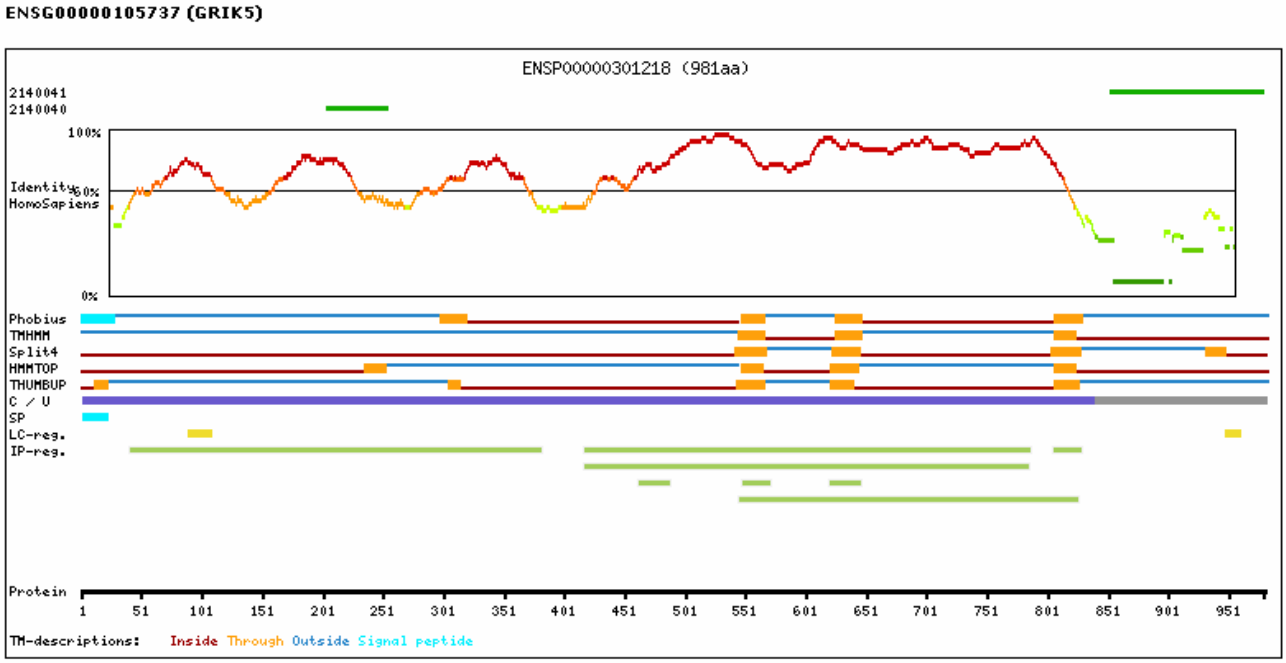

Figure 16. An example of the PrEST design tool with a protein predicted to have 3-5 TM regions and conflicting topology results. Intracellular loops are red, extracellular loops are blue and TM regions orange. Phobius and SignalP also predict a signal peptide, displayed in turqoise.

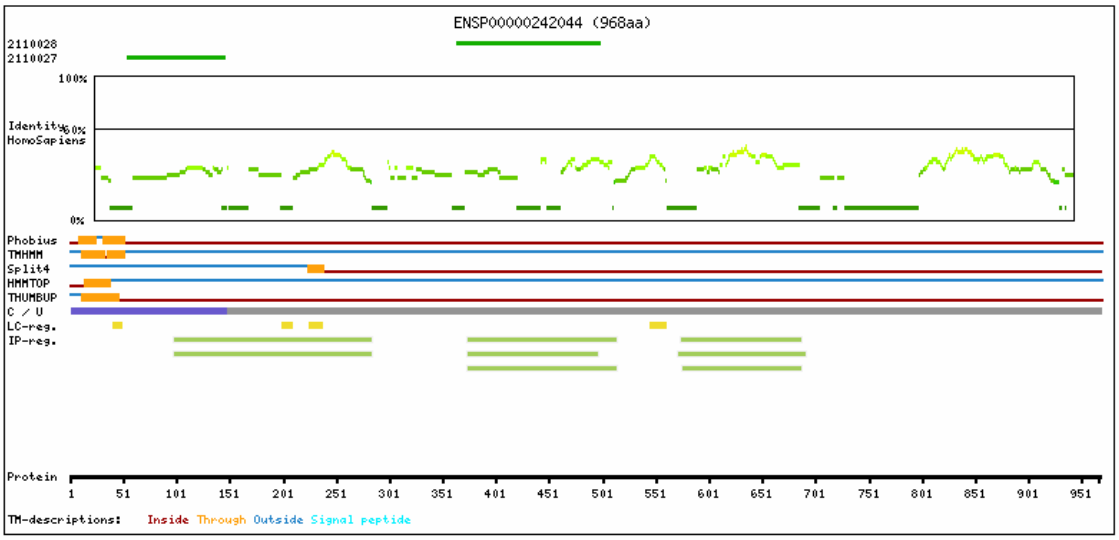

Figure 17. An example of the PrEST design tool with a protein predicted to have 1-2 TM regions and conflicting topology results.

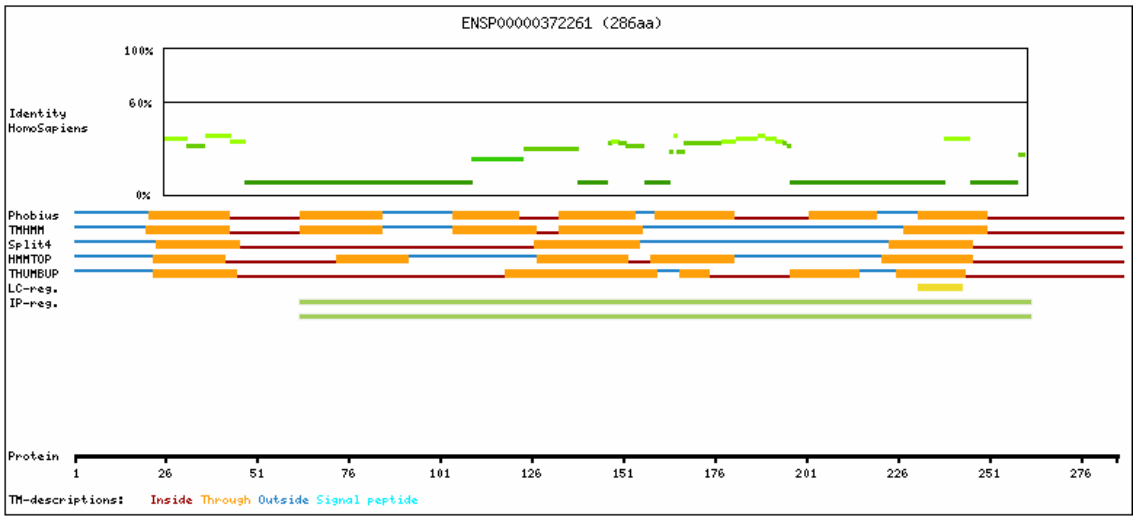

Figure 18. An example of the PrEST design tool with a protein with very non-agreeing prediction results ranging from three to seven TM regions predicted.

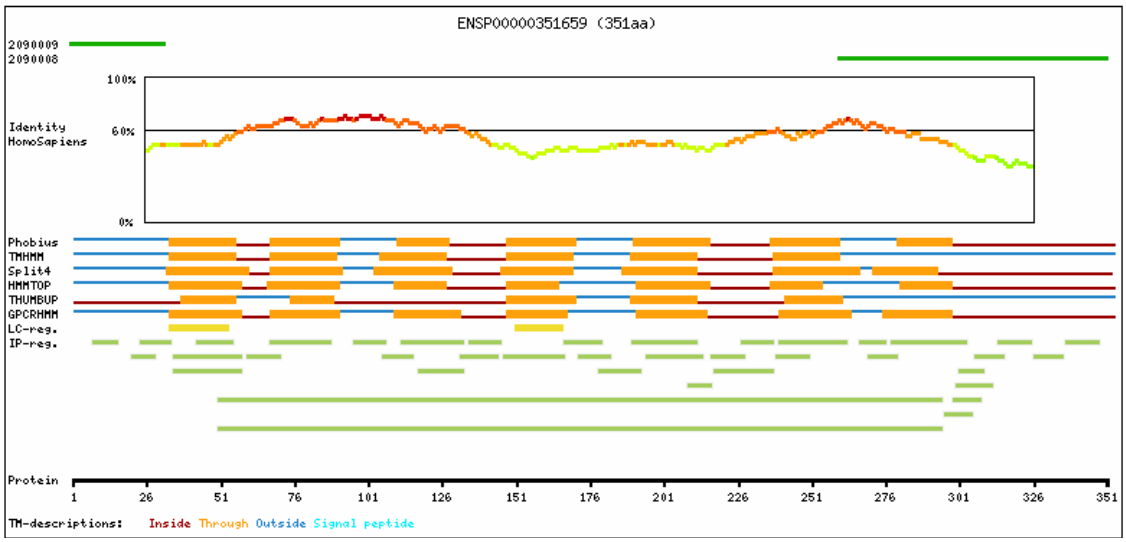

Figure 19. An example of the PrEST design tool showing the consequence of a PrEST designed using only TMHMM in the ProteinWeaver software.

5.5.3 PrEST design on membrane proteins

The criteria for PrEST design on membrane proteins are to some extent different than the criteria for non-membrane proteins. Transmembrane regions are avoided and consequently the PrEST is either designed in a loop or in the C-terminal or N-terminal region. The shortest PrEST length allowed for a membrane protein is 25 amino acids. If it is possible to design two PrESTs, it is desired to place one on an extracellular region and one on the intracellular region of the protein. One motive for this is that epitopes on the inside of the plasma membrane will not be accessible to the antibodies for *in vivo* studies, but will still work for localization purposes in, for example, TMAs. However, it is not yet clear how well an antibody will be able to bind to a short loop region on the outside of the cell, depending on glycosylations or other post-translational effects.

For each protein in the PrEST design tool, a decision is made on how to evaluate the results of the prediction methods. If all predictions agree, the decision is easy, but since this is not always the case, other factors must be considered. A PrEST will only be placed on a sequence region predicted by one or more methods as membrane spanning if there are no other alternative regions where all putative TM regions can be avoided. Since prediction methods, even when the predicted helices overlap, often predict different start and end positions of the helices, the loop available for PrEST design will have different predicted lengths depending on the method.

6. Validation of membrane protein topology predictions

The purpose of the experimental section of this project was to analyze if it is possible to validate the results of the prediction methods using a set of HPA antibodies. The selected methods for the validation were flow cytometry and confocal microscopy. Flow cytometry was used since it is a time efficient technique to analyze surface proteins on thousands of cells with high sensitivity. By labeling live cells with antibodies towards proteins predicted to be integral to the cell membrane, and using fluorescent activated cell sorter (FACS) analysis, it is possible to investigate if an antibody has been able to bind to a protein on the cell surface. An antibody will only bind to its target protein if the corresponding position of the PrEST is on extracellular parts of the protein, since the FACS analysis will be performed on non-permeabilized, live cells.

Confocal microscopy provides 3D-images of the localization of an antibody bound to its PrEST antigen in a labeled cell. As a pilot study for high throughput analysis, it was decided to analyze 25 antibodies targeting membrane proteins in the confocal microscope. Unfortunately, problems with the setup of the confocal microscopy had the consequence that the experiments were delayed and the results from that study are thus not available.

6.1 Selection of suitable HPA antibodies and cell lines

The first step in selecting a dataset with antibodies towards potential membrane proteins was to the number of available antibodies. The FACS analysis was performed on several cell lines, implying the necessity to know if the antigens corresponding to the selected antibodies are expressed in each cell line being analyzed. The easiest way of getting this information is by analyzing results of the cell microarrays (CMAs) in the TMA module of the HPA program. By the time of this selection, 741 antibodies had been analyzed in CMAs on 22 cell lines. The sequences of the proteins corresponding to the 741 antibodies where run in Phobius, which was the only prediction method implemented locally at that time. 56 of these were predicted to have at least one transmembrane region.

Next, it was decided which cell lines to use for the FACS analysis. Five of the eight cell lines selected to be part of the confocal microscopy study were known to be fast growing and easy to work with. The CMA images of these five cell lines were visually inspected for each of the 56 predicted membrane proteins, and each antibody was given a comment depending on the sub-cellular expression pattern; cytoplasmic, membranous, nuclear or other. An example of membranous expression in a cell line (Figure 20) and in a normal tissue (Figure 21) by a membrane protein targeting antibody is displayed below. The brown staining displays the location where the antibody has been able to bind to its corresponding protein and the blue staining visualizes the cells.

In some cases, more than one approved antibody towards the same gene was found, and these were also added to the list, even if no CMA images were available for the additional antibodies. The literature annotations were studied for all target proteins, and those annotated as present in the ER, Golgi or nuclear membrane were removed from the list of interesting candidates, as only plasma membrane proteins are valid for this project.

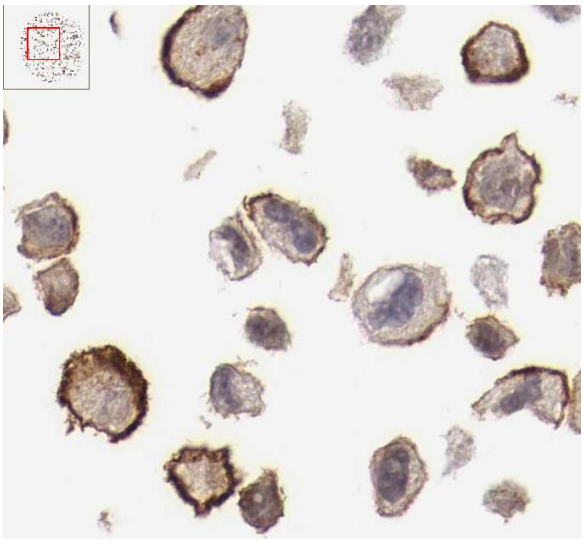

Figure 20. Expression of the antibody HPRK330043 in the cell line PC-3.

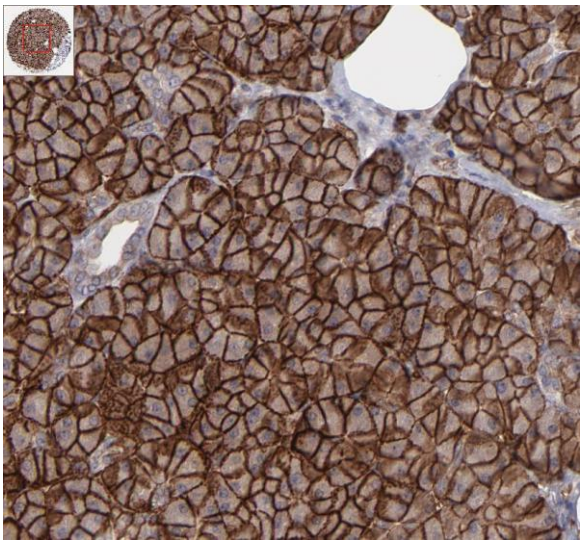

Figure 21. Expression of the antibody HPRK330043 in pancreas tissue.

A number of prediction methods additional to Phobius were run manually for this smaller set of antibodies and the topology prediction for the PrEST region was noted. After the literature annotations were analyzed, less than 25 antibodies remained in the list. By performing a database search of Gene Ontology (GO) ids for 'plasma membrane' (GO:0005886) and 'integral to plasma membrane' (GO:0005887) some extra antibodies were added. There were no available CMA images to study for these antibodies, but the values from an automated image analysis for the interesting cell lines were obtained from the database.

Antibody #	Antibody name	Gene_name	Phobius # TM/ PrEST location	HMMTOP # TM/ PrEST location	THUMBUP # TM/ PrEST location	TMHMM # TM/ PrEST location
1	HPRK230178	GABRA3	4/in	4/out	3/in	4/out
2	HPRK370006	SLC9A3R2	0/-	0/-	0/-	0/-
3	HPRK230039	TRPC5	8/out	9/in	8/in	7/out
4	HPRK300001	PTPRC	1/out	2/out	1/out	2/out
5	HPRK330043	ICAM1	1/out	1/out	1/out	1/out
6	HPRK330072	ICAM1	1/out	1/out	1/out	1/out
7	HPRK230269	KCND1	6/out	6/in	6/in	5/out
8	HPRK330040	ICAM2	1/out	1/out	1/out	1/out
9	HPRK330020	ECE1	1/out	1/out	1/out	1/out
10	HPRK320032	ERBB2	1/out	3/out	1/out	2/out
11	HPRK300030	ERBB2	1/in	1/in	1/in	2/out
12	HPRK250030	CTAGE5	1/in	1/in	1/in	1/out
13	HPRK250080	CTAGE5	1/in	1/in	1/in	1/out
14	HPRK230289	GPR101	7/in	7/in	7/in	7/in
15	HPRK220066	NP_443138.1	1/in	1/in	2/in	1/in
16	HPRK230140	IL1RAPL1	1/in	0/-	1/out	1/in
17	HPRK330053	SELP	1/out	1/out	3/out	1/out
18	HPRK330084	SELP	1/out	1/out	3/out	1/out
19	HPRK300029	SILV	1/out	1/out	1/out	1/out
20	HPRK300011	TYRP1	1/out	1/out	1/out	1/out
21	HPRK370022	ITGB5	1/out	1/out	2/out	1/out
22	HPRK330029	TJP1	2/out	0/-	1/in	0/-
23	HPRK330030	TJP1	2/out	0/-	1/out	0/-
24	HPRK260011	EPHA7	1/out	2/out	3/out	1/out
25	HPRK230337	GPR64	8/out	8/out	8/out	7/out

Table 2. The selected dataset of 25 antibodies. The table shows the HPR-id and gene name for each antibody, plus the resulting in/out location of the corresponding PrEST and number of predicted TM regions according to four prediction methods.

For the final list of 25 antibodies (Table 2), the three cell lines K-562, U-251mg and U-2OS showed the highest expression and were selected for the FACS analysis. K-562 consists of erythroleukemia cells derived from a chronic myeloid leukemia patient. U-2OS are human bone osteosarcoma cells derived from bone tissue, whereas U-251mg is a human malignant glioma cell line. Both U-251mg and U-2OS are adherent cells that grow attached to a surface, whereas K562 are non-adherent cells as they are derived from cells that exist in the blood stream.

6.2 FACS analysis

The FACS instrument can detect a range of properties of cells and allows high-throughput analysis of cells according to size, granularity, DNA content and surface markers. In this validation study the cells were analysed according to fluorescence intensity, which include fluorescent emission and scattering of light. The light source is a laser (or several lasers) with photodetectors measuring the forward scatter (light passing through the single cells in the stream), and the side scatter (light reflected on the cell surfaces and detected at side detectors). Once light signals have been detected, the stored data files can be analyzed in a number of ways on the connected computer. An example of the principles of a flow cytometer is displayed in Figure 22.

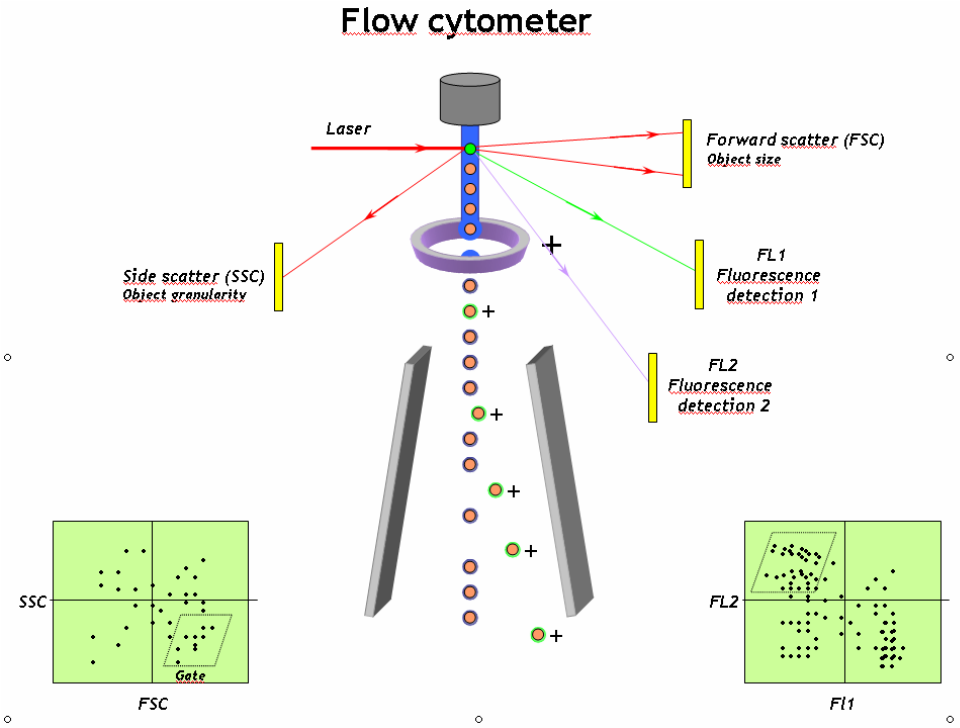

Figure 22. The principles of a flow cytometer.

Since the FACS is the most sensitive and commonly used high throughput cell sorter, it was suitable for this study with a dataset of 25 antibodies and three cell lines. One cell line was examined at a time and different positive controls were used for each FACS run. A fluorescent secondary antibody was used to detect the binding of the HPA antibody. A positive signal from the FACS analysis would indicate that the antibody had been able to bind to the PrEST, suggesting that this part of the protein is located on the outside of the cell, i.e. since the cells were alive and non-permeabilized, the antibodies cannot penetrate the plasma membrane and bind to the cell interior. However, the HPA antibodies have not been used on proteins with their native structure so a negative signal would not provide any definite information. Additionally, the number of surface proteins to be detected is not known, meaning that the sensitivity of detection might not be high enough to detect the low abundant surface proteins. To be able to detect the level of fluorescence from the cells and the background fluorescence when labeled with the secondary antibody, two negative controls were used throughout the study. As shown in Table 3, negative control ‘a’ consists of unlabelled cells and negative control ‘b’ of cells only labeled with the secondary antibody to show background binding. A picture of the FACS instrument is displayed in Figure 23.

For analysis and correct interpretation of the FACS signals, it was necessary to add a positive control, for each of the cell lines, to the dataset. However, finding a commercial rabbit antibody for a receptor of these three cell lines proved to be difficult. To be able to use the same secondary antibody as for the primary antibodies, it was

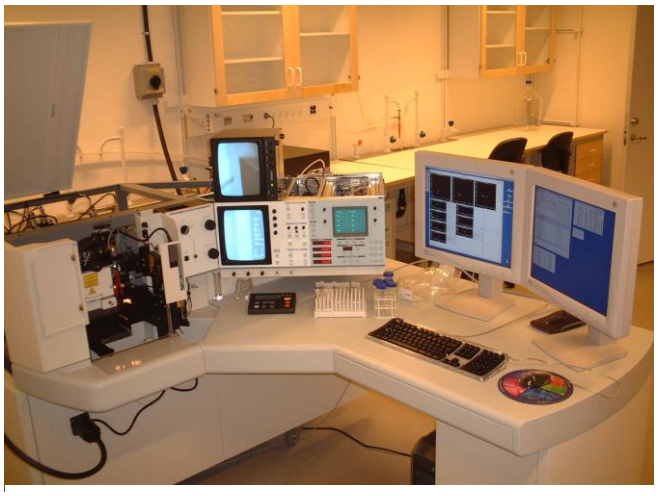

Figure 23. FACS Vantage SE (BD Biosciences).

crucial that the positive control also was a rabbit antibody. In the end, two HPA antibodies were selected as positive controls after investigating literature of expressed receptors for the cell lines (control 'e'). For K-562, the positive control selected was an HPA antibody against CD44, and for U-2OS an HPA antibody against Interleukin-17 receptor B precursor. No suitable HPA antibody as positive control was found for the U-256mg cell line. The lack of reliable positive controls makes interpretation of the results more difficult. The abbreviations used for the positive and negative controls are showed in Figure 24.

As another type of control only used for a few of the antibodies, the corresponding PrEST antigen was added to the well with the antibody, before labeling the cells. This was performed to try to prevent the antibody from binding to the cell by blocking it with the PrEST. This control 'c' therefore consists of cells + corresponding PrEST + primary antibody + secondary antibody. For one of the antibodies, a control 'd' with a PrEST corresponding to another antibody, i.e. an "irrelevant PrEST" was used to check for unspecific binding to the PrEST.

a	negative control: cells only
b	negative control: cells + secondary ab
e (K-562)	positive control: (cd44, HPRK330078)
e (U-2OS)	positive control: (I17RB_HUMAN, HPRK510032)
c	cells + correct PrEST + primary ab + secondary ab
d	cells + irrelevant PrEST + primary ab + secondary ab

Figure 24. The positive and negative controls

6.2.1 Materials and Methods

Cell cultures: The K-562 cells were grown in culture medium RPMI 1640 + 2mM Glutamine + 10 % Foetal Bovine Serum (FBS) + 1 % antibiotics/antimyotics (aa) and had a split ratio of 1:7-1:10. The culture medium for U-2OS was McCoy 5a + 10 % FBS + 1 % aa, with a split ratio of 1:3-1:6. The U-251mg culture medium was MEM + Earle's + L-glutamine + 10 % FBS and 1% aa, and the split ratio was 1:10. The two adherent cell lines were released from the plate with trypsin (0,25 %). The trypsin was washed off by adding medium followed by centrifugation at 1700 rpm for 3 minutes, before splitting the cells. All cells were incubated at 37º C and 5 % CO_2. The concentration was ~10^6 cells/ml and they were passed every 3rd-4th day.

Labeling of cells: The primary antibodies had a concentration of 5 µg/ml and a volume of 70 µL in a 96-well microtiter plate, with negative and positive controls included among the samples. For the control 'c' samples, the PrESTs were diluted 10x or 100x in 1xPBS+1%BSA and added to the primary antibodies with a concentration of 100x the concentration of primary antibodies. Cells were prepared in another 96-well plate with a concentration of ~150000 cells / well. The adherent cells were first released from the plate, either by using trypsin or a cell scraper depending on FACS run. 1xPBS (Phosphate Buffered Saline)+1%BSA (Bovine Serum Albumine) was used as a wash buffer to wash the cells and they were pelleted by centrifugation for 3 min at 1700 rpm. After two wash- and centrifuge steps, the cell pellets were resuspended with the primary antibodies or 70 µL 1xPBS+1%BSA for the negative controls, and incubated for 45 minutes at room temperature (RT). After washing and centrifugation, a secondary antibody was added to all wells except for the well with control 'a' and incubated for another 45 minutes at RT. The secondary antibody was the fluorescently labelled Goat-anti-rabbit IgG Alexa Fluorofor 488 with a final concentration of 5 µg/ml (concentration at delivery: 2mg/ml). The excess secondary antibodies were washed off by centrifugation and addition of 1xPBS+1%BSA. For each sample, a FACS tube was prepared with 150 µL 1xPBS+1%BSA and the labelled cells were transferred to the tubes with another 150 µL 1xPBS+1%BSA.

FACS analysis: The flow cytometer used to analyse the samples was a FACS Vantage SE (BD Biosciences). The sorting was performed at a rate of approximately 300-400 cells/s and at 488 nm.

6.3 Experimental Results

The first cell line to be analyzed in the FACS was K-562. Since the cells are non-adherent, there was no need to use trypsin or cell scrape. In Figure 28a, an example of the cell distribution of this cell line in the FACS is shown. The results of the two separate runs are displayed in Figure 25 as Mean Fluorescence Intensity (MFI) values. Sample 5 gave the highest MFI value, but most of the samples had MFI values of ~20. Sample 5 is predicted to be situated on the outside of the cell and is compared to the positive control for K-562 in Figure 28b. Although the positive control shown in green was not the most suitable control, it can clearly be concluded that sample 5 gives a higher signal than this control. Sample 13, displayed in Figure 28c and believed to be on the inside of the membrane, also resulted in a low signal in the K-562 runs and is found to the left of the positive control.

PrEST samples were selected for the second run after the results of the first run were analyzed. For K-562, a sample with high MFI (sample 5), and a sample with low MFI (sample 21) were selected. A molar excess of 10x the concentration of the antibodies was used and the results of this run can be viewed to the rightmost part of Figure 25. For sample 5, the MFI decreased from 121 to 56 by adding the PrEST, indicating that blocking was successful, at least to some extent. For sample 21, adding the corresponding PrEST only changed the MFI value from 35 to 33, indicating that this either is not a positive signal, or that the PrEST concentration was too low to affect the binding of the antibody to the target protein.

To be able to compare the signals (approximate comparison since exact numbers cannot be compared) between the different runs and cell lines, the negative control ‘b’ containing cells labelled with the secondary antibody was always set to MFI values of 12-14. The difference between control ‘b’ and ‘a’ (containing unlabelled cells), can be used to check the auto-fluorescence of unlabelled cells, i.e. secondary antibodies binding unspecifically.

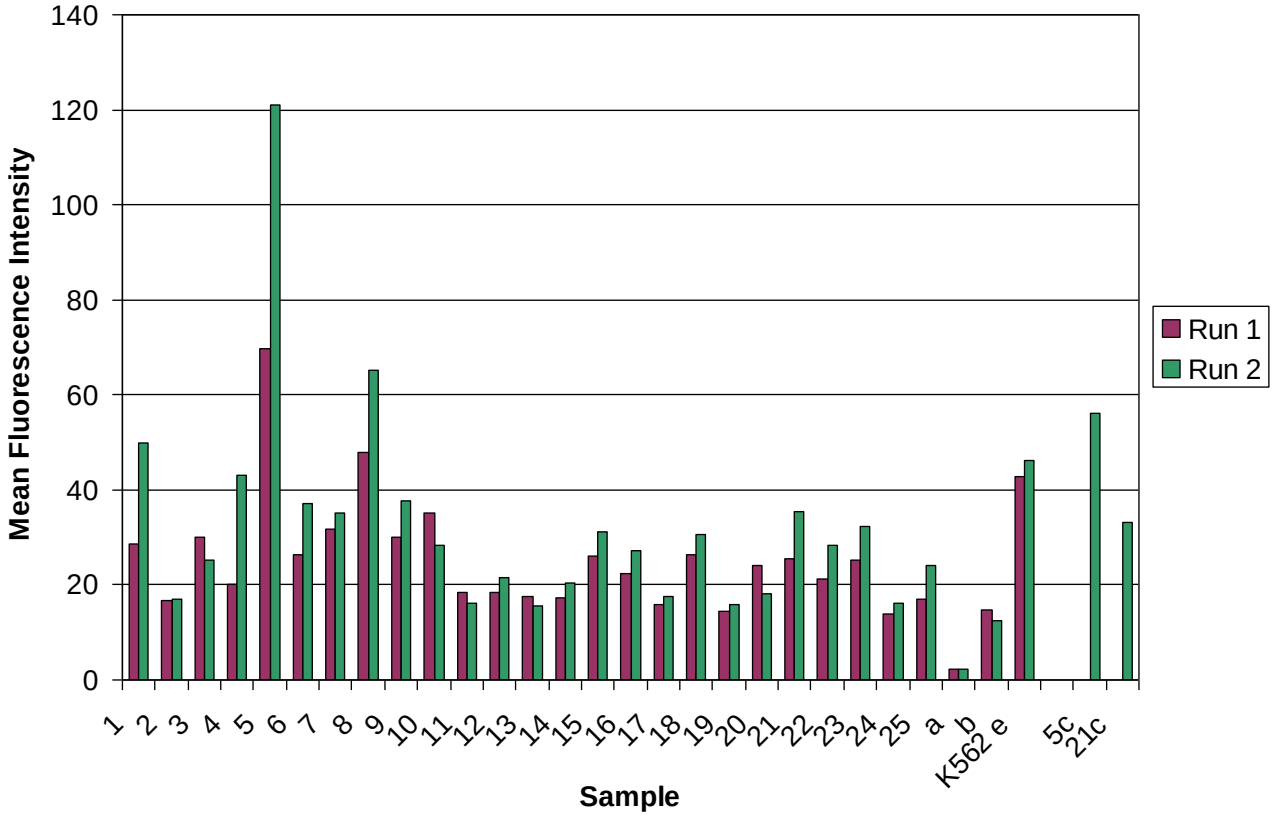

Figure 25. FACS results from labeled K-562 cells.

The second cell line analyzed was U-2OS and the results are displayed in Figure 26. This cell line showed less variation in MFI than K-562, but had one sample with a high MFI, sample 1. In the first two runs, the cells were released from the plate using trypsin, but since this might damage the receptors of the cells; some samples were run a third time and released with a cell scraper to be able to compare the two release methods. The MFI for sample 1 was more than doubled when using the cell scraper.

The second attempt to block the antibodies by adding PrESTs was performed using samples 1 and 10, after the results from the first run were analyzed. This time the concentration of PrESTs was increased to 100x the concentration of antibodies, since 10x did not appear enough to block the antibody completely in the K-562 run. Sample 1 generated the highest MFI value for U-2OS, and adding the PrEST decreased the MFI from 147 to 43. Sample 10 was chosen for its low MFI and adding the PrEST changed the signal from 20 to 18. Hence, sample 1 can be presumed to be a positive signal, but sample 10 is uncertain for the same reasons as sample 21 described above.

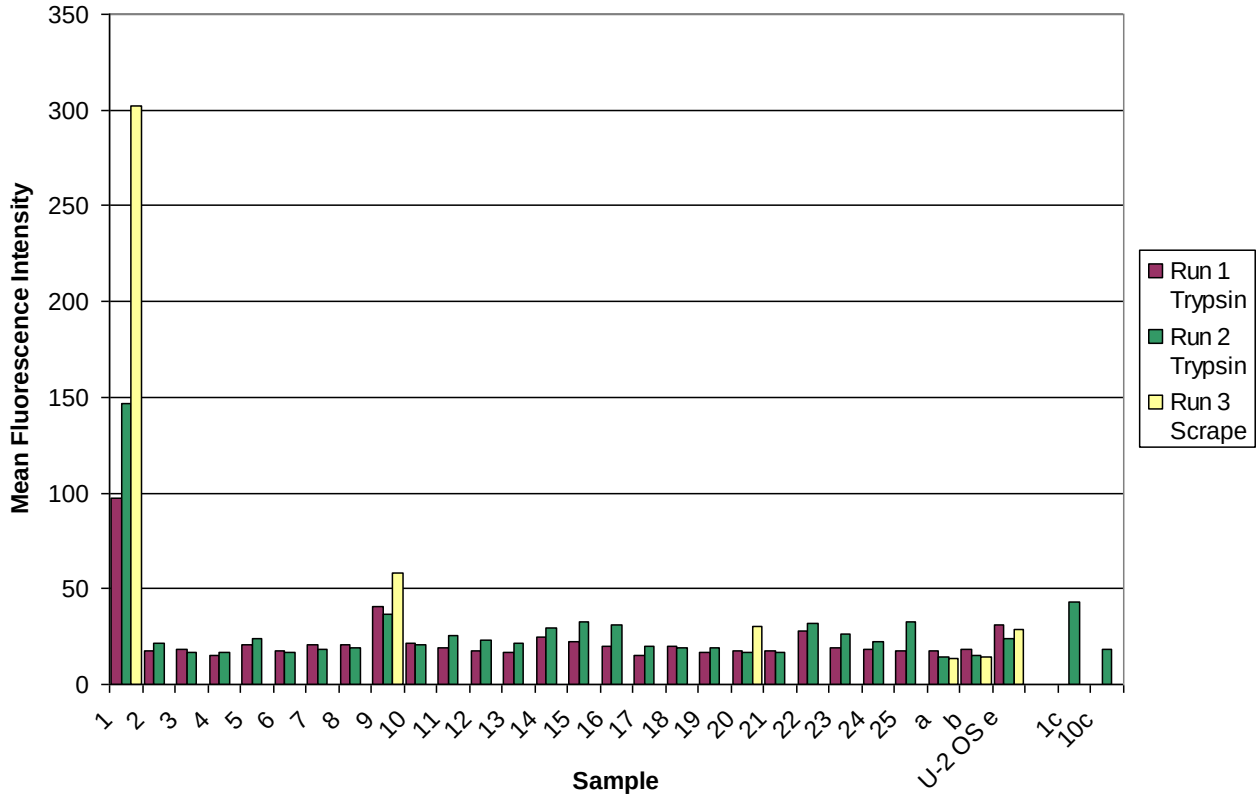

Figure 26. FACS results from labeled U-2OS cells.

The last cell line was U-251mg. In the first two runs, a cell scraper was used to release the cells, and in a third run, trypsin was used for a few of the samples to compare the outcome. The average signal in scraped U-251mg was much higher than for the other cell lines; even for the signals believed to be negative (Figure 27). This may be due to the usage of cell scaper in both two first runs. There are a number of signals ranging between MFI 80-200, and these may be positive results but are hard to interpret. For example, sample 14, a GPCR, is predicted as inside by all prediction methods, and sample 11, ERBB2, as inside by all but one.

Sample 1 was further analyzed with PrEST blocking in the FACS run with U-251mg. Here, adding the corresponding PrEST decreased the MFI from 660 to 100. To rule out that adding any PrEST affects the binding of the antibody and check for unspecific PrEST binding, the PrEST corresponding to sample 21 was added as a control ‘d’. This resulted in a MFI of 600, which indicates that the antibodies still can bind to the receptors, ruling out unspecific binding.

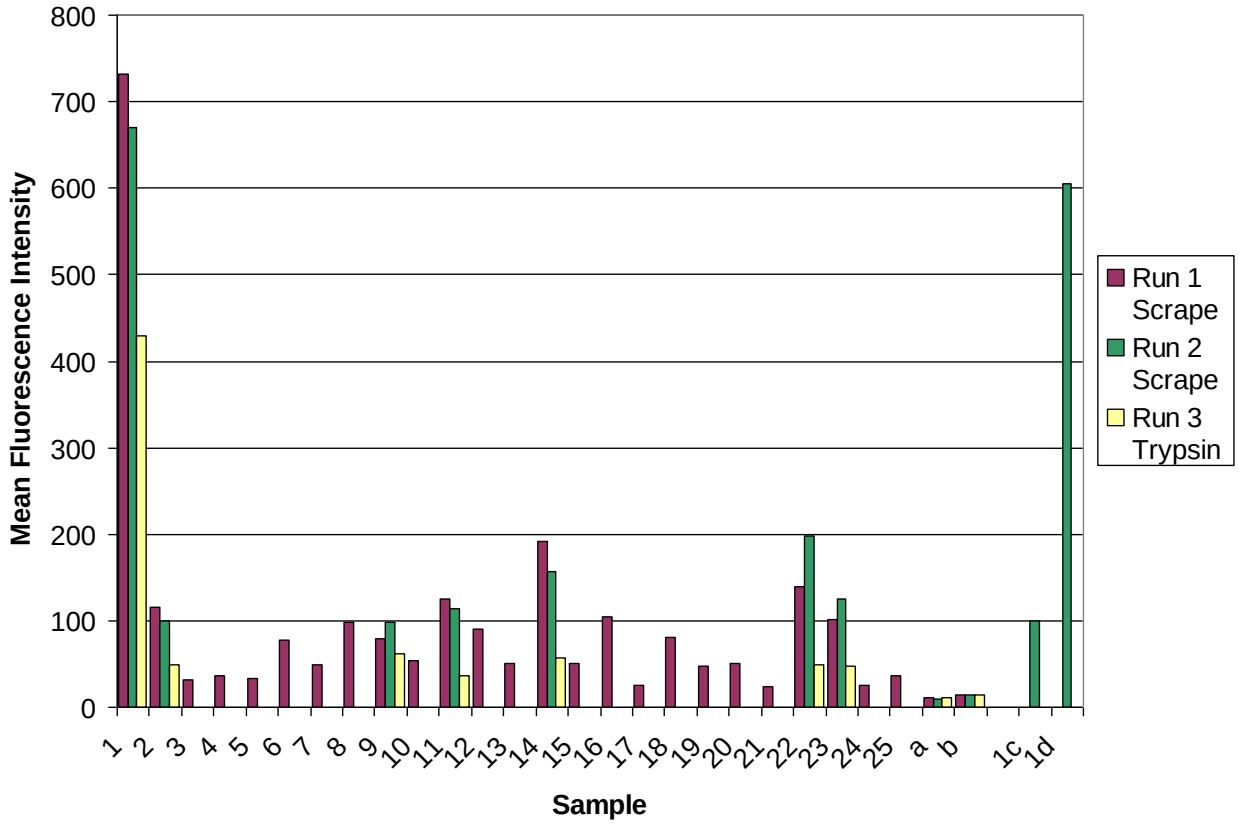

Figure 27. FACS results from labeled U-251mg cells.

In Figure 28, an example of the cell distribution and three examples of FACS histograms are provided. Sample 1 gave the highest MFI value in the U-251mg cell line, and the histogram in Figure 28d) includes the controls 'c' and 'd'. It is clear that the signals from the normal sample and from adding the irrelevant PrEST are very similar. Sample 5 and 13 are predicted to be on the inside versus outside and their results from cell line K-562 are compared to the positive control in Figure 28b and c.

Since sample 1 resulted in such a strong signal in U-251mg and U-2OS, this antibody was analyzed in more detail. It corresponds to the gene GABRA3, or the alpha-3 subunit of the gamma-aminobutyric-acid (GAB(A)) receptor, which is a ligand-gated chloride channel. There are 18 known human GABR subunits[59] and since several genes of this family show high sequence similarity, a BLAST search[58] was performed of the PrEST sequence against all human proteins. A multiple alignment was generated with the genes that got the highest BLAST score and 8 of these were so similar in the region corresponding to the PrEST that the epitopes that antibody binds to most likely can be found in any of them. All of these high similarity genes were part of the GAB(A) family and form a multisubunit channel in the plasma membrane. One explanation for the high MFI value for this antibody can therefore be that it binds to all GAB(A) receptors of the cell, which would also explain the strong staining seen in the CMA images.

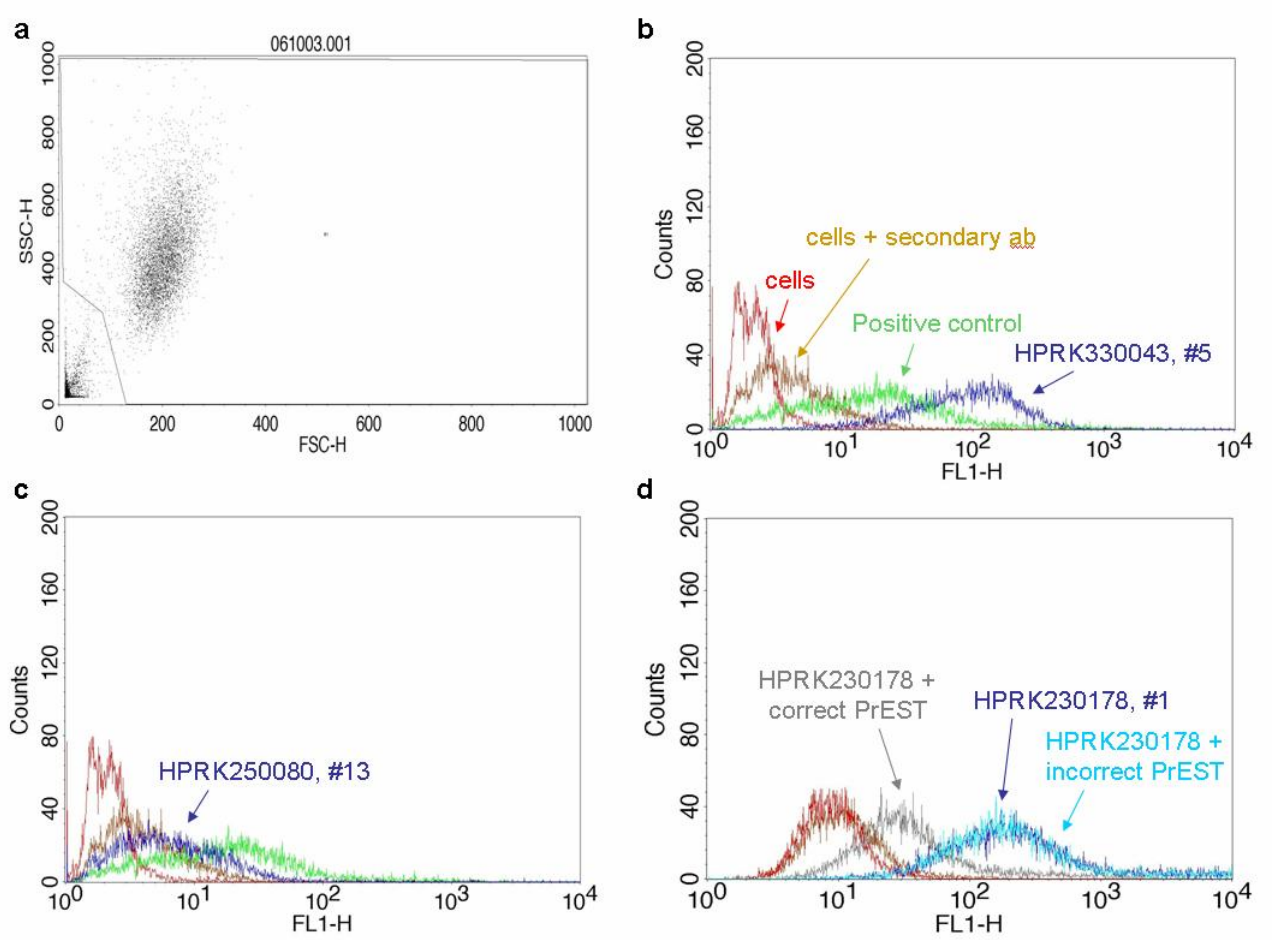

Figure 28. FACS results with one example of the cell distribution and histograms for three antibodies **a)** the cell distribution for the K-562 run **b) t**he result for sample 5 with negative controls 'a' and 'b' displayed in red and brown, positive control in green and the antibody in blue **c)** the results for sample 13 displayed in the same colors as the histogram for sample 5 **d)** the results for sample 1 in the U-251mg run. Control 'c' (corresponding PrEST added) in grey, control 'd' (irrelevant PrEST added) in turquoise, and the antibody in blue. Negative controls are displayed in red and brown.

7. Discussion

7.1 Prediction methods

As described in section 4.4, analyzing and comparing prediction methods for membrane topology involves several difficulties. The lack of reliable measures of the performance and accuracy of prediction methods consequently made selection of methods to use in the HPA PrEST selection software complicated. It is important to study publications about evaluation of prediction methods in detail, since it is easy to miss aspects such as prior removal of signal peptides from a test set, or that a method can not distinguish between membrane and globular proteins.

The result of the comparison of prediction methods to a set of TMHMM predicted 6TM proteins in Figure 12, show that the distribution of the number of TM regions is wide. There are many variations between the methods ranging from three to over ten predicted TM segments. However, most of the proteins are in the 6 and 7 TM groups. When comparing the results of the methods protein by protein, it was estimated that approximately only 1/8 of the proteins had all methods agreeing in the number of TM regions. Of the total number of 281 proteins, 76 were predicted as GPCRs, and the consequence was that GPCR-HMM was added to the list of selected prediction methods. Since one of the goals for the HPA program is to find interesting biomarkers and candidate genes for the pharmaceutical industry, the GPCRs is one of the most fascinating gene families to work with.

The results of the whole-genome scan in section 5.4 show a major difference in number of membrane protein coding genes predicted by the selected methods (Figure 14). The Venn diagram in Figure 15 displays the overlap between Phobius, TMHMM and THUMBUP, and especially the first two methods predict a similar number of

genes to code for membrane proteins. Since Phobius and TMHMM are built on similar models, this is not very surprising. The overlap between all three methods is 4930 genes, which is a large part of the total number of predicted genes, and it is reassuring to find that although these methods use different techniques, they are comparable.

The high number of predicted membrane protein coding genes for HMMTOP can be explained by the fact that the method is not constructed to discriminate between non-membrane and membrane proteins, and hence it is overpredictive. Some of the discrepancy between the different methods can be explained by the confusion of signal peptides with transmembrane regions. The overlap between Split and HMMTOP was 7476 membrane protein coding genes and 1126 of these overlap with the proteins that Phobius predicts to only have a SP region.

Given that Split 4.0 was run on a dataset only with proteins less than 1000 amino acids long, it can not be compared directly to the numbers of the other methods. However, the predictions by Split4.0 on the smaller dataset shows that this method is overpredictive compared to Phobius, TMHMM and THUMBUP. A final comparison of all methods including Split4.0 cannot be done until a new version of Split4.0 is released that can handle long protein sequences. Thus, Split4.0 probably also only will be visualized when at least one more method other than HMMTOP predicts a protein to have TM regions.

One of the main problems in implementing prediction methods for the purposes of this project, in addition to that many of the methods are not suitable for whole-genome research, is the lack of synchronized input and output formats. A standard output format for displaying the result of a prediction would provide easy parsing of the text files. Although most programs take a FASTA-formatted file with protein sequences as input, this is not always the case. This is something all developers of prediction methods should consider.

During the past years, new prediction methods have been developed, and due to the fact that most comparative evaluation studies were published in 2001 and 2002 [19, 48, 53], no comprehensive evaluation of the newer methods is available. Today there is more structural membrane protein data available, which can provide additional information and more accurate predictions. Fortunately, the design of the HPA database and the way the prediction methods were implemented make adding or removing of methods straightforward. For example, if a new comparative study is published, presenting better, more advanced or more suitable methods, these can easily be incorporated into the PrEST design tool, to replace some more poorly performing programs.

7.2 PrEST design on membrane proteins

In the PrEST design of the HPA program each gene is manually considered, giving the opportunity to simultaneously view information from various types of input. The visualization of the result of the prediction methods enables a manual evaluation of the results of multiple prediction methods, as opposed to consensus prediction methods where each method is weighted in the final decision with a fixed value.

The first draft of the PrEST design tool is now finished, but before it can be used for PrEST design, more functionality needs to be added, and it needs to be integrated with the rest of the LIMS system. The strength of the PrEST design tool is that any desired feature can easily be added and the user might also choose what features to be displayed. The visualization can also give important additional information in the evaluation of experimental results in the other modules of the pipeline, i.e. in the annotation of TMAs.

One of the difficulties in selecting PrEST regions for membrane proteins is when the different prediction methods do not agree. When it is possible to avoid all predicted putative transmembrane regions, this is of course done. Nonetheless, sometimes it might be necessary to place the PrEST in a region where some of the methods predict a TM helix. The strategy of the PrEST design is to rather make an attempt with a PrEST containing a putative TM region, than to fail the gene. To know how a presumed TM helix affects the results of the PrEST in the rest of the HPA pipeline, more experimental data from the pipeline is needed.

Figures 16-19 were examples of visualization of proteins and PrESTs in the PrEST design tool. These examples show both conflicting results, but also TM regions predicted the same by all methods. Figure 16 showed a protein predicted to contain a signal peptide by Phobius and SignalP. Three of the TM regions have agreeing results from all prediction methods, but there are also TM regions only predicted by one or two methods. The conflicting inside/outside predictions in the whole protein show the difficulties in predicting topology, and the consequences of having an extra TM segment inserted. In Figure 17, a protein with one or two TM regions illustrates how the prediction methods disagree on the length of the first segment; it is predicted as either two shorter regions (Phobius and TMHMM) or one long (THUMBUP and HMMTOP), whereas Split4.0 finds a

completely different TM region. Figure 18 illustrates the difficulty in finding a suitable PrEST region when the predictions are very disagreeing. Almost the whole protein is covered in predicted membrane spanning helices. Figure 19 is an illustration of the consequence of only using the results from one method, as these PrESTs were designed in ProteinWeaver using TMHMM. According to the prediction results of the other methods, the second PrEST most likely contains a membrane spanning region. This Figure also shows the results from GPCR-HMM, indicating that this protein is a 7-helix GPCR and predicted so by all methods except two.

Results from approved HPA antibodies towards membrane proteins are a prerequisite to be able to analyze whether PrESTs on the extracellular or intracellular parts of membrane-spanning proteins are the most suitable for the HPA program. PrESTs on the extracellular parts are preferred to enable *in vivo* studies with the HPA antibodies but if it turns out that antibodies directed to short loops on the outside of the plasma membrane have a low success rate, the strategy might have to be changed. Short PrESTs yield poorer immune response and lower concentration of antibodies after immunization. On the other hand, shorter proteins are easier to express in *E. coli*. Another potential problem with PrESTs on extracellular domains is post-translational modifications, such as glycosylations, that occurs only on the outside of the plasma membrane. Post-translational modifications are not identical between the bacterial expression system and the mature human protein and this might therefore disturb the recognition site of the generated HPA antibody.

7.3 Analysis of Experimental Results

Projects that are part of the HPA program are unique due to the availability of a high number of antibodies towards human proteins and corresponding data about expression patterns in tissues and cell lines, a fact that this master thesis project benefits from and makes use of.

Several issues make analysis of the FACS results difficult. In an ideal FACS experiment, the target protein would be over-expressed and the interactions easily monitored. However, in this case, it is unclear how many of the corresponding proteins that are actually highly expressed in the selected cell lines. Combined with the lack of a suitable positive control for the cell lines, the levels of the signals for the antibodies are hard to interpret. Another difficulty is that the antibodies have not been tested against proteins with native structures, so even if the CMAs show that the target is expressed in the same cell line and the antibody binds to the denaturated protein, it may not be able to bind to the native proteins in the living cells used for the FACS experiment.

When analyzing the results for all three cell lines, it is clear that the main problem lies in interpreting the limit between negative and positive signals. There are a few antibodies that clearly show high signals, e.g. sample 5 in K-562 and sample 1 in both U-2OS and U-251mg. However, although many signals are higher than the negative control, it is not clear which of these really indicate binding of the antibody. For the adherent cell lines, both trypsin and cell scrape to release the cells were used. The results from this show that using the cell scrape gave higher signals; in some cases the signal was more than doubled. The results from the FACS experiments for U-251mg in particular resulted in many high signals. The question is if this has to do with receptors being damaged when trypsin is used (maybe the most probable cause), or if the high signals using the cell scrape can be explained with an uneven grade of labelling of the cell population. The latter phenomena might be caused by heterogeneous cell samples or protocol-specific issues (labeling of the cells) that perhaps could be optimized.

The attempt to block the antibodies from binding by introducing the corresponding PrEST antigens turned out well. The increase in MFI by adding the PrEST for sample 5, indicates that this antibody induces a positive signal and hence that the protein is situated on the outside of the cell membrane. All topology prediction methods also predicted the PrEST as located on the outside. Interestingly, sample 6 is an antibody towards the same protein and situated on the same side of the TM segment as sample 5. The reason for the weaker signal of sample 6 could be that this part of the sequence is not accessible to the antibody in its native form. Structural analysis of this exoplasmic loop may provide an answer to this, and is possible since the structure of the corresponding protein ICAM is available in PDB. The PrEST blocking of the samples 21 and 10, which are predicted as outside but seem to give negative signals, did not result in any major changes in MFI and conclusively do not indicate a positive binding of the antibodies.

If the high signal of sample 1, or the GABRA3-antibody, can be explained by binding to the several genes from the same protein family, the question is which of the other medium signals actually are positive signals. For the U-251mg results, there are several MFI values >100. More than one is for a PrEST that all prediction methods predict to be situated on the inside of the membrane. Therefore additional analysis of these antibodies is needed to investigate the binding. As expected, the FACS results seem to be reproducible since the two runs of each cell

line is comparable to each other, when using the same method to release the adherent cells from the bottom of the plate.

For further analysis using FACS, the protocol for labelling cells can be optimized in several ways, i.e. by adding more wash steps for longer time periods or test different secondary antibodies. With few proteins expressed in each cell, increasing the signal-to-noise ratio is important to detect weaker signals. It would also be interesting to use permeabilized cells and compare the results to the live cells. Permeabilizing cells would allow antibodies to bind to either side of the membrane, i.e. antibodies can enter the interior of the cell. Also, other cell lines with different characteristics could be added to the study, since the results proved to be very cell line specific. To choose a Her-2 positive cell line and run Her-2 antibodies (both commercial and HPA antibodies), a positive signal in this cell line could be obtained. It would be optimal to have those kinds of positive controls for all cell lines studied. Changing the secondary antibody may affect the signal-to-noise ratio as well.

Only using the results of these FACS runs will unfortunately not give enough information to validate the prediction methods. However, optimizing the protocol may provide stronger positive signals and more information about the position of the PrESTs. Once the results from the confocal microscopy are available, further analysis of the results can be performed. If the confocal microscopy pilot study turns out well and provides a suitable high-throughput platform, then larger datasets with membrane proteins can be analyzed.

8. Conclusion

The answer to the question of how many membrane proteins the human genome codes for probably will not be found until more structures are discovered. The conclusion from the results of the whole-genome scan is that probably more than 25% of all genes contain membrane spanning regions. The newly discovered complexity to membrane helices has already changed the view of membrane proteins and there is a need for improved prediction methods. It will be interesting to see what techniques the next generation of prediction methods will be based on, and how the different types of helices will be dealt with.

The incorporation of membrane protein topology into the PrEST design tool provides a convenient way of analyzing the probable positions of TM segments as predicted by the selected prediction methods. This enables design of PrESTs on membrane proteins, hopefully avoiding TM regions and with a choice of extracellular or intracellular parts. Until more information about the success rate of PrESTs on the inside/outside of the membrane has been analyzed, it is desired to select one of each if possible. More analysis of the consequence of short PrESTs (<50 residues) on loops in the HPA pipeline also has to be performed.

Unfortunately the experimental results did not give enough information for a reliable validation of the methods. However, this pilot project shows promising novel applications of FACS for determination of membrane protein topology. A combination of flow cytometry and confocal microscopy would provide a reliable way of obtaining results of the inside/outside localization of antibodies towards eukaryotic membrane proteins.

Membrane proteins are in focus for research world wide. The findings of dual topology proteins and helices not behaving as expected show the complexity of this class of proteins. There are an endless number of areas in which the HPA antibodies can be applied. The strategy for PrEST design on membrane proteins and the implementation of prediction methods performed in this master thesis serves to allow for successful generation of antibodies directed towards membrane proteins. The experimental study has been a first step to show the possibilities for antibody-based proteomics in this field, and may in the future serve as important information for improving topology predictions.

9. Acknowledgements

I would like to thank all of the people that helped me during this project. Mathias Uhlén, for letting me do my master thesis in the HPA program and for giving me such an interesting topic to explore. Lisa Berglund, for supervising me and for giving excellent support. Sara Lindström, for all help with the FACS validation and for making me like the multipipette. Laurent Barbe, for performing the confocal microscopy study. Erik Sonnhammer for being my scientific reviewer and for advice about prediction methods. Erik, Kalle, Mattias, Per and Jan for all computer and LIMS-associated questions. Åsa, Cristina, Johan and Pawel for being excellent roommates and for nice discussions. I also want to thank my family for always supporting me.

10. Abbreviations

BLAST	Basic Local Alignment Search Tool
DB	Database
ER	Endoplasmic Reticulum
FACS	Fluorescent Activated Cell Sorting
IH	Immunohistochemistry
GPCR	G Protein-Coupled Receptor
HMM	Hidden Markov Model
HPA	Human Protein Atlas
HPR	Human Proteome Resource
LIMS	Laboratory Information Management System
mAb	Monoclonal antibody
msAb	Monospecific antibody
MFI	Mean Fluorescence Intensity
pAb	Polyclonal antibody
PDB	Protein Data Bank
PrEST	Protein Epitope Signature Tag
SP	Signal Peptide
TM	Transmembrane
TMA	Tissue Micro Array
TRAP	Translocon-Associated Protein Complex

11. References

1. Lindskog, M. in School of Biotechnology, Vol. Doctoral Thesis (Royal Institute of Technology, Stockholm; 2005).
2. Sonnhammer, E.L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6**, 175-182 (1998).
3. Cuthbertson, J.M., Doyle, D.A. & Sansom, M.S. Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng Des Sel* **18**, 295-308 (2005).
4. White, S.H. & von Heijne, G. The machinery of membrane protein assembly. *Curr Opin Struct Biol* **14**, 397-404 (2004).
5. Osborne, A.R., Rapoport, T.A. & van den Berg, B. Protein translocation by the Sec61/SecY channel. *Annu Rev Cell Dev Biol* **21**, 529-550 (2005).
6. Menetret, J.F. et al. Architecture of the ribosome-channel complex derived from native membranes. *J Mol Biol* **348**, 445-457 (2005).
7. White, S.H. & von Heijne, G. Transmembrane helices before, during, and after insertion. *Curr Opin Struct Biol* **15**, 378-386 (2005).
8. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
9. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422**, 193-197 (2003).
10. Agaton, C., Uhlen, M. & Hober, S. Genome-based proteomics. *Electrophoresis* **25**, 1280-1288 (2004).
11. Uhlen, M. et al. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics* **4**, 1920-1932 (2005).
12. Uhlen, M. & Ponten, F. Antibody-based proteomics for human tissue profiling. *Mol Cell Proteomics* **4**, 384-393 (2005).
13. Persson, A., Hober, S. & Uhlen, M. A human protein atlas based on antibody proteomics. *Curr Opin Mol Ther* **8**, 185-190 (2006).
14. Birney, E. et al. Ensembl 2006. *Nucleic Acids Res* **34**, D556-561 (2006).
15. Agaton, C. et al. Affinity proteomics for systematic protein profiling of chromosome 21 gene products in human tissues. *Mol Cell Proteomics* **2**, 405-414 (2003).
16. Uhlén, M. in Human Protein Atlas (http://www.proteinatlas.org, 2006).
17. Chen, C.P. & Rost, B. State-of-the-art in membrane protein prediction. *Appl Bioinformatics* **1**, 21-35 (2002).
18. Drews, J. Drug discovery: a historical perspective. *Science* **287**, 1960-1964 (2000).
19. Moller, S., Croning, M.D. & Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**, 646-653 (2001).
20. Melen, K., Krogh, A. & von Heijne, G. Reliability measures for membrane protein topology prediction algorithms. *J Mol Biol* **327**, 735-744 (2003).

21. Spieker-Polet, H., Sethupathi, P., Yam, P.C. & Knight, K.L. Rabbit monoclonal antibodies: generating a fusion partner to produce rabbit-rabbit hybridomas. *Proc Natl Acad Sci U S A* **92**, 9348-9352 (1995).
22. Lodish, L. Molecular Cell Biology, Edn. 4. (W.H. Freeman, New York; 2001).
23. Lindskog, M., Rockberg, J., Uhlen, M. & Sterky, F. Selection of protein epitopes for antibody production. *Biotechniques* **38**, 723-727 (2005).
24. Wimley, W.C. The versatile beta-barrel membrane protein. *Curr Opin Struct Biol* **13**, 404-411 (2003).
25. Wistrand, M., Kall, L. & Sonnhammer, E.L. A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein Sci* **15**, 509-521 (2006).
26. Attwood, T.K. A compendium of specific motifs for diagnosing GPCR subtypes. *Trends Pharmacol Sci* **22**, 162-165 (2001).
27. von Heijne, G. Recent advances in the understanding of membrane protein assembly and structure. *Q Rev Biophys* **32**, 285-307 (1999).
28. Aurora, R. & Rose, G.D. Helix capping. *Protein Sci* **7**, 21-38 (1998).
29. Heijne, G.v. Membrane-protein topology. *Nature Reviews: Molecular Cell Biology* **7**, 909-918 (2006).
30. Goder, V. & Spiess, M. Topogenesis of membrane proteins: determinants and dynamics. *FEBS Lett* **504**, 87-93 (2001).
31. Van den Berg, B. et al. X-ray structure of a protein-conducting channel. *Nature* **427**, 36-44 (2004).
32. Mingarro, I., Nilsson, I., Whitley, P. & von Heijne, G. Different conformations of nascent polypeptides during translocation across the ER membrane. *BMC Cell Biol* **1**, 3 (2000).
33. Heijne, G.V. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *Embo J* **5**, 3021-3027 (1986).
34. Magnus Monné, I.N., Marie Johansson, Niklas Elmhe, Gunnar von Heijne Positively and negatively charged residues have different effects on the position in the membrane of a model transmembrane helix. *Journal of Molecular Biology* **Volume 284**, 1177-1183 (1998).
35. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580 (2001).
36. Liu, J. & Rost, B. Comparing function and structure between entire proteomes. *Protein Sci* **10**, 1970-1979 (2001).
37. Kall, L., Krogh, A. & Sonnhammer, E.L. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**, 1027-1036 (2004).
38. Viklund, H. & Elofsson, A. Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci* **13**, 1908-1917 (2004).
39. Kyte, J. & Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105-132 (1982).
40. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol* **179**, 125-142 (1984).
41. von Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225**, 487-494 (1992).
42. Jones, D.T., Taylor, W.R. & Thornton, J.M. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33**, 3038-3049 (1994).
43. Rost, B., Casadio, R., Fariselli, P. & Sander, C. Transmembrane helices predicted at 95% accuracy. *Protein Sci* **4**, 521-533 (1995).
44. Rost, B. PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol* **266**, 525-539 (1996).
45. Krogh, A., Brown, M., Mian, I.S., Sjolander, K. & Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235**, 1501-1531 (1994).
46. Kall, L. in Center for Genomics and Bioinformatics, Vol. Doctoral thesis (Karolinska Institute, Stockholm; 2006).
47. Tusnady, G.E. & Simon, I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* **283**, 489-506 (1998).
48. Chen, C.P., Kernytsky, A. & Rost, B. Transmembrane helix predictions revisited. *Protein Sci* **11**, 2774-2791 (2002).
49. Tusnady, G.E. & Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17**, 849-850 (2001).
50. Nielsen, H. & Krogh, A. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol* **6**, 122-130 (1998).
51. Zhou, H. & 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).

52. Juretic, D., Zoranic, L. & Zucic, D. Basic charge clusters and predictions of membrane protein topology. *J Chem Inf Comput Sci* **42**, 620-632 (2002).
53. Kall, L. & Sonnhammer, E.L. Reliability of transmembrane predictions in whole-genome data. *FEBS Lett* **532**, 415-418 (2002).
54. Chen, C.P. & Rost, B. Long membrane helices and short loops predicted less accurately. *Protein Sci* **11**, 2766-2773 (2002).
55. Kahsay, R.Y., Gao, G. & 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).
56. Arai, M. et al. ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Res* **32**, W390-393 (2004).
57. Ikeda, M., Arai, M., Lao, D.M. & 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).
58. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
59. Simon, J., Wakimoto, H., Fujita, N., Lalande, M. & Barnard, E.A. Analysis of the set of GABA(A) receptor genes in the human genome. *J Biol Chem* **279**, 41422-41435 (2004).

Appendix 1: Example of output files

TMHMM Output

```
ENSP00000166244 len=1005        ExpAA=35.36     First60=1.41    PredHel=1       Topology=o541-563i
ENSP00000363776 len=1005        ExpAA=35.36     First60=1.41    PredHel=1       Topology=o541-563i
ENSP00000363775 len=495         ExpAA=2.02      First60=0.84    PredHel=0       Topology=o
ENSP00000262502 len=1030        ExpAA=245.24    First60=0.00    PredHel=11      Topology=i146-168o172-194i215-
237o257-279i286-308o340-362i375-397o452-474i506-523o528-550i571-605o
ENSP00000264926 len=495         ExpAA=0.05      First60=0.05    PredHel=0       Topology=o
ENSP00000373348 len=495         ExpAA=0.05      First60=0.05    PredHel=0       Topology=o
ENSP00000352054 len=1258        ExpAA=175.30    First60=0.00    PredHel=7       Topology=o377-399i419-441o855-
```

THUMBUP Output

```
>ENSP00000356745 ENSG00000000457

    2 maxB= 0.798  length=   688 ( is  TM )
 N-Term:in
 Helix 1:  99 113
 Helix 2:  397 406

>ENSP00000286031 ENSG00000000460

    4 maxB= 0.806  length=   853 ( is  TM )
 N-Term:in
 Helix 1:  159 170
 Helix 2:  417 429
 Helix 3:  447 461
 Helix 4:  498 503

>ENSP00000001380 ENSG00000000938

    0 maxB= 0.782  length=   529 ( not TM )
 N-Term:in
```

Phobius Output

```
SEQENCE ID              TM SP PREDICTION
ENSP00000003603          4  0 i21-42o62-80i92-114o211-233i
ENSP00000003616          1  0 i31-49o
ENSP00000363117          0  0 o
ENSP00000001567          0  Y n3-13c18/19o
ENSP00000356557          1  Y n9-21c30/31o54-76i
ENSP00000229416          0  0 o
ENSP00000374141          9  0 o34-54i75-96o102-121i133-152o172-194i201-229o241-259i271-293o299-320i
ENSP00000003912          7  0 o12-39i51-70o90-112i119-147o159-177i189-211o217-238i
ENSP00000363521          1  0 o27-45i
ENSP00000311573          0  0 o
```

HMMTOP Output

```
>HP: 334 ENSP00000363530 ENSG00000001460
 OUT   0
>HP: 406 ENSP00000374140 ENSG00000001461
 OUT   9   34   53   78   97  102  121  134  155  172  191  204  223  240  259  272  291  300  319
>HP: 324 ENSP00000003912 ENSG00000001461
 OUT   7   17   39   52   74   89  111  120  142  159  178  187  209  218  239
>HP: 141 ENSP00000363521 ENSG00000001461
  IN   1   27   46
>HP: 675 ENSP00000363937 ENSG00000001497
 OUT   0
>HP: 418 ENSP00000351284 ENSG00000002016
 OUT   1   93  111
```

GPCRHMM Output

First run: GPCR detection

Sequence identifier	global	local	pred
ENSP00000003603	-104.94	-	No
ENSP00000343380	Too short sequence		No
ENSP00000374140	17.87	31.10	GPCR
ENSP00000343549	19.66	31.10	GPCR
ENSP00000350722	-41.87	-	No
ENSP00000363520	16.44	31.10	GPCR
ENSP00000363521	Too short sequence		No
ENSP00000311573	-171.79	-	No

Second run: GPCR transmembrane segment localization prediction

SEQENCE ID	TM	SP	PREDICTION
ENSP00000374140	7	0	o101-121i134-156o167-187i202-223o239-260i271-293o300-321i
ENSP00000003912	7	0	o19-39i52-74o85-105i120-141o157-178i189-211o218-239i
ENSP00000363520	7	0	o107-127i140-162o173-193i208-229o245-266i277-299o306-327i
ENSP00000353385	7	0	o146-173i190-216o240-262i277-296o313-336i355-375o391-411i
ENSP00000352561	7	0	o146-173i187-207o224-246i261-280o297-320i339-359o375-395i
ENSP00000046967	7	0	o57-83i94-116o136-157i177-198o223-248i322-346o358-378i
ENSP00000358301	7	0	o57-83i94-116o136-157i177-198o223-248i322-346o358-378i
ENSP00000051619	7	Y	n5-15c20/21o141-161i171-192o205-223i237-257o276-298i317-338o342-365i