

# **Computational analysis of membrane proteins: genomic occurrence, structure prediction and helix interactions**

Ursula Lehnert<sup>a</sup>, Yu Xia<sup>a</sup>, Thomas E. Royce, Chern-Sing Goh, Yang Liu<sup>#</sup>, Alessandro Senes, Haiyuan Yu, ZhaoLei Zhang, Donald M. Engelman, Mark Gerstein<sup>\*</sup>

Department of Molecular Biophysics & Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA

<sup>#</sup> Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA

**Running title:** Computational genomics

\*Author to whom correspondence should be addressed: [mark.gerstein@yale.edu](mailto:mark.gerstein@yale.edu), phone: +1 203 432 6105; fax: +1 203-432-6946

<sup>a</sup>: These authors contributed equally to this paper.

Word count: 8877

**Abstract:**

We review recent computational advances in the study of membrane proteins, focusing on those that have at least one transmembrane helix. Transmembrane protein regions are, in many respects, easier to investigate computationally than experimentally, due to the uniformity of their structure and interactions (e.g. consisting predominately of nearly parallel helices packed together) on one hand and presenting the challenges of solubility on the other. We present the progress made on identifying and classifying membrane proteins into families, predicting their structure from amino acid sequence patterns (using many different methods), and analyzing their interactions and packing. The total result of this work allows us for the first time to begin to think about the membrane protein interactome, the set of all interactions between distinct transmembrane helices in the lipid bilayer.

1	Introduction.....	4
2	Genomic classification and analysis of transmembrane sequences.....	6
2.1	Advances in prediction of helical membrane protein topologies.....	6
2.1.1	Advances in 3D structure prediction of membrane helical proteins.....	11
2.2	Genome wide classification of membrane proteins .....	14
2.3	Integrative database systems.....	16
2.4	Co-evolutionary analysis of membrane proteins .....	17
2.5	Membrane proteins and pseudogenes .....	20
3	Structural characteristics of membrane proteins.....	22
3.1	Amino acid composition of membrane proteins.....	22
3.2	Helix packing and characterization of helical interfaces .....	24
3.3	Analysis of helix-helix interfaces in transmembrane proteins.....	26
4	Membrane protein interactions .....	28
4.1	The current excitement about protein networks.....	29
4.2	Identification of protein complexes with experimental techniques .....	31
4.2.1	Screening methods .....	31
4.2.2	Helix-helix interaction motifs .....	33
4.3	How many helix-helix interactions exist in a genome? .....	34
5	Perspectives.....	35
6	References.....	37

# 1 Introduction

Helical membrane proteins represent about 20-30% of all open reading frames (ORF) in sequenced genomes. To obtain the total number of membrane-associated proteins, one would add  $\beta$ -barrel proteins, proteins anchored by lipidic groups, and non-hydrophobic proteins that are bound in membrane complexes to this percentage as well. Thus, a large portion, perhaps even a majority of genes are related to membrane functions. In this review, we focus on the proteins having at least one putative transmembrane (TM) helix. In general, the TM regions comprise 18 or more amino acids with a largely hydrophobic character. These sequence features can be identified in primary sequences using hydrophobicity scales (Kyte & Doolittle, 1982; Steitz et al., 1982; Engelman et al., 1986; von Heijne, 1992; Wimley & White, 1996). Recent advances in the field of membrane protein assembly and structure have been reviewed (von Heijne, 1999). In the following, we discuss current advances in genomic, structural and functional aspects within the field of membrane proteins, largely from a computational point of view.

Membrane proteins are often found in oligomeric complexes, where they enable functions such as active transport, ion flows, energy transduction, and signal transduction. Many fundamental cellular processes involve protein-protein interactions, and membrane proteins are no exception. Comprehensively identifying complexes is important to systematically defining protein function (Eisenberg et al., 2000), (Lan et al., 2003), and hints about the function of an unknown protein can be obtained by investigating its interaction with other proteins of known function. Moreover, protein-protein interactions have obvious medical importance. Some forms of cancer, for instance, are associated with integral membrane protein-protein interactions, which lead

to aberrant downstream signal transductions important for cell growth regulation (Surti et al., 1998; DiMaio & Mattoon, 2001). Special attention has been drawn to G-protein coupled receptors (GPCR), because of their importance in therapeutic applications (see for example (Horn et al., 2003)). In the human genome approximately 2% (800) of the genes are GPCRs in which olfactory receptors constitute the largest gene family (Crasto et al., 2002).

Membrane proteins, nonetheless, pose something of a paradox. On the one hand, studying them is difficult experimentally. For instance, high-resolution structures of only about 70 membrane proteins exist compared with thousands of water soluble proteins ([http://blanco.biomol.uci.edu/Membrane\\_Proteins\\_xtal.html](http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html), (Berman et al., 2002)). These structures are highly dominated by  $\alpha$ -helical proteins and relatively fewer  $\beta$ -sheet structures. Also, high throughput techniques like the yeast 2 hybrid method, which identifies protein-protein interactions, cannot be easily applied to membrane proteins. On the other hand, from a computational standpoint, membrane proteins are actually easier to study than soluble ones. This is because their transmembrane regions have a much more limited diversity of potential structures, i.e. helical membrane proteins being mostly confined to parallel or anti-parallel orientations in relation to the membrane plane. In fact, accurate structural predictions of membrane helical proteins have been made in many cases (Adams et al., 1996; Pappu et al., 1999; Kim et al., 2003). Similarly, it may be easier to predict membrane protein interactions than soluble protein interactions. This is because membrane proteins interact in more restricted ways (i.e. cylinder to cylinder packing) than soluble ones, which can project a wide variety of different interfaces. The relation between the interacting sequence of membrane proteins and the type of

interaction present may be more direct because of the more restricted interface structures than for soluble proteins. Finally, there are fewer potential interactions for membrane proteins than for soluble proteins, since soluble proteins are relatively free to move within the cell and, therefore, have the ability to interact with many proteins at different times, angles, and locations. In contrast, a membrane protein's mobility is largely limited by the two dimensional constraint of the membrane and the number of nearest neighbors it can have is more limited.

## **2 Genomic classification and analysis of transmembrane sequences**

### ***2.1 Advances in prediction of helical membrane protein topologies***

While the need to solubilize them makes membrane proteins notoriously difficult experimental subjects for structural and biophysical studies, computational studies have been much more successful at predicting helical membrane protein topologies, i.e., identifying helical TM domains and predicting their in/out orientation relative to the membrane. The idea that it should be possible to estimate whether a polypeptide chain codes for a TM segment or not was formulated in the early 80's (Kyte & Doolittle, 1982; Steitz et al., 1982). It was based on the hypothesis that hydrophobic protein portions could form stable structures across the bilayer using hydrophobic helices (Engelman & Steitz, 1981). This structural arrangement would be stable if the gain in free energy

arising from burying hydrophobic residues into the bilayer exceeds the cost of burying charged and hydrogen-bonding groups. Although some aspects of the proposed insertions events have not been verified, notable results arose from the study. First free energy calculations of the insertion of  $\alpha$ -helices with a defined length of 21 amino acids were performed on bacteriorhodopsin and glycophorin A and accurately predicted the seven and one TM helices, respectively (Steitz et al., 1982; Engelman & Steitz, 1984).

In general, topology prediction algorithms make use of the observations that membrane helical proteins follow a special topology where TM helical segments are connected by alternating cytoplasmic (inside) and periplasmic (outside) loop segments, and that different amino acid distributions are associated with different segments. Such sequence patterns can be characterized by analyzing membrane protein sequences with experimentally determined topology, and they can in turn be used to predict TM regions and topologies for other proteins where only the primary sequence information is known. In particular, two general observations have been useful for predicting TM regions and their topologies.

- (i) Hydrophobic residues are enriched in TM helical segments where they traverse the hydrophobic region of a membrane. This observation forms the basis of hydrophobicity scanning algorithms for predicting TM regions. These algorithms use a sliding window scheme and calculate the mean residue hydrophobicity for each window (Kyte & Doolittle, 1982). Windows with the mean residue hydrophobicity above a certain threshold are candidates for TM regions. The window size is chosen to be consistent with the observed size of TM helical segments. Many different hydrophobicity scales have been

proposed (Kyte & Doolittle, 1982; Engelman et al., 1986; Wimley & White, 1996; Jayasinghe et al., 2001). In addition to computing the mean hydrophobicity of a window, a directional coefficient can be introduced to the averaging procedure, and this can be used to quantify the amphiphilic nature of helices (Eisenberg et al., 1984). Amino acid preferences in TM helical segments can also be inferred from scales that measure amino acid properties other than hydrophobicity (Deber et al., 2001; Zhou & Zhou, 2003), or estimated directly from a set of TM sequences with known topologies (Hofmann & Stoffel, 1993). These scales can be subsequently processed (Klein et al., 1985) and combined (Hirokawa et al., 1998; Juretic et al., 2002) to improve prediction results. Methods more sophisticated than simple window averaging have been proposed, such as neural networks (Rost et al., 1995) and wavelets (Lio & Vannucci, 2000). Prediction results can be improved by using a sequence profile instead of a single sequence as input. The sequence profile can be computed from multiple sequence alignments (Rost et al., 1995; Persson & Argos, 1996). Alternatively, a global profile can be constructed from pairwise alignments between the query sequence and all membrane protein sequences with known topology (Cserzo et al., 1997), which can in turn be used for predicting TM regions.

- (ii) Cytoplasmic segments often contain significantly more positive charges than periplasmic segments. This observation, termed the positive-inside rule, can be used to improve predictions (von Heijne, 1992). The cellular localization of N- and C- termini can be predicted and it gives an indication if the number of



TM segments is even or odd. Further, the positive-inside rule can help to decide whether an uncertain TM segment can be considered as “real”. Statistical analysis showed that the positive-inside rule is very likely to apply to most organisms from all three kingdoms (Wallin & von Heijne, 1998).

Improved prediction results can be achieved by simultaneously making use of the above two observations. This can be done in a straightforward way by first identifying putative TM helical regions using the sliding window approach, followed by quality checking using the positive-inside rule (Nakai & Kanehisa, 1992; von Heijne, 1992; Rost et al., 1996). Several prediction methods have been developed that fit membrane protein topological models to the entire query sequence and search for a grammatically correct topological model that best explains the given sequence. This can be done for example by using expectation maximization with dynamic programming (Jones et al., 1994) or with a hidden Markov model (HMM) (Tusnady & Simon, 1998; Krogh et al., 2001). One advantage of HMM is that length constraints on TM helical regions can be modeled in a consistent way together with hydrophobicity and charge bias.

Many of these prediction algorithms have been implemented as Web servers. A subset of these Web servers is listed in Table 1. This list is by no means complete; for a detailed survey of membrane protein topology prediction methods, see (Chen & Rost, 2002). Recently, several studies have been carried out to assess the accuracy of membrane protein topology prediction methods. In an analysis by Möller et al. (Moller et al., 2001), HMM-based methods such as TMHMM and HMMTOP performed the best. When tested on a dataset not used in training, the accuracy of the best algorithm was 85% for

predicting individual TM helical regions, and 59% for identifying all TM helical regions of a membrane protein correctly. However, the sidedness of TM helices is not well predicted: just 63% of these predictions predicted the sidedness of TM helices correctly. TMHMM is particularly good at distinguishing between membrane and soluble proteins. On the contrary, many hydrophobicity scanning algorithms cannot effectively discriminate TM helices from buried helices in soluble proteins. Many topology prediction algorithms tend to confuse signal peptides and transit peptides with TM helices, and it is recommended that these algorithms be used together with signal sequence prediction algorithms (Nielsen et al., 1999). In another analysis done by Ikeda et al. (Ikeda et al., 2002), model-based algorithms performed best. In a third analysis by Chen et al. (Chen et al., 2002), no method performed consistently as the best, but three methods stood out as more often better than worse: HMMTOP, PHDpsihtm (a version of PHDhtm based on PSI-BLAST profiles), and PHDhtm. The accuracy of the best method for identifying all TM helical regions of a membrane protein correctly is 84% for a high-resolution membrane protein dataset with known 3D structures, and 72% for a low-resolution membrane protein dataset. In addition, 66-85% of these predictions gave the sidedness of TM helices correctly. However, since the dataset used in training is not excluded from the test set, these numbers are likely to be an overestimation. Finally, in an analysis by Melén et al. (Melen et al., 2003), reliability scores were derived for five widely used membrane protein topology prediction methods, and TMHMM and MEMSAT were shown to have the best prediction characteristics in terms of prediction accuracy versus cumulative coverage of the test set. Furthermore, it was estimated that only 53-59% of all genome-wide membrane protein topology predictions are correct in

predicting both the number and the sidedness of TM segments. However, this number can be improved to ~70% if the in/out location of a protein's C-terminus is known from experiments.

It is apparent from the above analysis that further efforts are needed to improve current topology prediction methods. In addition, since different methods have different strengths and weaknesses, combining them and looking for a consensus prediction can often improve prediction results (Nilsson et al., 2000). TM helices of membrane proteins have been predicted with an accuracy greater than 99% based on the Wimley & White whole-residue hydropathy scale (Jayasinghe et al., 2001). The strength of this approach is that it also takes into account the cost of dehydrating the helix backbone and the energy for salt-bridge formation.

Using these computational methods, thousands of putative TM helical domains have been annotated in the SwissProt database. Further statistical analysis of these putative TM domain sequences has been very valuable for the identification of motifs that are important for the folding and function of membrane helical proteins.

### **2.1.1 Advances in 3D structure prediction of membrane helical proteins**

Significant progress has been made over the last several years on all fronts of protein structure prediction. The most dramatic example is the performance of *ab initio* structure prediction at CASP, a double-blind community-wide experiment on assessing structure prediction methods. In 1996, no group had sustained success in predicting generally correct structures over a range of targets (Lesk, 1997). Today, it is possible to construct

crude ( $\sim 5$  Å) models for diverse single domain proteins (Bonneau & Baker, 2001; Lesk et al., 2001; Keasar & Levitt, 2003). Progress is also evident at CAPRI, a community-wide experiment on assessing protein docking methods (Mendez et al., 2003). Methods tested at CASP and CAPRI are generally optimized for soluble proteins. However, they can also be modified for membrane protein 3D structure predictions.

3D structure prediction of membrane helical proteins may be simpler than that of soluble proteins for two reasons. First, helices are more stable in membrane environments than in aqueous solution, and the folding of membrane helical proteins can be approximated as the assembly of preformed TM helices. Second, the lipid bilayer environment imposes restrictions on the possible geometry of TM helix-helix packing. Since the location of TM helices in the primary sequence can be predicted reasonably well using topology prediction methods, recent efforts in membrane helical protein structure prediction have been focused on predicting the 3D assembly of TM helices.

The first step in membrane helical protein structure prediction is the development of an accurate energy function. Early methods model TM helix-helix association *in vacuo* using molecular mechanics force fields (Kerr et al., 1994; Adams et al., 1995; Adams et al., 1996). Predictions made by these methods are in good agreement with experiments despite the fact that protein-lipid interaction is not modeled. Furthermore, in some cases reasonable structural models can be generated by optimizing interhelical van der Waals interactions only (Pappu et al., 1999; Kim et al., 2003). These studies highlight the importance of TM helix packing in membrane helical protein folding. Recently, implicit solvent models have been introduced for efficient treatment of the membrane environment (Im et al., 2003; Lazaridis, 2003). In addition to physical potentials, other

forms of energy functions have also been developed, including a simple scoring function based on qualitative insights into TM helix interaction (Fleishman & Ben-Tal, 2002), and knowledge-based energy functions based on statistical analyses of membrane proteins sequences and structures (Pilpel et al., 1999; Adamian & Liang, 2001; Dobbs et al., 2002).

The second step in membrane protein structure prediction is the development of effective sampling methods that can generate low energy, native-like conformations of TM helix-helix interactions. In some methods, the interaction energy of preformed helices is optimized by restrained molecular dynamics and simulated annealing (Adams et al., 1995; Adams et al., 1996), potential smoothing (Pappu et al., 1999), or Monte Carlo minimization (Kim et al., 2003). In addition, membrane protein structures can be assembled from structural fragments using a simulated annealing protocol (Pellegrini-Calace et al., 2003). In other methods, solved structures for membrane proteins can serve as homology modeling templates for close homologs (Capener et al., 2000), and as fold recognition templates for detecting and aligning remote homologs (Bowie et al., 1991; Jones et al., 1992; Dastmalchi et al., 2001). The power of fold recognition can be augmented by computationally generating a representative set of plausible membrane helical protein folds (Bowie, 1999). This set of new folds can then be added to the fold library, and fold recognition methods can be used to predict if a protein sequence adopts a fold in the library.

These predictions can be improved in several ways. First, experimental or phylogenetic information can be incorporated (Adams et al., 1996; Pinto et al., 1997). Second, low energy conformations can be clustered and the representative conformation from the

largest cluster tends to be more native-like (Kim et al., 2003). Third, prediction results can be improved by looking for consensus predictions for homologous proteins (Briggs et al., 2001) or by combining different constraints derived from homologous proteins (Pogozheva et al., 1997).

Impressive 3D structure prediction results have been reported for individual cases of membrane helical proteins. For example, using a simple physical energy function, Pappu et al. constructed an *ab initio* model for the glycoporphin A TM dimer that is very close to the experimental NMR structure (root mean square deviation for superposition over all C<sub>α</sub> atoms is 0.59 Å for 36 residues) (Pappu et al., 1999). Unfortunately, current 3D structure prediction methods are complex and time consuming, and a quantitative comparison of different methods has not been carried out at this time. Despite the recent progress in membrane helical protein 3D structure prediction, it is clear that major efforts are needed to make these methods reliable and fast before they can be applied on a genomic scale.

## **2.2 Genome wide classification of membrane proteins**

Genome-wide analysis of protein structures provides a powerful method for understanding functional and evolutionary relationships in proteins. However, this kind of analysis has mostly been applied to soluble proteins, in part due to the paucity of structures of membrane proteins (Gerstein, 1997; Gerstein, 1998; Paulsen et al., 1998; Paulsen et al., 2000). However, a number of efforts has been made to use computational tools despite the absence of a large structural database, taking advantage of the relatively

simple architecture found in the TM region. Therefore, it seems timely to consider computational methods as alternatives for the analysis of TM helical regions in membrane proteins.

The occurrence of helical membrane proteins in genome sequences has been surveyed in several organisms (Goffeau et al., 1993; Rost, 1996; Arkin et al., 1997; Gerstein, 1997; Boyd et al., 1998; Gerstein, 1998; Jones, 1998; Wallin & von Heijne, 1998; Krogh et al., 2001). In general, the overall number of membrane proteins found depends on the prediction method used, but most studies report values in a range of 20-30% of the open reading frames in microbial genomes, with yeast having a slightly larger fraction. There is a progression in the number of occurrences from single helix proteins, which are most abundant, in a generally monotone decreasing fashion with number of TMs. However, there are some notable departures from the trend. An analysis of the worm genome (Gerstein et al., 2000) showed a much greater relative prevalence of 7-TM proteins in comparison to the other completely sequenced genomes, which are not of metazoans. In contrast, *E. coli* has a preference for 6 and 12 TM proteins.

Polytopic membrane proteins have multiple membrane spanning TM segments. Based on Pfam classification of protein domain families, TM prediction and sequence similarity these polytopic membrane protein domains have been classified further (Liu et al., 2002; Liu et al., 2004). Some interesting trends have been identified, such as

- (i) That there is an approximately linear relationship between the number of classified membrane protein domains and the number of ORFs, and
- (ii) That the majority of integral membrane proteins have only a single polytopic membrane domain. About 78% of integral membrane proteins in archaea and prokarya

and 67% in eukarya contain only a single classified membrane domain (Liu et al., 2004), suggesting recombination of domains is not common inside membranes. Distinct from soluble proteins, which gain new functions by recombination of different domains in the course of evolution (about 65% domains in prokarya and 80% in eukarya are combined with other domains), membrane proteins might achieve the same goal by more frequent use of non-covalent oligomeric associations within the membrane.

(iii) That the number of families of polytopic membrane proteins is small compared with the number of soluble protein families, i.e. 526 membrane protein families have been characterized (Liu et al., 2002; Liu et al., 2004) which corresponds to about 9% of the existing Pfam families (Bateman et al., 2002).

### ***2.3 Integrative database systems***

Data arising from the above mentioned studies are partly available through two interlinked and integrated database systems, PartList.org (Qian et al., 2001) and GeneCensus.org (Lin et al., 2002). GeneCensus.org also contains an integrated viewer of TM helix motifs and the expression levels of all membrane proteins in sequenced genomes. In general, GeneCensus takes a more sequence and less structural view of genome comparisons than PartsList, focusing on expression data, pathway activities, and protein interactions.

These integrated database systems have been used to discover a number of interesting correlations related to membrane proteins. The prediction of TM helices in yeast has been connected with a number of datasets giving measurements of whole genome expression



levels (Jansen & Gerstein, 2000). ORFs coding for membrane proteins were identified using the standard hydropathy scale and sliding window approach. This produced the notable result that membrane proteins are expressed at a considerably lower level than soluble proteins, by ~22%. Moreover, certain broad groups of membrane proteins are expressed more highly than others, e.g. 4-TMs are expressed at a higher level than 1 or 2-TMs. In a second step this analysis has been extended to fully relate subcellular localization (i.e. ER, cytoplasm, membrane, etc.) with gene expression level. A relationship between gene expression levels and subcellular localization was found indicating that cytoplasmic proteins have high expression levels (absolute expression =14.4) whereas nuclear (1.7) and membrane proteins (2.4) have relatively low ones (Drawid et al., 2000). In a new strategy, the localization of proteins in yeast has been greatly enhanced (Huh et al., 2003). Proteins are fused to the green fluorescent protein and their localization is determined by fluorescence microscopy. Although the detection of the subcellular localization is limited by the resolution of the microscopy, the advantage of this method is that protein expression is minimally perturbed. Thus, 70% of previously unlocalized proteins have been assigned to compartments.

## ***2.4 Co-evolutionary analysis of membrane proteins***

Using current concepts of protein evolution helps in understanding both the structural and the functional aspects of protein families. Divergent evolution relates to the evolution of organisms, which are linked to a common ancestor, into new species (Ohno, 1970; Zuckerkandl, 1975; Hood et al., 1977; Doolittle & Feng, 1990; Li, 1991). Many studies

have incorporated evolutionary information in order to identify functionally important residues that confer binding specificity (Casari et al., 1995; Lichtarge et al., 1996b; Lichtarge et al., 1996a; Lichtarge et al., 1997; Pazos et al., 1997; Landgraf et al., 2001). Additionally, other studies show that correlated mutation information can be used to predict proximal pairs of residues (Gobel et al., 1994; Olmea & Valencia, 1997) and to aid in structure prediction (Olmea et al., 1999; Ortiz et al., 1999). Co-evolutionary analysis of protein families has also been useful in identifying protein interaction partners.

It is generally believed that the functional diversification of genes within a gene family should be reflected in their interacting partners in another gene family (Fryxell, 1996; Pazos et al., 1997; Goh et al., 2000; Pazos & Valencia, 2001; Goh & Cohen, 2002; Pazos & Valencia, 2002). Studies of the co-evolution of binding specificity between homologous ligands and receptors (Moyle et al., 1994; Atwell et al., 1997; Jespers et al., 1999) show that protein-protein interfaces can adapt to mutations as they co-evolve and new interactions can be formed. Based on this hypothesis, co-evolution has been quantified between gene families that are known to interact (Goh et al., 2000). The co-evolutionary score is quantified by calculating the linear correlation coefficient between the sequence similarity matrices constructed from the multiple alignments of the two gene families. This method is described in further detail by Goh *et al* (Goh *et al.*, 2000). Using this co-evolutionary algorithm, binding partners were identified for proteins with previously unknown interaction partners (Goh & Cohen, 2002). Pazos *et al.* (Pazos & Valencia, 2001; Pazos & Valencia, 2002) extended this idea by applying it to large sets of proteins and protein domains to identify pairs of interacting proteins.

We have chosen one example of a membrane protein, the photosynthetic reaction center, to illustrate the usefulness of the co-evolutionary method. Since co-evolutionary analysis does not require structural information, it can be readily applied to study the structure and function of membrane proteins. The photosynthetic reaction center (RC) complex in purple bacteria is composed of subunits L, M, H, and in some species, a cytochrome subunit (Thornber et al., 1980; Michel et al., 1985; Michel et al., 1986; Weyer et al., 1987; Nagashima et al., 1994). The RC from *Rhodospseudomonas viridis* was the first integral membrane protein complex where well ordered three-dimensional crystals were obtained for X-ray structure analysis (Michel, 1982; Michel, 1983). Since then, only one other photosynthetic reaction center has been structurally determined (Allen et al., 1987; Chang et al., 1991), which is found in *Rhodobacter sphaeroides*. The L and M subunits form the central part of the RC. The L-M complex forms a flat surface parallel to the membrane surface where the cytochrome subunit binds at the periplasmic side and the H subunit at the cytoplasmic side of the membrane.

Figure 1 shows how surfaces that have a greater interfacial contact area have a correspondingly higher co-evolution correlation score. For example, the L and M subunits have a large interfacial area and co-evolve with a correlation score of 0.94, whereas the much smaller interface between the H and other cytochrome subunits results in a correlation score of 0.43. Figure 1 shows that there is a general correlation between intersubunit surface area and the score obtained from the co-evolutionary analysis. These results demonstrate the utility of applying co-evolutionary analyses to characterize the domain-domain interactions in membrane proteins.

## **2.5 Membrane proteins and pseudogenes**

Pseudogenes are disabled copies of functional genes in the genome; these sequences have close similarities to one or more paralogous functional genes, but in general are unable to be transcribed (Vanin, 1985; Mighell et al., 2000). There are three major groups of pseudogenes, having different origins:

- (i) Duplicated pseudogenes, created by gene duplications
- (ii) Processed pseudogenes, created by reverse-transcription of mRNA transcripts and
- (iii) Disabled genes, created by spontaneous loss of function.

Complete genome sequences have recently become available for many prokaryotes and eukaryotes including two mammals (International Human Genome Sequencing Consortium, 2001; Waterston et al., 2002), large-scale computational surveys have been performed on these genomes to identify and characterize potential pseudogenes, which also revealed many pseudogenes that used to code for membrane proteins (Cole et al., 2001; Glusman et al., 2001; Harrison et al., 2001; Harrison et al., 2002; Homma et al., 2002; Zhang et al., 2002; Harrison et al., 2003). The largest protein family in the nematode worm *C. elegans* is the 7-TM receptor, which has ~ 800 members (The *C. elegans* Sequencing Consortium, 1998). The *C. elegans* genome has approximately ~ 2100 pseudogenes, about one for every eight functional genes (Harrison et al., 2001). A substantial proportion (22%) of these pseudogenes initially coded for membrane proteins, especially 7-TM receptors. Substantial numbers of membrane protein pseudogenes are

also present in the genomes of some other eukaryotes such as the fruit fly (Harrison et al., 2003) and yeast (Harrison et al., 2002).

The human genome has about 30,000 functional genes and olfactory receptors (OR) constitute one of the largest gene super-families (International Human Genome Sequencing Consortium, 2001). Among the ~ 700 full-length OR genes identified in the genome, more than half of the sequences (359) contain frame disruptions (stop codons, frame shifts), indicating that these are pseudogenes (Glusman et al., 2001). Most of these OR genes and pseudogenes are located in gene clusters that range from 100 kb to 1Mb. The majority of the OR pseudogenes became disabled following a random and spontaneous process (Glusman et al., 2001).

A recent whole-genome survey has identified more than 10,000 pseudogenes in the human genome and about 5,000 pseudogenes in the mouse genome (Zhang et al.). Many membrane protein pseudogenes are also present in the mammalian genomes. A good example is cytochrome *b* (*cytb*), which is a ubiquitous 8-TM protein that catalyzes a crucial step in the mitochondrial oxidative phosphorylation process (Zhang et al., 1998; Zhang et al., 2000). The functional gene of this protein is in the mitochondrial genome, but more than 70 copies of its *cytb* pseudogenes are present in the nuclear genome due to a DNA-mediated process (Tourmen et al., 2002; Woischnik & Moraes, 2002). Cytochrome *c* (*cyc*), another important protein in the mitochondrial electron-transfer chain that interacts with cytochrome *b*, also has 49 pseudogenes in the human genome (Zhang & Gerstein, 2003).

The omnipresence of these pseudogenes has allowed a tracing of the evolution and phylogeny of membrane proteins. However, because of their close sequence similarities

to the functional genes, they also pose potential problems in the experimental studies of the functional membrane protein genes (Ruud et al., 1999).

### **3 Structural characteristics of membrane proteins**

The thermodynamics of membrane protein stability suggest that a division can be made between those factors that stabilize helices in a lipid environment and those that cause them to interact to form higher order structure (for review see (Popot & Engelman, 2000)). A proposal was made more than a decade ago that TMs might be independently stable across a bilayer, in response to a net hydrophobicity of the side chains and the influence of backbone hydrogen bonding in a low dielectric milieu (Engelman & Steitz, 1981; Popot & Engelman, 1990). Helices could then interact with each other to form higher order structures. These two thermodynamic stages might be a pathway for folding and oligomerization in vivo. Recently, this two stage model has been extended to a three stage model (Engelman et al., 2003), where ligand binding, and the re-entry of extramembranous loops would follow the assembling of TMs. The formation of oligomeric quaternary structure could take place at the transitions between the first and second stage, between the second and third, or between later stages.

#### **3.1 *Amino acid composition of membrane proteins***

Polytopic membrane protein domains have been classified on the basis of sequence similarities and topology using existing families assigned by a combination of a HMM

and sequence analysis of Pfam (Liu et al., 2002; Liu et al., 2004). Some amino acids, such as glycine, proline, and tyrosine, are found to be more frequent in conserved positions in TM regions than it is expected from their composition, whereas isoleucine, valine, and methionine are less conserved (Fig.2).

Based on the SwissProt database, the TM distributions of single types of amino acids and pairwise correlated amino acids in TM domains have been investigated further. A tendency for Cys, Tyr and Trp residues to appear close to one another has been pointed out (Arkin & Brunger, 1998).

Particular interest has been drawn to the occurrence of glycine pairs in  $\alpha$ -helical membrane proteins. The GxxxG motif, two glycines at position  $i$  and  $i+4$ , is known to be a key structural element in the dimeric association interface of glycophorin TMs (MacKenzie et al., 1997), to date the best characterized example of TM helix interaction. A separation of four residues places a pair on the same face of an  $\alpha$ -helix. The GxxxG motif led to the structural notion that small residues presented on the same face of the helix, and next to larger side chains, can increase the relatively small packing area that two helices can present to each other. The same motif was found to be a strong determinant for association in genetic screens that selected for strong helix association (Russ & Engelman, 1999; Russ & Engelman, 2000).

An independent study using statistical analysis has shown the abundance of the GxxxG motif in the domain of membrane proteins. Using the large number of TM domains annotated in SwissProt, Arkin and Brunger (1998) found a sharp peak of Gly pairs at a distance of 4 residues (GxxxG). This study was extended further on a later SwissProt version (v.37) (Senes et al., 2000), where the occurrence of residue pairs and triplets in

TM helices has been surveyed using the TMSTAT method. This method gives a parent distribution, and permits an evaluation of the significance of observed pair frequencies with respect to the distribution ranking over- and under-represented pairs by the significance difference from their expectations. The GxxxG motif was the most significant over-represented pair, with 32% more occurrences observed in the sequence than their random expectation ( $p=10^{-33}$ , (Senes et al., 2000)). Moreover, all other combinations of two small residues (Gly, Ala and Ser) at  $i$ ,  $i+4$  resulted significantly over-represented. Pairs of two large residues (Ile, Val, Leu) tend also to be spaced four residues apart while large and small residues are more frequent at  $i+1$  and  $i+2$ , with a strong correlation for GxxxG motifs and a neighboring  $\beta$ -branched Ile and Val residues made apparent from the triplet data. Further, it has been pointed out that these motifs are well conserved in families annotated as transporter, symporter and channels (Liu et al., 2002).

The combination of experimental and statistical analysis studies clearly establishes that there is a high selectivity in the use of particular relationships between amino acids along a TM helix, and suggests that further studies of motifs are likely to be both informative and predictive.

### ***3.2 Helix packing and characterization of helical interfaces***

Much attention has been drawn to structural characteristics of  $\alpha$ -helical membrane proteins. Characteristics such as protein packing, packing density and residue volumes



are important criteria in the context of helix-helix interactions, protein stability and function.

Protein packing calculations that measure the volume occupied by protein constituents (packing efficiency) were developed for soluble proteins a long time ago (Richards, 1974; Richards, 1985), but have only recently been applied to helix interfaces (Eilers et al., 2002). An example of calculations of the packing in membrane proteins (Gerstein & Chothia, 1999) is shown in table 2 and further information is available at <http://bioinfo.mbb.yale.edu/geometry/membrane>.

Table 2 shows examples of volumes of buried atoms in various membrane protein structures and compares it to a standard reference volume of the same atoms in a soluble protein structure. A clear tendency toward tighter packing in membrane proteins compared to soluble ones can be deduced from this study. In this study, the protein volume is calculated by surrounding each atom with a Voronoi polyhedron. The faces of the Voronoi polyhedron are perpendicular to vectors connecting the centers of different atoms, and the edges of the polyhedron result from the intersection of these planes (for detailed description see (Gerstein & Richards, 2001)). This method relies on several parameters such as the set of used van der Waals radii and the criteria for selecting buried atoms in the calculation. Therefore the sensitivity of the methodology of packing calculations has been investigated further and led to the development of a new set of parameters concerning the van der Waals radii and standard volumes (Tsai et al., 2001). The results are available at <http://www.molmovdb.org/geometry/>, (Tsai & Gerstein, 2002).

A comparison of packing of helical segments in membrane proteins and soluble proteins confirmed that on average membrane proteins pack more tightly (packing value of 0.431) compared to their soluble counterparts (0.405) despite the fact that TM proteins cannot make use of hydrophobic effects in folding within the bilayer (Eilers et al., 2000). An interesting finding in these studies is that, on average, smaller residues pack tighter and occur more often in TM proteins whereas larger residues tend to pack more tightly and occur more frequently in soluble proteins (Eilers et al., 2000). This is consistent with the finding that small sidechains participate in packing motifs, as noted above. An additional distinction is that proline occurs frequently in TMs, and is found preferentially in the center of the membrane, whereas prolines are quite rare within helices of soluble proteins (Cordes et al., 2002).

### ***3.3 Analysis of helix-helix interfaces in transmembrane proteins***

Helix-helix interfaces are described by characteristics, such as helical packing angles and interfacial motifs. Inter-helical packing angles are typically defined as the dihedral angle measured around the helices' mutually perpendicular vector of closest approach (Bowie, 1997a; Walther et al., 1998). A major aim in studying packing angles is to define categories in which to bin helical interactions. Early studies with this purpose predicted three preferred packing angles ( $-52^{\circ} - 37^{\circ}$ ,  $75^{\circ} - 83^{\circ}$ , and  $22^{\circ} - 23^{\circ}$ ) (Chothia et al., 1981; Walther et al., 1996) and a companion survey reported a preference for the  $-37^{\circ}$  angle (Walther et al., 1996), although this has been challenged on statistical grounds (Bowie, 1997a). If TM helix pairs are studied separately, angle preferences become more

prevalent. A survey of 88 TM interfaces identified packing angles as low as  $-56^\circ$  and as high as  $67^\circ$  with a strong preference for left-handed crossing angles in the  $15^\circ$ - $25^\circ$  range (Bowie, 1997b). In contrast, 30% of the 2145 soluble helix-packing angles studied at the time fall outside of this range and have a much broader distribution (Bowie, 1997b). In a study separating parallel and anti-parallel interactions it was observed that the bias for left-handed crossing interactions was mostly due to the anti-parallel component (Senes et al., 2001). Such packing angle constraints in TM helix packing can potentially aid in the development of membrane protein folding algorithms, as this greatly reduces potential search spaces. Tools for calculating helix-helix packing angles are described in (Bansal et al., 2000) and in (Dalton et al., 2003).

Differences have been studied between residues participating in inter-helical contacts and those that do not, inter-helical residues in helices having right- or left-handed packing angles, and in inter-helical residues within parallel and anti-parallel orientations (Eilers et al., 2002). Analysis of the available structures indicates that Gly residues tend to be found at packing interfaces (Javadpour et al., 1999; Eilers et al., 2002), permitting close approaches of the backbones, and formation of interhelical networks of weak hydrogen bonds between  $C\alpha$ -H donors and oxygen acceptors, an interaction that has been hypothesized to be quite favorable in an apolar membrane environment (Senes et al., 2001). Since Gly residues are strong helix breakers in solution, it was somewhat surprising to find that the GxxxG motif can mediate interactions at the interface of soluble dimers, with a similar geometry to the right-handed glycophorin motifs and formation of  $C\alpha$  hydrogen bonds (Kleiger et al., 2001; Kleiger et al., 2002). Moreover, helices with left-handed crossing angles are often more tightly packed (packing value

0.518) than helices with right-handed crossing angles (0.508) (Eilers et al., 2002). It has also been pointed out that larger residues such as Phe, Tryp and His have a higher propensity for appearing in TM voids and pockets while smaller residues (Ser, Gly, Ala) do not. These studies have been extended to amino acid triplet motifs, which could be involved in the formation of interhelical interactions (Adamian et al., 2003) and it has been pointed out that the pair motifs such GG4 can be a part of these triplets.

## **4 Membrane protein interactions**

Protein-protein interactions play a role in nearly all events that take place in a cell. The set of all such interactions carried out by proteins encoded in a genome has been dubbed the *interactome*. An important idea emerging in post-genomic biology is that the cell can be understood as a complex network of interacting proteins (Hartwell et al., 1999; Eisenberg et al., 2000). Complex networks have also been used elsewhere to describe such diverse systems as the internet, power grids, the ecological food web and scientific collaborations. Despite the seemingly huge differences among these systems, it has been shown that they all share similar network topology (Watts & Strogatz, 1998; Albert et al., 1999; Barabasi & Albert, 1999; Huberman & Adamic, 1999; Albert et al., 2000; Amaral et al., 2000; Albert & Barabasi, 2001; Jeong et al., 2001; Girvan & Newman, 2002).

However, defining protein interactions, which involve membrane proteins, presents many challenges, such as the low abundance of membrane proteins and the difficulty of detecting interacting partners.

## **4.1 The current excitement about protein networks**

A great variety of genome-wide information related to protein networks has been accumulated in recent work, especially in the yeast *Saccharomyces cerevisiae*. There are datasets of explicit protein-protein interactions (Ito et al., 2000; Uetz et al., 2000; Gavin et al., 2002; Ho et al., 2002) and also of experimentally derived regulatory relationships (Lee et al., 2002). Furthermore, there are databases collecting a wide variety of manually curated interactions from individual experiments (i.e. MIPS, BIND, and DIP (Mewes et al., 2002; Xenarios et al., 2002; Bader et al., 2003)) and systems for automatically finding interactions in the literature (Friedman et al., 2001). In addition to the experimentally-derived interaction networks, there are also predicted interactions (Jansen et al., 2002; Valencia & Pazos, 2002; Jansen et al., 2003).

Protein-protein interaction networks are often globally characterized by a number of parameters from graph theory, such as degree distribution, clustering coefficient, characteristic path length and diameter (Watts & Strogatz, 1998; Albert & Barabasi, 2001; Jeong et al., 2001). Furthermore, these networks are undirected networks. Within undirected networks, the statement “node A is connected to node B” is the same as “node B is connected to node A”. Protein networks are quite complex and can often be divided into many quite substantial sub-networks.

The most common methods are based on “guilt-by-association”. Two proteins are more likely to interact if they share several correlated genomic features. Examples of these genomic features are gene expression profiles (DeRisi et al., 1997), phylogenetic profiles (Pellegrini et al., 1999), essentiality (Winzeler et al., 1999), localisation (Kumar et al.,

2002), and gene neighborhood (Tamames et al., 1997), among others. In addition, comparative genomics provides an efficient way to map genome-wide interaction datasets between different organisms (Walhout et al., 2000).

This body of work has resulted in the identification of many types of possible networks and sub-networks. For example, it has been known that interaction data produced by different methods are of different qualities. The topology of the interaction network determined by yeast two-hybrid experiments is quite different from that determined by *in vivo* pull-down experiments (Jansen et al., 2002; von Mering et al., 2002), probably reflecting the different selection principles involved. Proteins can be divided into different classes based on their biological properties, such as expression level, amino acid composition, subcellular localization, solubility, and so on. Therefore, different sub-networks can be generated by selecting different classes or groups of protein nodes. For instance, membrane proteins can be subdivided by the number of TM helices. A challenging research question is to compare the topologies of these sub-networks, looking for global differences in the networks associated with different types of proteins. TopNet (Yu et al., 2003) is an automated web tool designed to calculate and compare topological parameters for different sub-networks derived from any given protein network. The number of interaction partners for soluble proteins and membrane proteins within the interaction network has been examined. In general, soluble proteins have many more interaction partners than membrane proteins. Interestingly, the number of interaction partners for membrane proteins does not seem to have any correlation with the number of TM helices that they have.

## ***4.2 Identification of protein complexes with experimental techniques***

### ***4.2.1 Screening methods***

Several approaches to the study of interactomes have emerged recently. Using an adaptation of a “two-hybrid” assay (Ito et al., 2000; Uetz et al., 2000) pairwise interactions were mapped on a large scale in yeast. Microarray technology has also been used to study interactions (Zhu et al., 2001) and the idea of using proteins carrying a tag that can be separated on an affinity column has been developed as a screen (Gavin et al., 2002; Ho et al., 2002). Tagged proteins, bound to a column or bead and bringing with them associated proteins, are analyzed by electrophoresis, mass spectrometry, and bioinformatics to give the identity of proteins in the complex.

Many soluble protein complexes have been identified using these approaches, although problems with false positives and negatives persist. These are likely to arise from failures to control the biochemistry, for example two-hybrid screens require artificially elevated concentrations and exploit binding events that promote interactions, and column separations are at high effective dilution.

Membrane proteins remain to be explored in any systematic way, and many of the experimental techniques for directly assaying protein-protein interactions that have been applied on a genomic scale are thought to be biased against membrane proteins. For instance, the yeast two-hybrid system (Fields & Song, 1989) is difficult for integral membrane proteins, because the interaction must take place in the cell nucleus, as the reassembled functional transcription factor becomes bound to its target promoter for the

activation of the corresponding reporter gene in a consecutive step. However, integral membrane proteins are anchored in the membrane and cannot be transported into the nucleus. Related considerations apply for other methodologies such as the proteome chip (Zhu et al., 2001) and large-scale pull-down experiments (Gavin et al. 2002; Ho et al. 2002).

One way to circumvent the problems related to membrane proteins is to express a truncated form of the membrane protein. The use of only the cytoplasmic or extracellular domain is a strategy, which has been applied to single pass TM domains (Ozenberger & Young, 1995; Keegan & Cooper, 1996; Borg et al., 2000). However, this strategy is not suitable for multipass TM domains with binding interfaces composed of several cytoplasmic loops, or for the detection of interactions inside the membrane. The need for the development of new approaches for detecting membrane protein interactions is necessary.

Several systems, which are mainly variations of the two-hybrid method, have been set up. The Ras recruitment system (RRS) and the reversed Ras recruitment system are based on the Ras pathway in yeast (Broder et al., 1998; Hubsman et al., 2001). It allows the study of protein interactions between a membrane and a cytoplasmic protein.

Another approach is based on the characteristics of ubiquitin-specific proteases. Ubiquitin functions as a tag for protein degradation and the split-ubiquitin system takes advantage of the specific ubiquitin cleavage (Johnsson & Varshavsky, 1994). The advantage of the split-ubiquitin system is that it can detect protein-protein interactions in various cell locations and is applicable to nuclear, cytoplasmic and integral membrane proteins. Different reporters (rURA3 and trans-activator) have been attached to this



system for investigating integral membrane protein interactions (Dunnwald et al., 1999; Laser et al., 2000).

The G protein based screening system is based on the G-protein signaling process (Ehrhard et al., 2000). Here, the bait X is an integral membrane protein and its interaction partner Y (a soluble protein) is expressed as a fusion to the G $\gamma$  subunit. If X and Y interact, G $\gamma$  recruits to the membrane and binds the G $\beta$  subunit. In the following, the G-protein signaling is blocked. Interaction between two known interaction partners syntaxin 1 and neuronal Sec1 and the fibroblast-derived growth factor receptor 3 with SNT-1 have been demonstrated by this method.

#### **4.2.2 *Helix-helix interaction motifs***

Several assays have been developed for biophysical and genetic studies of membrane protein interactions. Characteristics such as the oligomeric state of TM helices, interaction motifs and energetic considerations about the helix association process have been investigated. Widely used methods include SDS gels (Lemmon et al., 1992), Förster resonance energy transfer (Fisher et al., 1999) and analytical ultra centrifugation (Fleming et al., 1997). These biochemical assays use pure systems and generally exploit detergent solubilized states. They have the advantage of permitting detailed analysis of the chemical interactions and energies, as well as defining the oligomeric state of the proteins.

Genetic assays have the advantage of permitting the observation of interactions in a natural membrane and can permit genetic screening and selection procedures, however

they report less clearly on stoichiometry and energy. They have been developed for establishing helix-helix interaction between specific TM sequences (Langosch et al., 1996; Russ & Engelman, 1999; Schneider & Engelman, 2003). To date, they are limited to homo- and hetero-oligomerization of parallel helices, and thus cannot serve adequately to survey all possibilities found in membrane helix associations. These techniques led to significant results in identifying interaction motifs as detailed in previous chapters. For example a milestone in the application of the TOXCAT assay for homo oligomerization is the identification of interaction motifs in the glycoporphin TM segment (Russ & Engelman, 1999).

### **4.3 How many helix-helix interactions exist in a genome?**

With the emergence of whole genome sequences and the annotation of potential TM segments in the sequences, we can speculate on the number of potential protein-protein and helix-helix interactions of membrane proteins in genomes. In the following, we will describe a rough estimation of the number of potential interactions in membrane proteins. We compared TM sequences from three different organisms with each other: (i) *M. genitalium* (MG) (Fraser et al., 1995), (ii) *E. coli* (EC) (Blattner et al., 1996) and (iii) *S. cerevisiae* (SC) (Goffeau et al., 1996). The numbers of membrane proteins, TM-helices, and potential helix-helix interaction pairs are shown in Fig. 3 and Table 3. In this distribution we did not take into account any mobility and geometrical aspects which have been discussed in the introduction. The numbers are given for the individual organisms as well as for orthologous membrane proteins that are present in all three

genomes (i.e. EC-SC-MG) or just in two out of the three (e.g. EC-SC). The orthologous proteins across the organisms have been assigned to using the database of Clusters of Orthologous Groups of proteins (COGs) (Tatusov et al., 2003).

The distribution of TM helices is shown in Figure 3 for the different groups of orthologous proteins. The figure illustrates how all of the possible subsets are fairly consistent in their distribution of membrane proteins.

Table 3 estimates roughly the possible number of helix-helix interactions involving only membrane proteins. These numbers (Table 3, row E-F) correspond to the upper limit for potential helix-helix interaction pairs. In the membrane each helix can only have a small number of interaction partners, because of the structural arrangement of the protein in the membrane. If one focuses on orthologs present in the different organisms with a known function the number of potential helix pairs shrinks down to a quite manageable size (~10000 pairs). In particular, the “virtual organism EC-SC-MG” could be used as a starting point to study helix-helix interaction further, using both computational and experimental methods.

## **5 Perspectives**

The past few years have produced a steep increase in our knowledge of membrane protein occurrence, structure and interactions. The number of high-resolution structures of membrane proteins has increased tremendously from the late 90’s onward. This, in turn has stimulated discussion about structural characteristics of TM segments and has led to a number of useful models and the subsequent development of tools dealing with

the “look” of a typical helix and its particularities. Currently, projects combining crystallographic and NMR techniques and innovative bioinformatics, are underway to increase the number of known 3D structures of integral membrane proteins. Thus, an important task for bioinformatics will be, for example, to provide tools such as prediction methods for finding the most appropriate crystallization and structure determination methods.

Although hydrophobicity scales and topology prediction tools for TM sequences go back as far as the early 80's, improved tools have been developed and refined subsequently. The current deluge of available sequence data has added an incentive for method development.

What might be expected next? A systematic study of protein-protein interactions on a genomic scale needs to be developed. Hopefully, with the current advances in the modifications of the two hybrid systems, a method will become available to study membrane protein interactions. This would open a new area of understanding protein networks and interactions, stimulating current discussions, for example, about interaction motifs in membrane proteins.

### **Acknowledgement**

DME and MG thank the NIH for support (P01 GM54160). UL thanks the German Academic Exchange Service (DAAD) for a Postdoctoral fellowship.

## 6 References

- ADAMIAN, L., JACKUPS, R., JR., BINKOWSKI, T. A. & LIANG, J. (2003). Higher-order interhelical spatial interactions in membrane proteins. *Journal of Molecular Biology* **327**, 251-272.
- ADAMIAN, L. & LIANG, J. (2001). Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J Mol Biol* **311**, 891-907.
- ADAMS, P. D., ARKIN, I. T., ENGELMAN, D. M. & BRUNGER, A. T. (1995). Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nat Struct Biol* **2**, 154-162.
- ADAMS, P. D., ENGELMAN, D. M. & BRUNGER, A. T. (1996). Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching. *Proteins* **26**, 257-261.
- ALBERT, R. & BARABASI, A. L. (2001). Statistical Mechanics of Complex Networks. *arXiv:cond-mat/0106096*, 1-53.
- ALBERT, R., JEONG, H. & BARABASI, A. L. (1999). Diameter of the World-Wide Web. *Nature* **401**, 130-131.
- ALBERT, R., JEONG, H. & BARABASI, A. L. (2000). Error and attack tolerance of complex networks. *Nature* **406**, 378-382.
- ALLEN, J., FEHER, G., YEATES, T., KOMIYA, H. & REES, D. (1987). Structure of the reaction center from Rhodobacter sphaeroides R-26: the protein subunits. *Proc Natl Acad Sci U S A* **84**, 6162-6166.
- AMARAL, L. A., SCALA, A., BARTHELEMY, M. & STANLEY, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 11149-11152.
- ARKIN, I. T., BRUNGER, A. T. & ENGELMAN, D. M. (1997). Are there dominant membrane protein families with a given number of helices? *Proteins* **28**, 465-466.
- ARKIN, L. & BRUNGER, A. T. (1998). *Biochim. Biophys. Acta* **1429**, 113-128.
- ATWELL, S., ULTSCH, M., DE VOS, A. M. & WELLS, J. A. (1997). Structural plasticity in a remodeled protein-protein interface. *Science* **278**, 1125-1128.
- BADER, G. D., BETEL, D. & HOGUE, C. W. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research* **31**, 248-250.
- BANSAL, M., KUMAR, S. & VELAVAN, R. (2000). HELANAL: a program to characterize helix geometry in proteins. *Journal of Biomolecular Structure & Dynamics* **17**, 811-819.
- BARABASI, A. L. & ALBERT, R. (1999). Emergence of Scaling in Random Networks. *Science* **286**, 509-512.
- BATEMAN, A., BIRNEY, E., CERRUTI, L., DURBIN, R., ETWILLER, L., EDDY, S. R., GRIFFITHS-JONES, S., HOWE, K. L., MARSHALL, M. & SONNHAMMER, E. L. (2002). The Pfam protein families database. *Nucleic Acids Res* **30**, 276-280.
- BERMAN, H. M., BATTISTUZ, T., BHAT, T. N., BLUHM, W. F., BOURNE, P. E., BURKHARDT, K., FENG, Z., GILLILAND, G. L., IYPE, L., JAIN, S., FAGAN, P., MARVIN, J., PADILLA, D., RAVICHANDRAN, V., SCHNEIDER, B., THANKI, N., WEISSIG, H., WESTBROOK, J. D. & ZARDECKI, C. (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* **58**, 899-907.

- BONNEAU, R. & BAKER, D. (2001). Ab initio protein structure prediction: progress and prospects. *Annu Rev Biophys Biomol Struct* **30**, 173-189.
- BORG, J. P., MARCHETTO, S., LE BIVIC, A., OLLENDORFF, V., JAULIN-BASTARD, F., SAITO, H., FOURNIER, E., ADELAIDE, J., MARGOLIS, B. & BIRNBAUM, D. (2000). ERBIN: a basolateral PDZ protein that interacts with the mammalian ERBB2/HER2 receptor. *Nat Cell Biol* **2**, 407-414.
- BOWIE, J. U. (1997a). Helix packing angle preferences. *Nature Structural Biology* **4**, 915-917.
- BOWIE, J. U. (1997b). Helix packing in membrane proteins. *Journal of Molecular Biology* **272**, 780-789.
- BOWIE, J. U. (1999). Helix-bundle membrane protein fold templates. *Protein Sci* **8**, 2711-2719.
- BOWIE, J. U., LUTHY, R. & EISENBERG, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170.
- BOYD, D., SCHIERLE, C. & BECKWITH, J. (1998). How many membrane proteins are there? *Protein Sci* **7**, 201-205.
- BRIGGS, J. A., TORRES, J. & ARKIN, I. T. (2001). A new method to model membrane protein structure based on silent amino acid substitutions. *Proteins* **44**, 370-375.
- BRODER, Y. C., KATZ, S. & ARONHEIM, A. (1998). The ras recruitment system, a novel approach to the study of protein-protein interactions. *Curr Biol* **8**, 1121-1124.
- CAPENER, C. E., SHRIVASTAVA, I. H., RANATUNGA, K. M., FORREST, L. R., SMITH, G. R. & SANSOM, M. S. (2000). Homology modeling and molecular dynamics simulation studies of an inward rectifier potassium channel. *Biophys J* **78**, 2929-2942.
- CASARI, G., SANDER, C. & VALENCIA, A. (1995). A method to predict functional residues in proteins. *Nat Struct Biol* **2**, 171-178.
- CHANG, C., EL-KABBANI, O., TIEDE, D., NORRIS, J. & SCHIFFER, M. (1991). Structure of the membrane-bound protein photosynthetic reaction center from Rhodobacter sphaeroides. *Biochemistry* **30**, 5352-5360.
- CHEN, C. P., KERNYTSKY, A. & ROST, B. (2002). Transmembrane helix predictions revisited. *Protein Sci* **11**, 2774-2791.
- CHEN, C. P. & ROST, B. (2002). State-of-the-art in membrane protein prediction. *Applied Bioinformatics* **1**, 21-35.
- CHOTHIA, C., LEVITT, M. & RICHARDSON, D. (1981). Helix to helix packing in proteins. *Journal of Molecular Biology* **145**, 215-250.
- CLAROS, M. G. & VON HEIJNE, G. (1994). TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci* **10**, 685-686.
- COLE, S. T., EIGLMEIER, K., PARKHILL, J., JAMES, K. D., THOMSON, N. R., WHEELER, P. R., HONORE, N., GARNIER, T., CHURCHER, C., HARRIS, D., MUNGALL, K., BASHAM, D., BROWN, D., CHILLINGWORTH, T., CONNOR, R., DAVIES, R. M., DEVLIN, K., DUTHOY, S., FELTWELL, T., FRASER, A., HAMLIN, N., HOLROYD, S., HORNSBY, T., JAGELS, K., LACROIX, C., MACLEAN, J., MOULE, S., MURPHY, L., OLIVER, K., QUAIL, M. A., RAJANDREAM, M. A., RUTHERFORD, K. M., RUTTER, S., SEEGER, K., SIMON, S., SIMMONDS, M., SKELTON, J., SQUARES, R., SQUARES,

- S., STEVENS, K., TAYLOR, K., WHITEHEAD, S., WOODWARD, J. R. & BARRELL, B. G. (2001). Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007-1011.
- CORDES, F. S., BRIGHT, J. N. & SANSOM, M. S. (2002). Proline-induced distortions of transmembrane helices. *J Mol Biol* **323**, 951-960.
- CRASTO, C., MARENCO, L., MILLER, P. & SHEPHERD, G. (2002). Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. *Nucleic Acids Res* **30**, 354-360.
- CSERZO, M., WALLIN, E., SIMON, I., VON HEIJNE, G. & ELOFSSON, A. (1997). Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng* **10**, 673-676.
- DALTON, J. A. R., MICHALOPOULOS, I. & WESTHEAD, D. R. (2003). Calculation of helix packing angles in protein structures. *Bioinformatics* **19**, 1298-1299.
- DASTMALCHI, S., MORRIS, M. B. & CHURCH, W. B. (2001). Modeling of the structural features of integral-membrane proteins reverse-environment prediction of integral membrane protein structure (REPIMPS). *Protein Sci* **10**, 1529-1538.
- DEBER, C. M., WANG, C., LIU, L. P., PRIOR, A. S., AGRAWAL, S., MUSKAT, B. L. & CUTICCHIA, A. J. (2001). TM Finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Sci* **10**, 212-219.
- DERISI, J. L., IYER, V. R. & BROWN, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686.
- DIMAIO, D. & MATTOON, D. (2001). Mechanisms of cell transformation by papillomavirus E5 proteins. *Oncogene* **20**, 7866-7873.
- DOBBS, H., ORLANDINI, E., BONACCINI, R. & SENO, F. (2002). Optimal potentials for predicting inter-helical packing in transmembrane proteins. *Proteins* **49**, 342-349.
- DOOLITTLE, R. F. & FENG, D. F. (1990). Nearest neighbor procedure for relating progressively aligned amino acid sequences. *Methods Enzymol* **183**, 659-669.
- DRAWID, A., JANSEN, R. & GERSTEIN, M. (2000). Genome-wide analysis relating expression level with protein subcellular localization. *Trends Genet* **16**, 426-430.
- DUNNWALD, M., VARSHAVSKY, A. & JOHNSON, N. (1999). Detection of transient in vivo interactions between substrate and transporter during protein translocation into the endoplasmic reticulum. *Mol Biol Cell* **10**, 329-344.
- EHRHARD, K. N., JACOBY, J. J., FU, X. Y., JAHN, R. & DOHLMAN, H. G. (2000). Use of G-protein fusions to monitor integral membrane protein-protein interactions in yeast. *Nat Biotechnol* **18**, 1075-1079.
- EILERS, M., PATEL, A. B., LIU, W. & SMITH, S. O. (2002). Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophysical Journal* **82**, 2720-2736.
- EILERS, M., SHEKAR, S. C., SHIEH, T., SMITH, S. O. & FLEMING, P. J. (2000). Internal packing of helical membrane proteins. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5796-5801.
- EISENBERG, D., MARCOTTE, E. M., XENARIOS, I. & YEATES, T. O. (2000). Protein function in the post-genomic era. *Nature* **405**, 823-826.
- EISENBERG, D., SCHWARZ, E., KOMAROMY, M. & WALL, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol* **179**, 125-142.

- ENGELMAN, D. M., CHEN, J., CHIN, C., CURRAN, R., DIXON, A. M., DUPUY, A., LEE, A., LEHNERT, U., MATHEWS, E., RESHETNYAK, Y., SENES, A. & POPOT, J. L. (2003). Membrane protein folding: beyond the two stage model. *FEBS Lett* **27740**, 1-4.
- ENGELMAN, D. M. & STEITZ, T. A. (1981). The spontaneous insertion of proteins into and across membranes: the helical hairpin hypothesis. *Cell* **23**, 411-422.
- ENGELMAN, D. M. & STEITZ, T. A. (1984). On the folding and insertion of globular membrane proteins. In *The protein folding problem* (ed. D. B. Wetlaufer). American Association for the Advancement of science.
- ENGELMAN, D. M., STEITZ, T. A. & GOLDMAN, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem* **15**, 321-353.
- FIELDS, S. & SONG, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-246.
- FISHER, L. E., ENGELMAN, D. M. & STURGIS, J. N. (1999). Detergents modulate dimerization, but not helicity, of the glycophorin A transmembrane domain. *J Mol Biol* **293**, 639-651.
- FLEISHMAN, S. J. & BEN-TAL, N. (2002). A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. *J Mol Biol* **321**, 363-378.
- FLEMING, K. G., ACKERMAN, A. L. & ENGELMAN, D. M. (1997). The effect of point mutations on the free energy of transmembrane alpha-helix dimerization. *J Mol Biol* **272**, 266-275.
- FRIEDMAN, C., KRA, P., YU, H., KRAUTHAMMER, M. & RZHETSKY, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17 Suppl 1**, S74-82.
- FRYXELL, K. J. (1996). The coevolution of gene family trees. *Trends Genet* **12**, 364-369.
- GAVIN, A. C., BOSCHE, M., KRAUSE, R., GRANDI, P., MARZIOCH, M., BAUER, A., SCHULTZ, J., RICK, J. M., MICHON, A. M., CRUCIAT, C. M., REMOR, M., HOFERT, C., SCHELDER, M., BRAJENOVIC, M., RUFFNER, H., MERINO, A., KLEIN, K., HUDAK, M., DICKSON, D., RUDI, T., GNAU, V., BAUCH, A., BASTUCK, S., HUHSE, B., LEUTWEIN, C., HEURTIER, M. A., COPLEY, R. R., EDELMANN, A., QUERFURTH, E., RYBIN, V., DREWES, G., RAID, M., BOUWMEESTER, T., BORK, P., SERAPHIN, B., KUSTER, B., NEUBAUER, G. & SUPERTI-FURGA, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147.
- GERSTEIN, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J Mol Biol* **274**, 562-576.
- GERSTEIN, M. (1998). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins* **33**, 518-534.
- GERSTEIN, M. & CHOTHIA, C. (1999). Perspectives: signal transduction. Proteins in motion. *Science* **285**, 1682-1683.
- GERSTEIN, M., LIN, J. & HEGYI, H. (2000). Protein folds in the worm genome. *Pac Symp Biocomput*, 30-41.
- GERSTEIN, M. & RICHARDS, F. M. (2001). Protein Geometry: Distances, Areas, and Volumes. *International Tables for Crystallography F*, 531-539.



- GIRVAN, M. & NEWMAN, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 7821-7826.
- GLUSMAN, G., YANAI, I., RUBIN, I. & LANCET, D. (2001). The complete human olfactory subgenome. *Genome Res.* **11**, 685-702.
- GOBEL, U., SANDER, C., SCHNEIDER, R. & VALENCIA, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-317.
- GOFFEAU, A., NAKAI, K., SLONIMSKI, P., RISLER, J. L. & SLOMINSKI, P. (1993). The membrane proteins encoded by yeast chromosome III genes. *FEBS Lett* **325**, 112-117.
- GOH, C. S., BOGAN, A. A., JOACHIMIAK, M., WALTHER, D. & COHEN, F. E. (2000). Co-evolution of proteins with their interaction partners. *J Mol Biol* **299**, 283-293.
- GOH, C. S. & COHEN, F. E. (2002). Co-evolutionary analysis reveals insights into protein-protein interactions. *J Mol Biol* **324**, 177-192.
- HARRISON, P., KUMAR, A., LAN, N., ECHOLS, N., SNYDER, M. & GERSTEIN, M. (2002). A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J Mol Biol* **316**, 409-419.
- HARRISON, P. M., ECHOLS, N. & GERSTEIN, M. B. (2001). Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res.* **29**, 818-830.
- HARRISON, P. M., MILBURN, D., ZHANG, Z., BERTONE, P. & GERSTEIN, M. (2003). Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res* **31**, 1033-1037.
- HARTWELL, L. H., HOPFIELD, J. J., LEIBLER, S. & MURRAY, A. W. (1999). From molecular to modular cell biology. *Nature* **402**, C47-52.
- HIROKAWA, T., BOON-CHIENG, S. & MITAKU, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**, 378-379.
- HO, Y., GRUHLER, A., HEILBUT, A., BADER, G. D., MOORE, L., ADAMS, S. L., MILLAR, A., TAYLOR, P., BENNETT, K., BOUTILIER, K., YANG, L., WOLTING, C., DONALDSON, I., SCHANDORFF, S., SHEWNARANE, J., VO, M., TAGGART, J., GOUDREAU, M., MUSKAT, B., ALFARANO, C., DEWAR, D., LIN, Z., MICHALICKOVA, K., WILLEMS, A. R., SASSI, H., NIELSEN, P. A., RASMUSSEN, K. J., ANDERSEN, J. R., JOHANSEN, L. E., HANSEN, L. H., JESPERSEN, H., PODTELEJNIKOV, A., NIELSEN, E., CRAWFORD, J., POULSEN, V., SORESENSEN, B. D., MATTHIESEN, J., HENDRICKSON, R. C., GLEESON, F., PAWSON, T., MORAN, M. F., DUROCHER, D., MANN, M., HOGUE, C. W., FIGEYS, D. & TYERS, M. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180-183.
- HOFMANN, K. & STOFFEL, W. (1993). TMbase - A database of membrane spanning protein segments. *Biol Chem Hoppe-Seyler* **374**.
- HOMMA, K., FUKUCHI, S., KAWABATA, T., OTA, M. & NISHIKAWA, K. (2002). A systematic investigation identifies a significant number of probable pseudogenes in the *Escherichia coli* genome. *Gene* **294**, 25.

- HOOD, L., HUANG, H. V. & DREYER, W. J. (1977). The area-code hypothesis: the immune system provides clues to understanding the genetic and molecular basis of cell recognition during development. *J Supramol Struct* **7**, 531-559.
- HORN, F., BETTLER, E., OLIVEIRA, L., CAMPAGNE, F., COHEN, F. E. & VRIEND, G. (2003). GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res* **31**, 294-297.
- HUBERMAN, B. A. & ADAMIC, L. A. (1999). Growth dynamics of the World-Wide Web. *Nature* **401**, 131.
- HUBSMAN, M., YUDKOVSKY, G. & ARONHEIM, A. (2001). A novel approach for the identification of protein-protein interaction with integral membrane proteins. *Nucleic Acids Res* **29**, E18.
- HUH, W. K., FALVO, J. V., GERKE, L. C., CARROLL, A. S., HOWSON, R. W., WEISSMAN, J. S. & O'SHEA, E. K. (2003). Global analysis of protein localization in budding yeast. *Nature* **425**, 686-691.
- IKEDA, M., ARAI, M., LAO, D. M. & SHIMIZU, T. (2002). 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.
- IM, W., FEIG, M. & BROOKS, C. L., 3RD. (2003). An implicit membrane generalized born theory for the study of structure, stability, and interactions of membrane proteins. *Biophys J* **85**, 2900-2918.
- INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.
- ITO, T., TASHIRO, K., MUTA, S., OZAWA, R., CHIBA, T., NISHIZAWA, M., YAMAMOTO, K., KUHARA, S. & SAKAKI, Y. (2000). Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 1143-1147.
- JANSEN, R. & GERSTEIN, M. (2000). Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res* **28**, 1481-1488.
- JANSEN, R., LAN, N., QIAN, J. & GERSTEIN, M. (2002). Integration of genomic datasets to predict protein complexes in yeast. *Journal of Structural and Functional Genomics* **2**, 71-81.
- JANSEN, R., YU, H., GREENBAUM, D., KLUGER, Y., KROGAN, N. J., CHUNG, S., EMILI, A., SNYDER, M., GREENBLATT, J. F. & GERSTEIN, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-453.
- JAVADPOUR, M. M., EILERS, M., GROESBEEK, M. & SMITH, S. O. (1999). Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. *Biophys J* **77**, 1609-1618.
- JAYASINGHE, S., HRISTOVA, K. & WHITE, S. H. (2001). Energetics, stability, and prediction of transmembrane helices. *J Mol Biol* **312**, 927-934.
- JEONG, H., MASON, S. P., BARABASI, A. L. & OLTVAI, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411**, 41-42.

- JESPER, L., LIJNEN, H. R., VANWETSWINKEL, S., VAN HOF, B., BREPOELS, K., COLLEN, D. & DE MAEYER, M. (1999). Guiding a docking mode by phage display: selection of correlated mutations at the staphylokinase-plasmin interface. *J Mol Biol* **290**, 471-479.
- JOHNSSON, N. & VARSHAVSKY, A. (1994). Split ubiquitin as a sensor of protein interactions in vivo. *Proc Natl Acad Sci U S A* **91**, 10340-10344.
- JONES, D. T. (1998). Do transmembrane protein superfolds exist? *FEBS Lett* **423**, 281-285.
- JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. (1992). A new approach to protein fold recognition. *Nature* **358**, 86-89.
- JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33**, 3038-3049.
- JURETIC, D., ZORANIC, L. & ZUCIC, D. (2002). Basic charge clusters and predictions of membrane protein topology. *J Chem Inf Comput Sci* **42**, 620-632.
- KEASAR, C. & LEVITT, M. (2003). A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *J Mol Biol* **329**, 159-174.
- KEEGAN, K. & COOPER, J. A. (1996). Use of the two hybrid system to detect the association of the protein-tyrosine-phosphatase, SHPTP2, with another SH2-containing protein, Grb7. *Oncogene* **12**, 1537-1544.
- KERR, I. D., SANKARARAMAKRISHNAN, R., SMART, O. S. & SANSOM, M. S. (1994). Parallel helix bundles and ion channels: molecular modeling via simulated annealing and restrained molecular dynamics. *Biophys J* **67**, 1501-1515.
- KIM, S., CHAMBERLAIN, A. K. & BOWIE, J. U. (2003). A simple method for modeling transmembrane helix oligomers. *J Mol Biol* **329**, 831-840.
- KLEIGER, G., GROTHE, R., MALLICK, P. & EISENBERG, D. (2002). GXXXG and AXXXA: common alpha-helical interaction motifs in proteins, particularly in extremophiles. *Biochemistry* **41**, 5990-5997.
- KLEIGER, G., PERRY, J. & EISENBERG, D. (2001). 3D structure and significance of the GPhiXXG helix packing motif in tetramers of the E1beta subunit of pyruvate dehydrogenase from the archeon *Pyrobaculum aerophilum*. *Biochemistry* **40**, 14484-14492.
- KLEIN, P., KANEHISA, M. & DELISI, C. (1985). The detection and classification of membrane-spanning proteins. *Biochim Biophys Acta* **815**, 468-476.
- KROGH, A., LARSSON, B., VON HEIJNE, G. & SONNHAMMER, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580.
- KUMAR, A., AGARWAL, S., HEYMAN, J. A., MATSON, S., HEIDTMAN, M., PICCIRILLO, S., UMANSKY, L., DRAWID, A., JANSEN, R., LIU, Y., CHEUNG, K. H., MILLER, P., GERSTEIN, M., ROEDER, G. S. & SNYDER, M. (2002). Subcellular localization of the yeast proteome. *Genes Dev* **16**, 707-719.
- KYTE, J. & DOOLITTLE, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105-132.

- LAN, N., MONTELIONE, G. T. & GERSTEIN, M. (2003). Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. *Curr Opin Chem Biol* **7**, 44-54.
- LANDGRAF, R., XENARIOS, I. & EISENBERG, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol* **307**, 1487-1502.
- LANGOSCH, D., BROSIG, B., KOLMAR, H. & FRITZ, H. J. (1996). Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator. *J Mol Biol* **263**, 525-530.
- LASER, H., BONGARDS, C., SCHULLER, J., HECK, S., JOHNSSON, N. & LEHMING, N. (2000). A new screen for protein interactions reveals that the *Saccharomyces cerevisiae* high mobility group proteins Nhp6A/B are involved in the regulation of the GAL1 promoter. *Proc Natl Acad Sci U S A* **97**, 13732-13737.
- LAZARIDIS, T. (2003). Effective energy function for proteins in lipid membranes. *Proteins* **52**, 176-192.
- LEE, T. I., RINALDI, N. J., ROBERT, F., ODOM, D. T., BAR-JOSEPH, Z., GERBER, G. K., HANNETT, N. M., HARBISON, C. T., THOMPSON, C. M., SIMON, I., ZEITLINGER, J., JENNINGS, E. G., MURRAY, H. L., GORDON, D. B., REN, B., WYRICK, J. J., TAGNE, J. B., VOLKERT, T. L., FRAENKEL, E., GIFFORD, D. K. & YOUNG, R. A. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799-804.
- LEMMON, M. A., FLANAGAN, J. M., HUNT, J. F., ADAIR, B. D., BORMANN, B. J., DEMPSEY, C. E. & ENGELMAN, D. M. (1992). Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices. *J Biol Chem* **267**, 7683-7689.
- LESK, A. M. (1997). CASP2: report on ab initio predictions. *Proteins Suppl* **1**, 151-166.
- LESK, A. M., LO CONTE, L. & HUBBARD, T. J. (2001). Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins Suppl* **5**, 98-118.
- LI, W. H. A. G., D. (1991). *Fundamentals of Molecular Evolution*. Sinauer Associates.
- LICHTARGE, O., BOURNE, H. R. & COHEN, F. E. (1996a). Evolutionarily conserved Galphabeta gamma binding surfaces support a model of the G protein-receptor complex. *Proc Natl Acad Sci U S A* **93**, 7507-7511.
- LICHTARGE, O., BOURNE, H. R. & COHEN, F. E. (1996b). An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **257**, 342-358.
- LICHTARGE, O., YAMAMOTO, K. R. & COHEN, F. E. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J Mol Biol* **274**, 325-337.
- LIN, J., QIAN, J., GREENBAUM, D., BERTONE, P., DAS, R., ECHOLS, N., SENES, A., STENGER, B. & GERSTEIN, M. (2002). GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing. *Nucleic Acids Res* **30**, 4574-4582.
- LIO, P. & VANNUCCI, M. (2000). Wavelet change-point prediction of transmembrane proteins. *Bioinformatics* **16**, 376-382.

- LIU, Y., ENGELMAN, D. M. & GERSTEIN, M. (2002). Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol* **3**, research0054.0051-0054.0012.
- LIU, Y., GERSTEIN, M. & ENGELMAN, D. M. (2004). Transmembrane protein domains rarely use covalent domain recombination as an evolutionary mechanism. *PNAS* **101**, 3495-3497.
- MACKENZIE, K. R., PRESTEGARD, J. H. & ENGELMAN, D. M. (1997). A transmembrane helix dimer: structure and implications. *Science* **276**, 131-133.
- MELEN, K., KROGH, A. & VON HEIJNE, G. (2003). Reliability measures for membrane protein topology prediction algorithms. *J Mol Biol* **327**, 735-744.
- MENDEZ, R., LEPLAE, R., DE MARIA, L. & WODAK, S. J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins* **52**, 51-67.
- MEWES, H. W., FRISHMAN, D., GULDENER, U., MANNHAUPT, G., MAYER, K., MOKREJS, M., MORGENSTERN, B., MUNSTERKOTTER, M., RUDD, S. & WEIL, B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Research* **30**, 31-34.
- MICHEL, H. (1982). Three-dimensional crystals of a membrane protein complex: the photosynthetic reaction centre from *Rhodospseudomonas viridis*. *J Mol Biol* **158**, 567-572.
- MICHEL, H. (1983). Crystallization of membrane proteins. *Trends Biochem Sci* **8**, 56-59.
- MICHEL, H., WEYER, K., GRUENBERG, H., DUNGER, L., OESTERHELT, D. & LOTTSPREICH, F. (1986). The "light" and "medium" subunits of the photosynthetic reaction centre from *Rhodospseudomonas viridis*: isolation of the genes, nucleotide and amino acid sequence. *EMBO* **5**, 1149-1158.
- MICHEL, H., WEYER, K., GRUENBERG, H. & LOTTSPREICH, F. (1985). The "heavy" subunit of the photosynthetic reaction centre from *Rhodospseudomonas viridis*: isolation of the gene, nucleotide and amino acid sequence. *EMBO* **4**, 1667-1672.
- MIGHELL, A. J., SMITH, N. R., ROBINSON, P. A. & MARKHAM, A. F. (2000). Vertebrate pseudogenes. *FEBS Lett.* **468**, 109-114.
- MOLLER, S., CRONING, M. D. & APWEILER, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**, 646-653.
- MOYLE, W. R., CAMPBELL, R. K., MYERS, R. V., BERNARD, M. P., HAN, Y. & WANG, X. (1994). Co-evolution of ligand-receptor pairs. *Nature* **368**, 251-255.
- NAGASHIMA, K., MATSUURA, K., OHYAMA, S. & SHIMADA, K. (1994). Primary Structure and Transcription of Genes Encoding B870 and Photosynthetic Reaction Center Apoproteins from *Rubrivivax gelatinosus*. *J Biol Chem* **269**, 2477-2484.
- NAKAI, K. & KANEHISA, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **14**, 897-911.
- NIELSEN, H., BRUNAK, S. & VON HEIJNE, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng* **12**, 3-9.
- NILSSON, J., PERSSON, B. & VON HEIJNE, G. (2000). Consensus predictions of membrane protein topology. *FEBS Lett* **486**, 267-269.
- OHNO, S. (1970). *Evolution by Gene Duplication*. Springer Verlag, New York.

- OLMEA, O., ROST, B. & VALENCIA, A. (1999). Effective use of sequence correlation and conservation in fold recognition. *J Mol Biol* **293**, 1221-1239.
- OLMEA, O. & VALENCIA, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des* **2**, S25-32.
- ORTIZ, A. R., KOLINSKI, A., ROTKIEWICZ, P., ILKOWSKI, B. & SKOLNICK, J. (1999). Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins Suppl* **3**, 177-185.
- OZENBERGER, B. A. & YOUNG, K. H. (1995). Functional interaction of ligands and receptors of the hematopoietic superfamily in yeast. *Mol Endocrinol* **9**, 1321-1329.
- PAPPU, R. V., MARSHALL, G. R. & PONDER, J. W. (1999). A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nat Struct Biol* **6**, 50-55.
- PASQUIER, C., PROMPONAS, V. J., PALAIOS, G. A., HAMODRAKAS, J. S. & HAMODRAKAS, S. J. (1999). A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm. *Protein Eng* **12**, 381-385.
- PAULSEN, I. T., NGUYEN, L., SLIWINSKI, M. K., RABUS, R. & SAIER, M. H., JR. (2000). Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J Mol Biol* **301**, 75-100.
- PAULSEN, I. T., SLIWINSKI, M. K. & SAIER, M. H., JR. (1998). Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J Mol Biol* **277**, 573-592.
- PAZOS, F., HELMER-CITTERICH, M., AUSIELLO, G. & VALENCIA, A. (1997). Correlated mutations contain information about protein-protein interaction. *Journal Of Molecular Biology* **271**, 511-523.
- PAZOS, F. & VALENCIA, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* **14**, 609-614.
- PAZOS, F. & VALENCIA, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219-227.
- PELLEGRINI, M., MARCOTTE, E. M., THOMPSON, M. J., EISENBERG, D. & YEATES, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-4288.
- PELLEGRINI-CALACE, M., CAROTTI, A. & JONES, D. T. (2003). Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3D structures. *Proteins* **50**, 537-545.
- PERSSON, B. & ARGOS, P. (1996). Topology prediction of membrane proteins. *Protein Sci* **5**, 363-371.
- PILPEL, Y., BEN-TAL, N. & LANCET, D. (1999). kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J Mol Biol* **294**, 921-935.
- PINTO, L. H., DIECKMANN, G. R., GANDHI, C. S., PAPWORTH, C. G., BRAMAN, J., SHAUGHNESSY, M. A., LEAR, J. D., LAMB, R. A. & DEGRADO, W. F. (1997). A functionally defined model for the M2 proton channel of influenza A virus

- suggests a mechanism for its ion selectivity. *Proc Natl Acad Sci U S A* **94**, 11301-11306.
- POGOZHEVA, I. D., LOMIZE, A. L. & MOSBERG, H. I. (1997). The transmembrane 7-alpha-bundle of rhodopsin: distance geometry calculations with hydrogen bonding constraints. *Biophys J* **72**, 1963-1985.
- POPOT, J. L. & ENGELMAN, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry* **29**, 4031-4037.
- POPOT, J. L. & ENGELMAN, D. M. (2000). Helical membrane protein folding, stability, and evolution. *Annu Rev Biochem* **69**, 881-922.
- QIAN, J., STENGER, B., WILSON, C. A., LIN, J., JANSEN, R., TEICHMANN, S. A., PARK, J., KREBS, W. G., YU, H., ALEXANDROV, V., ECHOLS, N. & GERSTEIN, M. (2001). PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Res* **29**, 1750-1764.
- RICHARDS, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *Journal of Molecular Biology* **82**, 1-14.
- RICHARDS, F. M. (1985). Calculation of molecular volumes and areas for structures of known geometry. *Methods in Enzymology* **115**, 440-464.
- ROST, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol* **266**, 525-539.
- ROST, B., CASADIO, R., FARISELLI, P. & SANDER, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Sci* **4**, 521-533.
- ROST, B., FARISELLI, P. & CASADIO, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci* **5**, 1704-1718.
- RUSS, W. P. & ENGELMAN, D. M. (1999). TOXCAT: a measure of transmembrane helix association in a biological membrane. *Proc Natl Acad Sci U S A* **96**, 863-868.
- RUSS, W. P. & ENGELMAN, D. M. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. *J Mol Biol* **296**, 911-919.
- RUUD, P., FODSTAD, O. & HOVIG, E. (1999). Identification of a novel cytokeratin 19 pseudogene that may interfere with reverse transcriptase-polymerase chain reaction assays used to detect micrometastatic tumor cells. *Int J Cancer* **80**, 119-125.
- SCHNEIDER, D. & ENGELMAN, D. M. (2003). GALLEX, a measurement of heterologous association of transmembrane helices in a biological membrane. *J Biol Chem* **278**, 3105-3111.
- SENES, A., GERSTEIN, M. & ENGELMAN, D. M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol* **296**, 921-936.
- SENES, A., UBARETXENA-BELANDIA, I. & ENGELMAN, D. M. (2001). The Calpha --- H...O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc Natl Acad Sci U S A* **98**, 9056-9061.
- STEITZ, T. A., GOLDMAN, A. & ENGELMAN, D. M. (1982). Quantitative application of the helical hairpin hypothesis to membrane proteins. *Biophys. J.* **37**, 124-125.
- SURTI, T., KLEIN, O., ASCHHEIM, K., DI MAIO, D. & SMITH, S. O. (1998). Structural models of the bovine papillomavirus E5 protein. *Proteins* **33**, 601-612.

- TAMAMES, J., CASARI, G., OUZOUNIS, C. & VALENCIA, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J Mol Evol* **44**, 66-73.
- TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E. V., KRYLOV, D. M., MAZUMDER, R., MEKHEDOV, S. L., NIKOLSKAYA, A. N., RAO, B. S., SMIRNOV, S., SVERDLOV, A. V., VASUDEVAN, S., WOLF, Y. I., YIN, J. J. & NATALE, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.
- THE C. ELEGANS SEQUENCING CONSORTIUM. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* Sequencing Consortium.[erratum appears in Science 1999 Jan 1;283(5398):35]. *Science* **282**, 2012-2018.
- THORNER, J., COGDELL, R., SEFTOR, R. & WEBSTER, G. (1980). Further studies on the composition and spectral properties of the photochemical reaction centers of bacteriochlorophyll-b containing bacteria. *Biochim Biophys Acta* **593**, 60-75.
- TOURMEN, Y., BARIS, O., DESSEN, P., JACQUES, C., MALTHIERY, Y. & REYNIER, P. (2002). Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics* **80**, 71-77.
- TSAL, J. & GERSTEIN, M. (2002). Calculations of protein volumes: sensitivity analysis and parameter database. *Bioinformatics* **18**, 985-995.
- TSAL, J., VOSS, N. & GERSTEIN, M. (2001). Determining the minimum number of types necessary to represent the sizes of protein atoms. *Bioinformatics* **17**.
- TUSNADY, G. E. & SIMON, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* **283**, 489-506.
- UETZ, P., GIOT, L., CAGNEY, G., MANSFIELD, T. A., JUDSON, R. S., KNIGHT, J. R., LOCKSHON, D., NARAYAN, V., SRINIVASAN, M., POCHART, P., QURESHI-EMILI, A., LI, Y., GODWIN, B., CONOVER, D., KALBFLEISCH, T., VIJAYADAMODAR, G., YANG, M., JOHNSTON, M., FIELDS, S. & ROTHBERG, J. M. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623-627.
- VALENCIA, A. & PAZOS, F. (2002). Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol* **12**, 368-373.
- VANIN, E. F. (1985). Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.* **19**, 253-272.
- VON HEIJNE, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225**, 487-494.
- VON HEIJNE, G. (1999). Recent advances in the understanding of membrane protein assembly and structure. *Quat. Rev. Biophys.* **32**, 285-307.
- VON MERING, C., KRAUSE, R., SNEL, B., CORNELL, M., OLIVER, S. G., FIELDS, S. & BORK, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403.
- WALHOUT, A. J., SORDELLA, R., LU, X., HARTLEY, J. L., TEMPLE, G. F., BRASCH, M. A., THIERRY-MIEG, N. & VIDAL, M. (2000). Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116-122.



- WALLIN, E. & VON HEIJNE, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* **7**, 1029-1038.
- WALTHER, D., EISENHABER, F. & ARGOS, P. (1996). Principles of helix-helix packing in proteins: the helical lattice superposition model. *Journal of Molecular Biology* **255**, 536-553.
- WALTHER, D., SPRINGER, C. & COHEN, F. E. (1998). Helix-helix packing angle preferences for finite helix axes. *Proteins* **33**, 457-459.
- WATERSTON, R. H., LINDBLAD-TOH, K., BIRNEY, E., ROGERS, J., ABRIL, J. F., AGARWAL, P., AGARWALA, R., AINSCOUGH, R., ALEXANDERSSON, M., AN, P., ANTONARAKIS, S. E., ATTWOOD, J., BAERTSCH, R., BAILEY, J., BARLOW, K., BECK, S., BERRY, E., BIRREN, B., BLOOM, T., BORK, P., BOTCHERBY, M., BRAY, N., BRENT, M. R., BROWN, D. G., BROWN, S. D., BULT, C., BURTON, J., BUTLER, J., CAMPBELL, R. D., CARNINCI, P., CAWLEY, S., CHIAROMONTE, F., CHINWALLA, A. T., CHURCH, D. M., CLAMP, M., CLEE, C., COLLINS, F. S., COOK, L. L., COPLEY, R. R., COULSON, A., COURONNE, O., CUFF, J., CURWEN, V., CUTTS, T., DALY, M., DAVID, R., DAVIES, J., DELEHAUNTY, K. D., DERI, J., DERMITZAKIS, E. T., DEWEY, C., DICKENS, N. J., DIEKHANS, M., DODGE, S., DUBCHAK, I., DUNN, D. M., EDDY, S. R., ELNITSKI, L., EMES, R. D., ESWARA, P., EYRAS, E., FELSENFELD, A., FEWELL, G. A., FLICEK, P., FOLEY, K., FRANKEL, W. N., FULTON, L. A., FULTON, R. S., FUREY, T. S., GAGE, D., GIBBS, R. A., GLUSMAN, G., GNERRE, S., GOLDMAN, N., GOODSTADT, L., GRAFHAM, D., GRAVES, T. A., GREEN, E. D., GREGORY, S., GUIGO, R., GUYER, M., HARDISON, R. C., HAUSSLER, D., HAYASHIZAKI, Y., HILLIER, L. W., HINRICHS, A., HLAVINA, W., HOLZER, T., HSU, F., HUA, A., HUBBARD, T., HUNT, A., JACKSON, I., JAFFE, D. B., JOHNSON, L. S., JONES, M., JONES, T. A., JOY, A., KAMAL, M., KARLSSON, E. K., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562.
- WATTS, D. J. & STROGATZ, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393**, 440-442.
- WEYER, K., LOTTSPEICH, F., GRUENBERG, H., LANG, F., OSTERHELT, D. & MICHEL, H. (1987). Amino acid sequence of the cytochrome subunit of the photosynthetic reaction centre from the purple bacterium *Rhodospseudomonas viridis*. *EMBO* **6**, 2197-2202.
- WIMLEY, W. C. & WHITE, S. H. (1996). Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat Struct Biol* **3**, 842-848.
- WINZELER, E. A., SHOEMAKER, D. D., ASTROMOFF, A., LIANG, H., ANDERSON, K., ANDRE, B., BANGHAM, R., BENITO, R., BOEKE, J. D., BUSSEY, H., CHU, A. M., CONNELLY, C., DAVIS, K., DIETRICH, F., DOW, S. W., EL BAKKOURY, M., FOURY, F., FRIEND, S. H., GENTALEN, E., GIAEVER, G., HEGEMANN, J. H., JONES, T., LAUB, M., LIAO, H., DAVIS, R. W. & ET AL. (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901-906.
- WOISCHNIK, M. & MORAES, C. T. (2002). Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Res* **12**, 885-893.

- XENARIOS, I., SALWINSKI, L., DUAN, X. J., HIGNEY, P., KIM, S. M. & EISENBERG, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* **30**, 303-305.
- YU, H., ZHU, X., GREENBAUM, D., KARRO, J. & GERSTEIN, M. (2003). TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *NAR in press*.
- ZHANG, Z., BERRY, E. A., HUANG, L. S. & KIM, S. H. (2000). Mitochondrial cytochrome bc1 complex. *Sub-Cellular Biochemistry* **35**, 541-580.
- ZHANG, Z. & GERSTEIN, M. (2003). The human genome has 49 cytochrome c pseudogenes, including a relic of a primordial gene that still functions in mouse. *Gene* **312**, 61-72.
- ZHANG, Z., HARRISON, P. & GERSTEIN, M. (2002). Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res* **12**, 1466-1482.
- ZHANG, Z., HARRISON, P., LIU, Y. & GERSTEIN, M. unpublished results.
- ZHANG, Z., HUANG, L., SHULMEISTER, V. M., CHI, Y. I., KIM, K. K., HUNG, L. W., CROFTS, A. R., BERRY, E. A. & KIM, S. H. (1998). Electron transfer by domain movement in cytochrome bc1. *Nature* **392**, 677-684.
- ZHOU, H. & ZHOU, Y. (2003). Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. *Protein Sci* **12**, 1547-1555.
- ZHU, H., BILGIN, M., BANGHAM, R., HALL, D., CASAMAYOR, A., BERTONE, P., LAN, N., JANSEN, R., BIDLINGMAIER, S., HOUFEK, T., MITCHELL, T., MILLER, P., DEAN, R., GERSTEIN, M. & SNYDER, M. (2001). Global analysis of protein activities using proteome chips. *Science* **293**, 2101-2105.
- ZUCKERKANDL, E. (1975). The appearance of new structures and functions in proteins during evolution. *J Mol Evol* **7**, 1-57.

### **Captions:**

#### Table Captions:

Table 1. A representative list of Web servers for predicting membrane helical protein topology.

Table 2: Volumes of different membrane proteins and their packing efficiency. The relative packing efficiency (rel. pack. eff.) is defined as  $V(\text{ref})/V - 100\%$  with  $V$  being the observed volume of the buried atoms in the structure and  $V(\text{ref})$  is the standard reference volume of the atoms in soluble proteins (Gerstein & Chothia, 1999).

Table 3: The scale of the membrane protein interactome. Orthologous proteins across the organisms have been assigned according to the COG database. In row C the number of proteins is narrowed down to the ones with a known function excluding poorly characterized functional groups named R and S in the COG database.

<b>Name, Reference</b>	<b>URL</b>	<b>Method</b>
ALOM, (Nakai & Kanehisa, 1992)	<a href="http://psort.nibb.ac.jp/">http://psort.nibb.ac.jp/</a>	Sliding window + positive-inside rule
DAS, (Cserzo et al., 1997)	<a href="http://www.sbc.su.se/~miklos/DAS/">http://www.sbc.su.se/~miklos/DAS/</a>	Dense alignment surface
HMMTOP, (Tusnady & Simon, 1998)	<a href="http://www.enzim.hu/hmmtop/">http://www.enzim.hu/hmmtop/</a>	Model-based, HMM
MEMSAT, (Jones et al., 1994)	<a href="http://www.psidpred.net/">http://www.psidpred.net/</a>	Model-based, Dynamic programming
PHDhtm, (Rost et al., 1996)	<a href="http://www.predictprotein.org/">http://www.predictprotein.org/</a>	Neural network
PRED-TMR, (Pasquier et al., 1999)	<a href="http://biophysics.biol.uoa.gr/PRED-TMR2/">http://biophysics.biol.uoa.gr/PRED-TMR2/</a>	Sliding window + edge detection
SOSUI, (Hirokawa et al., 1998)	<a href="http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html">http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html</a>	Sliding window + positive-inside rule
SPLIT, (Juretic et al., 2002)	<a href="http://pref.etfos.hr/">http://pref.etfos.hr/</a>	Sliding window + positive-inside rule
THUMBUP, (Zhou & Zhou, 2003)	<a href="http://theory.med.buffalo.edu/Softwares-Services_files/thumbup.htm">http://theory.med.buffalo.edu/Softwares-Services_files/thumbup.htm</a>	Sliding window + positive-inside rule
TMAP, (Persson & Argos, 1996)	<a href="http://www.mbb.ki.se/tmap/">http://www.mbb.ki.se/tmap/</a>	Multiple sequence alignments
TMFinder, (Deber et al.,	<a href="http://www.bioinformatics-">http://www.bioinformatics-</a>	Sliding window

<b>Name, Reference</b>	<b>URL</b>	<b>Method</b>
2001)	<a href="http://canada.org/TM/">canada.org/TM/</a>	
TMHMM, (Krogh et al., 2001)	<a href="http://www.cbs.dtu.dk/services/TMHMM/">http://www.cbs.dtu.dk/services/TMHMM/</a>	Model-based, HMM
Tmpred, (Hofmann & Stoffel, 1993)	<a href="http://www.ch.embnet.org/software/TMPRED_form.html">http://www.ch.embnet.org/software/TMPRED_form.html</a>	Sliding window + positive-inside rule
TopPred, (Claros & von Heijne, 1994)	<a href="http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html">http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html</a>	Sliding window + positive-inside rule

Table 1

<b>Membrane protein</b>	<b>PDB</b>	<b>Volume buried [Å<sup>3</sup>]</b>	<b>Ref. Vol. buried</b>	<b>Rel. Pack. Eff. [%]</b>
<b>Bacteriorhodopsin</b>	2brd	7889	8030	98.3
<b>Cytochrome bc<sub>1</sub> complex</b>	1bgy	130467	132866	98.2
<b>Glycophorin A</b>	1afo	1087	1221	89.0

Table 2

		<b>MG</b>	<b>EC</b>	<b>SC</b>	<b>MG- EC</b>	<b>MG- SC</b>	<b>EC-SC</b>	<b>MG- EC-SC</b>
<i>A</i>	<b>Total nr. membrane prot.</b>	110	1030	1140				
<i>B</i>	<b>Nr. of A with COG assignment</b>	57	875	513	44	24	243	23
<i>C</i>	<b>Nr. of B with known function (not R and S)</b>	45	685	425	40	22	221	22
<i>D</i>	<b>TM-helices in proteins from C</b>	261	4386	2874	236	138	1805	138
<i>E</i>	<b>Potential TM helix pairs</b>	35124	9639080	4144093	28835	10120	1638466	10120
<i>F</i>	<b>Nr. inter-molecular pairs from E</b>	34191	9620691	4131375	27966	9591	1629915	9591
<i>G</i>	<b>Nr. intra- molecular pairs from E</b>	933	18389	12718	869	529	8551	529

Table 3

### **Figure Captions:**

Fig. 1: The relationship between co-evolution correlation score and interfacial surface area. The correlation scores are shown for different complexes: L-M, cytochrome (C)-L, H-M, C-M, H-L and H-C.

Fig. 2: Amino-acid compositions of TM-helices (i) for all helical membrane protein sequences and (ii) for conserved positions of 168 Pfam-A families of polytopic membrane domains that contain more than 20 members. Only the most frequent amino acids are shown. Data are taken from (Liu et al., 2002).

Fig. 3: TM helix distribution for the virtual organisms MC-EC, MC-SC, EC-SC and MG-EC-SC. The number of helices is expressed as a fraction of the total number of helices to make the bars comparable between different organisms.



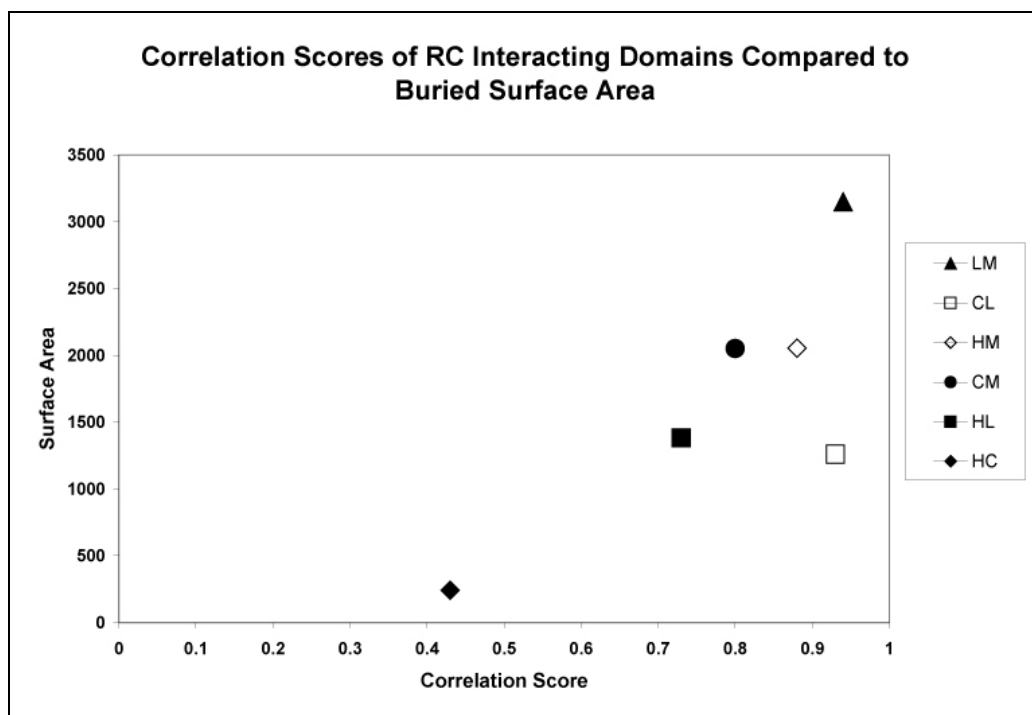


Figure 1

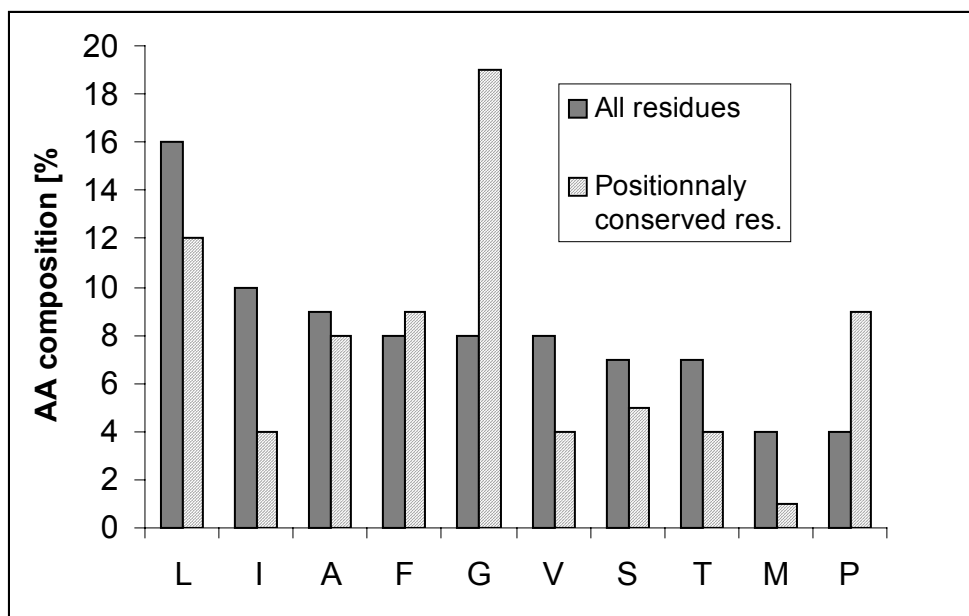


Fig. 2:

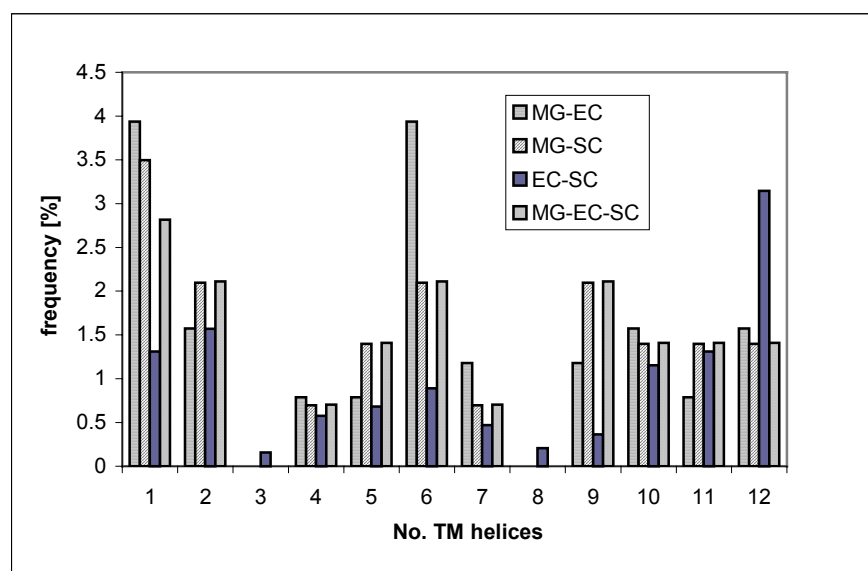


Fig. 3: