Integrated Bioinformatic Approaches to the Prediction of Protein-Protein Interactions

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy
in the Faculty of Life Sciences

2011

Lisa Li Chun Lee

Table of Contents[1]

[1] WORD COUNT: 41,854

List of Figures

List of Tables

List of Abbreviations

APC	The average product correction
AUC	Area Under ROC Curve
BIND	Biomolecular Interaction Network Database
BioGRID	The General Repository for Interaction Datasets
BLAST	Best Local Alignment Search Tool
DIP	Database of Interacting Proteins
EM	Electron Microscopy
ERS	Entropy Reduction Step
GFP	Green Fluorescent Protein
GO	Gene Ontology
GPCR	G Protein-Coupled Receptor
JTT	The Jones-Taylor-Thornton model
NMR	Nuclear Magnetic Resonance Spectroscopy
MI	Mutual Information
MIPS	Mammalian Protein-protein Interactions Database
MS	Mass spectrometry
MSA	Multiple Sequence Alignment
PDB	Protein Data Bank
PPI	Protein-Protein Interaction
PQS	Protein Quaternary Structure database
RDF	Random decision forest
RBH	Reciprocal Best BLAST Hits
ROC	Receiver Operating Characteristic Curve
rRNA	ribosomal RNA
SVM	Support Vector Machine
TAP	Tandem Affinity Purification
TM	Transmembrane Domain
Y2H	Yeast Two-Hybrid

Abstract

Protein-protein interactions (PPIs) play a fundamental role in many biological processes such as signal transduction from the extracelluar space to cytosol. Functions of less characterized proteins can often be deduced from PPI networks. Various sequence-based approaches were taken to predicting and understanding potential PPIs using bioinformatic means. Initially, the *mirrortree* method was comprehensively examined to derive a robust approach for PPI predictions. The analysis has revealed that *mirrortree* is extremely sensitive to many factors especially sequence diversity and the selection of orthologues. Indeed, higher sequence diversity improves the predictive power of the approach. In an attempt to improve prediction accuracy, various speciation signal correction methods were evaluated and the RNA-based approaches appear to be more effective in removing the speciation signal and ultimately produce more accurate predictions. The utility of *mirrortree* was further extended for domain-domain interactions in fibrillin-1. However, due to the low sequence diversity of the orthologues, poor prediction results were obtained. Furthermore, a residue based method utilizing the mutual information (MI) statistic was evaluated for intramolecular protein interaction predictions. Similar to the *mirrortree* method, removal of the background signal occurring from common ancestry improves the prediction accuracy. When MI of a third position was incorporated to facilitate the interaction prediction between two contacting positions, the prediction quality was increased. Moreover, in order to identify clusters consisting of three contacting residues, position combinations with the highest significant partial correlation coefficients were extracted and their atomic distances were compared to assess the accuracy of the prediction. Lastly, an analysis was carried out to study the association between PRINTS fingerprints and functionally important interaction sites in seven G protein-coupled receptor families. More than 50% of the functional sites acquired from literature were found to be in close proximity to fingerprint motifs. In the surface patch analysis, over 80% of the functional sites were shown to overlap a motif cluster. Overall, the approaches taken in this thesis have tackled interaction predictions from various directions and keenly provide some insights for protein-protein interactions and evolution.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's policy on presentation of Theses.

Acknowledgements

I would like to take this opportunity to extend my appreciation to a number of people who have played essential roles during the course of my PhD. Without their considerable assistance, it would have been very difficult for the completion of this study.

- Both of my supervisors, Professor Simon Hubbard and Professor Teresa Attwood, for the invaluable guidance and attention that they have been providing me so that I can exceed my own limitations.

- My advisor Dr. Simon Lovell for the inspirational questions during my quarterly meetings to help me tackle problems from various directions.

- My family for the emotional and financial support.

- Also all my officemates in room B1071 for helping me settle in and making me feel at home in the Bioinformatics Corridor.

1. Introduction

Proteins are essential parts of organisms and play a critical role in the primary machinery of cells. Composed of up to 20 different types of amino acid, proteins are linear polymers. However, in order to carry out biological functions properly, proteins fold into 3-dimensional structures and interact with other proteins to form a larger protein-protein interaction (PPI) network. As PPIs are responsible for most cellular functions, the interruption of PPIs often leads to diseases. In many early studies, one gene was studied for a disease. However, most diseases are not monogenic. Furthermore, studying individual enzymes does not provide a full understanding of the cellular organization of functions. Instead, pathways and networks should be studied. As more structures are catalogued, the attention of biological research is turning towards interactions and networks.

1.1. Types of Protein-Protein Interactions

Protein-protein interactions have diverse roles in biology, and differ in many aspects. Different types of PPIs have been described in literature (Nooren and Thornton, 2003), including stable vs. transient, ordered vs. disordered and domains vs. motifs.

PPIs can be classified as stable or transient. Stable complexes are permanent and irreversible (Jones and Thornton, 1996; Tsai *et al.*, 1998), and are associated with proteins that form multimers, which can consist of the same protein (homo-multimer) or different proteins (hetero-multimers). Examples of multi-subunits for stable complexes are hemoglobin and core RNA polymerase. Transient PPIs are temporary in nature and occur in a wide range of cellular process, such as transport and signaling. Proteins in transient PPIs interact to fulfill a specific function, and disassociate after the function is achieved. These proteins include enzyme-inhibitor or signaling-effector complexes. However, specific conditions are generally required to initiate transient interactions. It has been shown that the interface residues for stable interactions tend to evolve at a slower rate than transient interaction interface residues, permitting correlated mutations to occur between interacting partners (Mintseris and Weng, 2005). In contrast, residues

in transient interaction interfaces are more likely to have higher substitution rates, and therefore lead to little or no coevolution between the interaction partners. Moreover, stable and transient interactions were found to have different interface structure properties (Mintseris and Weng, 2003), and can be either strong or weak.

Some interactions are essentially unchanged in structural terms upon complex formation and, are referred to as ordered. Those without defined structures prior to interactions are defined as disordered. Generally, it is thought that proteins need to be in their natural structural form in order to function properly; however, approximately 30% of proteins in the eukaryotic system are classified as disordered (Ward *et al.*, 2004). It was found that disordered proteins are more prevalent in more complex organisms and are potentially involved in the evolution of complexity. In addition, disordered proteins are often found to be involved in cellular functions, such as cell signaling and the regulation of transcription. Meszaros and colleagues (2007) have found that ordered proteins have a higher proportion of hydrophobic residues, whereas disordered proteins contain more polar and charged residues. Furthermore, binding interfaces for disordered PPIs tend to be a single continuous stretch of residues, while the interaction interfaces for ordered PPIs are more fragmented. It has also been suggested that the flexibility of disordered proteins may enable the binding of various interacting partners to ensure functional diversity (Sandhu, 2009). In a recent study, Prakash (2011) analyzed the linear free-energy relationships for ordered and disordered PPIs and found that the binding affinity of ordered and disordered PPIs is linked to their disassociation and association rates respectively.

Interactions can occur at the protein level, domain level, or even at the motif level. Most high-throughput screening methods for identifying PPIs define the interaction at the protein level (i.e. yeast two-hybrid and tandem affinity purification), but the interaction may well be dictated by domain. For instance, when two multi-domain proteins interact, only two of the many domains may be involved. Protein domains are often defined as stable protein subunits that are structurally, functionally and evolutionarily independent of the rest of the protein. Considered as the building blocks of evolution, domains are often duplicated and shuffled to generate proteins with different functions (Bornberg-Bauer *et al.*, 2010). Similarly, motifs are a short segment of a protein sequence that may be functionally or structurally important, and are often

conserved in many sequences. They are usually shorter than domains and can sometimes coalesce to form domains. Short Linear Motifs (SLiMS) are functional protein subunits that are responsible for mediating functions such as protein interaction and post-translational modification. They are involved in biological pathways. These motifs are usually less than 10 amino acids long and are typically found in disordered parts of a protein. Given that SLiMS are defined by a pattern that typically has fewer than five defined positions, they certainly have an advantage over domains, as the flexibility of the motif sequences can facilitate the acquisition of new functionality to proteins (Diella *et al.*, 2008).

1.2. Experimental Methods for Detecting Protein-Protein Interactions

Different experimental methods have been developed to measure PPIs based on their genetic, biochemical and physical properties. While some methods are capable of detecting large protein complexes, other methods detect more focused binary relationships. Physical interactions can be detected by methods such as yeast two-hybrid (Fields and Song, 1989; Ito *et al.*, 2001; de Folter and Immink, 2011), affinity purification-mass spectrometry (Rigaut *et al.*, 1999; Ho *et al.*, 2002; Gavin *et al.*, 2006; Krogan *et al.*, 2006; Volkel *et al.*, 2010) and protein microarrays (MacBeath and Schreiber, 2000; Zhu *et al.*, 2000; Zhu *et al.*, 2001; Chen and Snyder, 2010). In addition, some methods, such as DNA microarray/gene expression (Eisen *et al.*, 1998) and synthetic lethality (Ye *et al.*, 2005; Nijman, 2011), infer PPIs via functional associations. A high-throughput experimental method opens the door to large scale PPI analyses, but in terms of sensitivity and specificity, each method has its own particular strengths and weakness. Since PPI data generated by physical interaction methods are probably more reliable and have lower false positive and false negative rates, these methods are discussed in further detail. Moreover, analyses utilizing high quality experimental PPI data generated using these methods can be found in Chapters 2 and 4.

1.2.1. Yeast Two-Hybrid

The yeast two-hybrid (Y2H, Figure 1.1) system has been the primary experimental method for detecting PPIs since its introduction in 1989 (Fields and Song, 1989). Unlike many other high-throughput methods, it is an *in-vivo* technique in that interactions can be detected in its natural physical state (von Mering *et al.*, 2002). Due to such an advantageous condition, Y2H is very sensitive and can detect transient and unstable interactions. Many large scale studies (Uetz *et al.*, 2000; Ito *et al.*, 2001; Ho *et al.*, 2002; Cornell *et al.*, 2004; Gavin *et al.*, 2006; Krogan *et al.*, 2006; Collins *et al.*, 2007) have successfully generated PPI data from the use of this system. It identifies potential binary PPIs by taking advantage of the delicate property of transcription factors. That is, a transcription activator is first separated into a DNA-binding domain (BD) and a transcription activation domain (AD). The DNA-binding domain is then fused to a "bait" protein and bound to the upstream activation sequence of a reporter gene while the transcription activation domain is fused to a "prey" protein. Upon the re-joining of the two domains, the transcription activation signal can be measured to determine physical interactions between the two proteins under test. Since Y2H is very sensitive, as it is capable of detecting transient interactions, it would not be surprising for a high level of false positives to be acquired. Indeed, it has been estimated that the false positive rate could be as high as 70% (Deane *et al.*, 2002). As the system is not the natural physical state for most proteins (i.e. non-yeast and non-nuclear), it is inevitable that non-biological interactions would be detected by mistake. Also, many bait proteins are capable of activating the transcription without even forming physical contact with a transcription activation domain. Moreover, activation sometimes occurs by random chance.

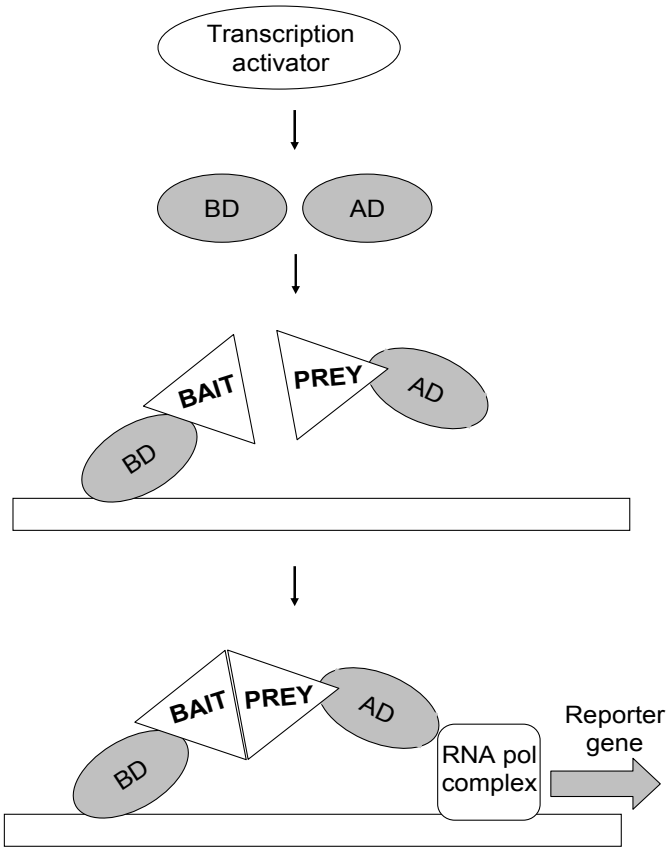


Figure 1.1. The yeast two-hybrid system. It detects interactions based on the principle that interaction between two fusion proteins BAIT and PREY can activate transcription.

1.2.2. Tandem Affinity Purification/Mass Spectrometry

As stated in the Y2H system, proteins rarely act on their own in a physiological system; instead, they often bind to other proteins or macromolecules to form a functional complex. However, it is quite challenging to separate each protein from a protein complex without destroying the interaction signal. To address this issue, the affinity purification technique was developed. The principle of the affinity purification technique is the use of inherent interactions between two proteins. One protein is covalently coupled to a matrix in a column, and afterwards protein extracts are passed over the column. Only proteins that interact and bind to the immobilized protein naturally will be retained; the rest of the proteins will flow through the column. Many different affinity purification methods, such as immunoprecipitation, GST-pulldown and tandem affinity purification (TAP; Puig *et al.*, 2001) have been developed for the

retrieval of protein complexes. Among these methods, TAP is often combined with mass spectrometry (MS) to retrieve high order interactions, and remarkably has become one of the most accurate methods for detecting PPIs. The MS method works by using enzymes such as trypsin to digest proteins to peptides in order to generate peptide fragments which are then separated based on their mass-to-charge ratios. For each protein, the masses of all peptide fragments are compared to a database containing masses of known protein sequences in order to determine the amino acid sequence of the protein. The high sensibility of affinity purification, and the precision of MS, greatly reduce the detection of false positive interactions, and is an effective tool for large scale experiments. Furthermore, TAP-MS has been shown to be very effective in detecting protein complexes in a large scale study (Ho *et al.*, 2002), as 3,617 associated proteins in the yeast proteome were detected in this study. However, prior to the potential interaction taking place, a tag is attached to the end of each protein; therefore, it is possible that these tags might interfere with the interaction and result in indirect interactions. Furthermore, this method is more likely to detect stable interactions rather than transient interactions, as loosely associated proteins may flow through the column during purification.

1.2.3. Protein Microarray

Protein arrays have emerged as a useful tool for the large scale screening of many types of interactions, such as protein-DNA, protein-RNA, protein-protein and protein-ligand interactions. It works by first covalently attaching proteins (probe) to a glass slide at separate locations to form a microscopic array. Subsequently, target proteins are hybridized to the probe proteins for potential interactions. As the first research group to construct a proteome microarray, Zhu and colleagues (Zhu *et al.*, 2000) have reported more than 5,800 yeast proteins, which is equivalent to 58% of the yeast proteome. Indeed, the advantages that protein arrays have are high sensitivity and the ability to test a large number of proteins or even the entire proteome in one experiment. However, the denatured state of the proteins being analyzed can be a critical factor for the success of PPI detections (Liotta *et al.*, 2003). For instance, many antibodies require denaturation of antigens for the linearization of the epitope. However,

PPIs can be interrupted by denaturation. In addition, this can eliminate the interaction between linearized epitopes and antibodies that require natural 3-dimensional structures.

1.2.4. Crystallography

Crystallography is a structural-based approach for studying protein-protein complexes. By solving structures that consist of two binding proteins, PPI interfaces can be detected. X-ray crystallography is the primary method for identifying proteins structures. Solving protein structures by X-ray crystallography involves several steps. First, proteins need to purified and crystallized; afterwards, an intense beam of X-rays strikes the crystal and diffracts into a characteristic pattern of spots. The diffraction data are then analyzed and arranged into a 3-dimensional electron density map which is then used to determine the arrangement of atoms within a crystal. As of July 2011, more than 74,000 protein structures had been solved and housed in the Protein Data Bank (PDB; Rose *et al.*, 2011) database. Indeed, the ability to provide detailed atomic information makes X-ray crystallography the most popular method, especially for identifying interactions for pharmaceutical drug targets (Scapin, 2006). However, the requirement of aligning molecules in the same order places a restriction on the crystallization of flexible proteins, as the electron density map of flexible regions cannot be determined. It has been challenging to crystallize membrane proteins because they are often very low in natural abundance, unstable in detergent solutions and have flexible structures. However, many membrane proteins are important for vital biological functions. For instance, G protein-coupled receptors (GPCRs) are membrane proteins that are involved in signal transduction pathways, and have important medical implications due to the fact that more than 60% of marketed drugs are associated with GPCRs. As a result of the common challenges to crystallize membrane proteins, the first human GPCR structure was not available until 2007. Nevertheless, the availability of crystal structures does facilitate the identification of PPIs tremendously. However, it is time-consuming, and as such, the number of PPIs that can be identified from structure still remains relatively low when compared to other methods.

As different datasets used for the analyses in this thesis were obtained from various sources that contained PPI data generated using the aforementioned high-

throughput techniques, it is necessary to understand their technological differences in order to better interpret the results of the analyses. Particularly for the benchmarking study (Chapter 2), different datasets obtained from different sources that include PPIs derived using different techniques were compared. For instance, the Hake+ dataset contains only structure-based interactions while the Tan+ dataset consists of PPIs derived from various methods, such as Y2H and TAP-MS. Certainly, a large discrepancy in the prediction results was observed for these two datasets. Without an understanding of the biological differences and methods used to generate the datasets, the results could be interpreted inappropriately.

It should be noted that all of these methods may be missing some ordered/disordered interactions that only occur *in-vivo* in particular subcelluar localizations of conditions. Furthermore, high-throughput methods are known to generate data with a significant fraction of false positives (Deane *et al.*, 2002), and distinguishing between direct and indirect (mediated by a intermediate proteins) PPIs can also be quite challenging (Edwards *et al.*, 2002).

1.3. Computational Approaches for Protein-Protein Interaction Predictions

High-throughput experimental methods for detecting PPIs generate a large amount of data. In addition to high false positive and negative rates, experimental methods are generally labour-intensive and time-consuming. Hence, development of computational methods for the prediction of PPIs is necessary to complement experimental interaction detection methods. As computational methods are not restricted to specific experimental conditions, the rate of interaction identification is much faster than experimental methods. Not only can the data generated by computational methods be used to expand the interaction repertoire, but it can also be used to validate the existing protein-protein interactions detected by experimental methods.

1.3.1. Support Vector Machine

One popular method that uses machine learning approaches is the support vector machine (SVM) system which involves training the physiochemical properties of sequence data (Bock and Gough, 2001; Bradford and Westhead, 2005). This method recognizes correlated patterns of interacting sequences and sub-structures in an automated fashion. The resulting patterns often consist of many functional residues in each protein. It has been reported that binding sites often have similar properties which enable the differentiation between these important sites from the rest of the protein (Chothia and Janin, 1975; Jones and Thornton, 1996). Hydrophobic residues tend to cluster at binding interfaces, which is also where polar residues are often found. Other important properties for binding interfaces are residue conservation, shape, solvent accessibility and surface tension. In the study carried out by Bock and Gough (2001), features such as charge, hydrophobicity, and the surface tension of each residue were utilized to capture important characteristics of the proteins in their training dataset. Proteins with unknown interacting partners were then compared using the trained feature patterns, and pairs with similar patterns were deemed as interacting. Consequently, 80% of correct protein interaction predictions were obtained. In another study (Bradford and Westhead, 2005), the authors used an approach that combines SVM and surface patch analysis to predict interface surface patches. In addition to the features used in the Bock and Gough study, conservation, shape and solvent accessibility were also used. As a result, the binding site locations of 76% of the proteins in their dataset were successfully identified.

1.3.2. Gene Neighbour

It has been suggested that genes present in the same order in multiple species are more likely to share similar functions (Tamames *et al.*, 1997); this constitutes the basis of the gene neighbour method. The conservation of gene order is largely observed in bacteria in the form of operons, a group of genes that is expressed and regulated together in a single unit. Two studies (Dandekar *et al.*, 1998; Huynen *et al.*, 2000) have shown that over 60% of co-regulated genes within three bacterial and archaeal genomes were found to interact physically. Similar results were observed for two eukaryotic genomes, *Saccharomyces cerevisiae* and *Caenorhabditis elegans* (Teichmann and Babu,

2002). In these two eukaryotic genomes, interactions were detected in over 90% of the conserved co-regulated gene pairs. However, conservation of gene order in eukaryotic genomes is not generally as prevalent as it is in prokaryotic genomes. Although it has been applied to few eukaryotic genomes successfully (Teichmann and Babu, 2002), great challenges still remain in applying this technique to a large number of eukaryotic genomes.

1.3.3. Gene Fusion

Through the observation that some protein pairs present in different genomes fuse into a single protein in another genome, protein interactions can be deduced with the assistance of gene fusion events. Some efforts have been made to predict PPIs through the use of this method, sometimes termed the Rosetta Stone method, with reasonable success (Enright *et al.*, 1999; Marcotte *et al.*, 1999). For instance, Enright and colleagues (Enright *et al.*, 1999) identified 6,809 putative protein interactions in *E. coli* and 45,502 in *S. cerevisiae*. In these studies, a significant sequence similarity was observed between putative interacting proteins in separate genomes, and in the genome where the fusion event took place, more than half of these proteins were found to be functionally linked. However, this method is limited to proteins with shared domains in distinct proteins (Sprinzak and Margalit, 2001), and appears to offer a rather limited general approach.

1.3.4. Phylogenetic Profile

Early attempts to predict PPIs focused on the presence/absence of proteins in complete genomes (Pellegrini *et al.*, 1999). When two genes are present or absent together in several species, it often indicates that they underwent similar evolutionary processes and therefore are strongly functionally linked. Although one could assume that two genes that are in the same biological process share similar functions, it does not necessarily mean a physical interaction between the two is always the case. A simple phylogenetic profile can be constructed by assigning a score of 1 for a protein presence and a score of 0 for a protein absence. This method is particularly useful when sequence similarity is not evident, since proteins that are part of the same biological process are

more likely to share the same functions regardless of the genomic content. However, the biggest drawback of this method is that a large number of complete genomes are required for accurate predictions. Furthermore, proteins that are common to most organisms, or specific to only one single organism, should be avoided as false predictions can be made through the use of these data.

1.3.5. *Mirrortree*

The *mirrortree* approach (Goh *et al.*, 2000; Pazos and Valencia, 2001) has emerged as a coevolution-based method for predicting PPIs, and variations of this approach have been developed vigorously to improve prediction accuracy (Jothi *et al.*, 2005; Noivirt *et al.*, 2005; Pazos *et al.*, 2005; Sato *et al.*, 2005; Craig and Liao, 2007). The standard *mirrortree* approach is based on the assumption that interacting proteins share similar evolutionary history. Hence, the phylogenetic trees of two proteins can be compared to determine the degree of coevolution, and subsequently a prediction of the interaction can be made. However, due to the complexity and intensive computation power required to compare tree topologies, an indirect approach to obtain intergenic distances is often applied instead. The *mirrortree* method has been shown to be prone to high false positives and negatives, which implies that simply comparing two phylogenetic trees cannot fully capture the coevolutionary signal between two protein families; rather, some underlying background signals could be embedded, and they need to be removed in order to unmask the true coevolutionary signal (Sato *et al.*, 2006). It has been observed that species in a phylogenetic tree are not independent of each other; instead, they are related due to the common ancestry constraint (Felsenstein, 1985; Harvey and Pagel, 1991). Hence, attempts to remove signals related to such phylogenetic relationships among the sequences used to build a phylogenetic tree were made by several groups (Pazos *et al.*, 2005; Sato *et al.*, 2005). Certainly, improvement in the prediction results has been reported. As each of these so-called improved methods was only applied to a limited dataset, the real utility of these methods is still unclear. As a result, they were further examined in the benchmarking study in Chapter 2, using various datasets that were generated by different experimental techniques.

1.4. Protein Interaction Databases

A model organism is a species that has been studied so extensively that its basic biology is well-understood. As a result, it can serve as an ideal reference model for studying other organisms. A substantial amount of protein interaction data generated using model organisms have been published, and still more are in progress (Marcotte *et al.*, 1999; Uetz *et al.*, 2000; Ito *et al.*, 2001; Ho *et al.*, 2002). To date, the most complete genetic interaction map (Costanzo *et al.*, 2010) was constructed for a model organism, *Saccharomyces cerevisiae,* which consists of approximately 75% of all genes in the budding yeast genome**.** The availability of these interaction data has provided many opportunities for examining and characterizing PPIs in many organisms where such relationships are normally not as well understood. Many databases have been developed to house PPIs generated by different methods for different organisms, and are discussed below.

The Database of Interacting Proteins or DIP (Xenarios *et al.*, 2000) is one of the largest protein interaction databases publically available, and contains more than 70,000 interactions from more than 68,000 experiments. These experiments include, but are not limited to: Y2H, protein microarray and TAP-MS. Approximately 23,000 proteins are included in DIP. Experimentally determined protein interactions are curated manually by experts to form a high quality core set, as well as by automated computational approaches. Additionally, three computational methods are implemented to assess the accuracy of protein interactions in the DIP database. The first method, the expression profile reliability (EPR) index (Deane *et al.*, 2002), compares the RNA expression profiles of the proteins under test with a high quality core set of the DIP database. In a similar fashion, the paralogous verficiation method or PVM (Deane *et al.*, 2002) evaluates the patterns of interactions between large-scale PPI datasets and the high quality DIP core set so that potential paralogues are identified. The domain verification (DPV) method (Deng *et al.*, 2002) uses evolutionarily conserved domains and the maximum likelihood estimation method to detect potential domain-domain interactions.

Another high quality PPI database, the MIPS mammalian protein-protein interaction database (Pagel *et al.*, 2005) acquires PPI data only through manual literature curation. The interaction information stored in this database is indeed of

superior quality as only data from individual experiments, rather than data generated from large-scale high-throughput experiments, are included. However, due to the nature of manual human curation, the size of the database is rather limited. As reported by Pagel and colleagues (Pagel *et al.*, 2005), 1,800 interactions with experimental evidence were extracted from more than 370 scientific journal articles for 10 species. This provided interaction information for more than 900 proteins. Unlike many other PPI databases, such as DIP, BIND and MINT, which contain very little mammalian data, more than 90% of the interactions in MIPS were derived from *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*. The availability of interaction data for these species is seen as a great benefit for studies with prospective medical implications.

The Biomolecular Interaction Network Database or BIND (Bader *et al.*, 2003) is a PPI database that contains binary interactions, molecular complexes and pathways. Interacting objects in this database are defined as DNA, RNA or protein. A large portion of the data come from yeast and humans, and as many as 300 of the interactions stored in BIND associate with mammalian proteomes. Most of the interaction data in the BIND database are extracted from PDB, and many large-scale high-throughput interaction methods, such as yeast two-hybrid, mass spectrometry, genetic interactions and phage display. User submission of individual experimental results also facilitated the expansion of this database to more than 6,000 interactions.

The largest interaction resource database at the moment is the General Repository for Interaction Datasets (BioGRID; Stark *et al.*, 2006), which contains more than 246,000 interactions derived from over 31,000 proteins for 17 different species. This database was originally designed to only accommodate interaction data from yeast two-hybrid and mass spectrometry experiments, but it is now extended to also include interaction data from literature and from other high-throughput experiments. As claimed by the authors, BioGRID is now the largest repository of PPI data for both the budding yeast, *Saccharomyces cerevisiae,* and the fission yeast, *Schizosaccharomyces pombe*. A useful feature of the BioGRID web server was the development of a visualization tool, Osprey. This tool allows for annotating interaction data using Gene Ontology (GO).

IntAct (Kerrien *et al.*, 2007) is a database developed and maintained by the European Bioinformatics Institute. Users can freely obtain interaction data and analyze

them by using the tools that are available on the website. All data are extracted from published literature or from user submissions. Disease-focused datasets based on proteins with an association to Alzheimer's and cancer are also available. Additionally, four tools were developed to assist the analysis of PPI data. While ProViz can be used to visualize interactions, MiNe is capable of computing minimal connecting network for a selected protein set. The Targets tool is specifically designed to predict targets for pull-down experiments, and Validator is a PSI-MI semantic validator for various PSI file formats. Moreover, gene ontology annotations are provided for all interactions in the database to help link protein networks with their functional aspects. What sets IntAct apart from most other protein interaction databases (but similar to BIND) is that it contains not only protein interactions, but also other complex interactions, such as DNA, RNA and other small molecule interactions. There are currently more than 229,000 binary interactions in the IntAct database.

MINT or Molecular INTeraction database (Ceol *et al.*, 2009), is a relational database that contains both direct contact (physical association) and indirect contact (association) PPIs. Like many high-quality PPI databases, it takes a literature-based approach to obtain PPI information through professional curators. For quality assessment purposes, a reliability score is assigned to each interaction in the MINT database. Interaction data are extracted from four peer-reviewed journals: FEBS Letters, EMBO Journal, EMBO Reports and FEBS Journal. Because of the fully manual process and the small curation team, the number of interactions in MINT is rather limited. Hence, to increase the size of the database, a less detailed 'light curation' has been applied. This strategy has significantly increased the coverage of interactions that are mediated by modular domains and interactions between viral and host proteins. Although the majority of the interactions in MINT are binary, MINT also supports molecular complexes and biological pathways, and consists of over 89,000 interactions. Interaction data for *Homo sapiens*, and for model organisms such as *Saccharomyces cerevisiae* and *Drosophila melanogaster*, account for most of the interaction entries in MINT. There are 26,517 and 22,338 interactions for *Saccharomyces cerevisiae* and *Drosophila melanogaster*, respectively.

1.5. Protein Structure

Protein structures have been shown to be more conserved than linear protein sequences (Chothia and Lesk, 1986). In particular, active sites of distantly related species have been found to have very similar geometries (Lesk and Chothia, 1980; Chothia and Lesk, 1982; Read *et al.*, 1984); this is thought to be the result of evolution for maintaining functional stability. Protein sequences that share a common ancestor are termed homologues and usually perform the same function. In order to determine homology, linear protein sequences are often compared, and a conclusion is made based on sequence similarity. However, low sequence similarity is often detected among distantly related species and therefore the homology relationship among them can be mistakenly ignored. Due to higher conservation between structures, protein structures can be compared instead of primary sequences in order to overcome such issue.

The three experimental methods that are currently used to solve protein structures are X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy (ER). The main advantage of X-ray crystallography over NMR is that X-ray crystallography is capable of solving structures for large molecules as opposed to small molecules (70 kDa) for NMR. The PDB databse contains the largest collection of 3-dimensional structures for proteins, RNA, DNA and other important biological macromolecules. It was first established in 1971 and has grown from 7 structures to over 74,000 structures in 2011. The structure data contained in the PDB database are generated using X-ray crystallography, NMR, EM and from theoretical modeling. However, the most common method used to determine a structure is X-ray crystallography, which accounts for approximately 86% of all structures in PDB (Figure 1.2).

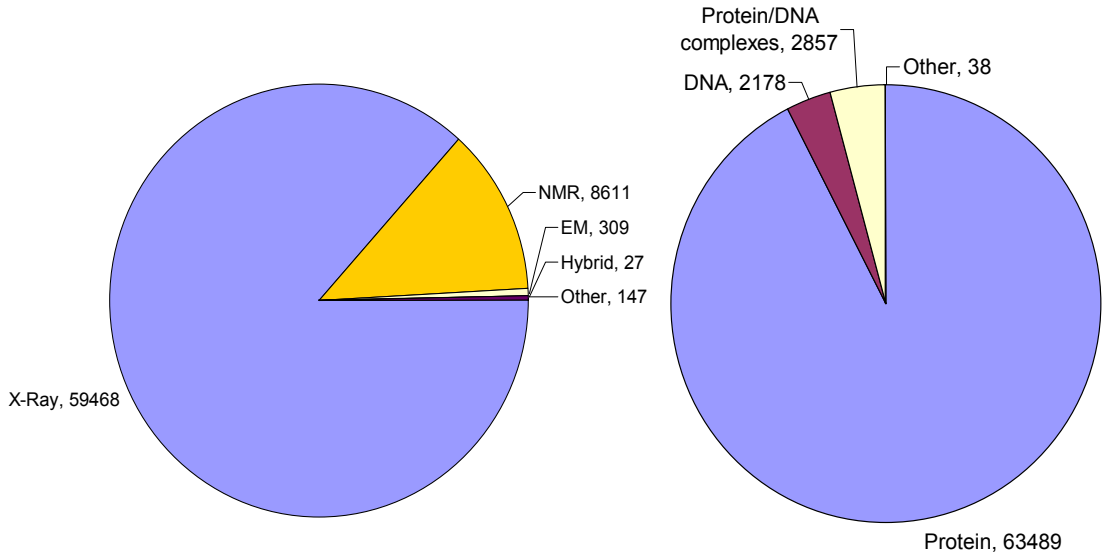


Figure 1.2. Pie charts for different types of structures in PDB. The pie chart on the left is based on different construction methods. The number of structures generated using each method is ranked in descending order: X-ray > NMR > EM > other > Hybrid. The pie chart on the right is based on molecule types. The number of structures generated for each molecule type is ranked in descending order: protein > protein/DNA complexes > DNA > other.

1.6. Homology

Protein sequences that share a common ancestor are termed homologues, and usually perform the same function. There are two types of homologues: orthologues and paralogues. As shown in Figure 1.3, orthologues are homologues that were separated by a speciation event, and are often found performing the same function in different species. However, the other type of homologue, paralogue, was developed through a duplication event. Paralogues are generally found performing similar functions in the same species and can provide insight into how genomes evolve. Although most paralogues are found in the same species, some paralogues exist in different species. For instance, the human hemoglobin and chimpanzee myoglobin are paralogues. For most evolution based studies in Bioinformatics, orthologues should be used instead of paralogues to represent speciation events. However, this is not an easy task. It is possible that homologues detected from different species are paralogues, particularly when multiple copies of duplicate genes are found in the same genome. Many tools

(Altschul *et al.*, 1990; O'Brien *et al.*, 2005) have been developed to help detect and reconstruct homology and some of the most commonly used tools are described below.

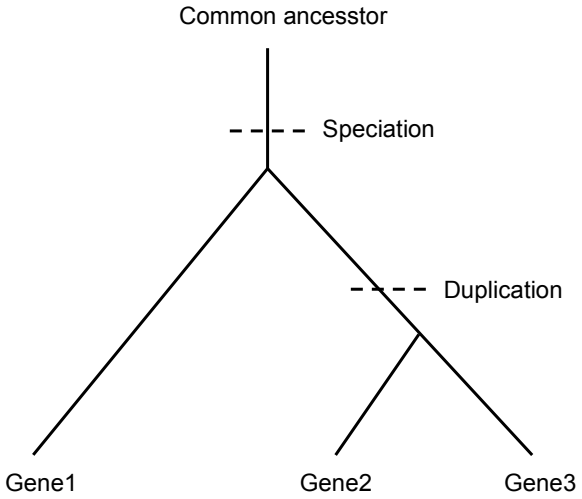


Figure 1.3. Evolutionary tree showing the formation of orthologues and paralogues. A duplication event generated paralogues Gene2 and Gene3, which are co-orthologues to Gene1 due to a speciation event.

The Basic Local Alignment Search Tool, or BLAST (Altschul *et al.*, 1990), compares sequences and determines the level of similarity between them. It is a quick and powerful method to detect homology for uncharacterized sequences. BLAST identifies statistically significant similarities between sequences by first using the dynamic programming technique to obtain an optimal alignment. Generally, two types of alignments are produced: global and local. For global alignments, the best alignment is found by aligning both sequences along their entire lengths, while in local alignments, only the best aligned regions are returned. The global alignment approach uses the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970), and the local alignment approach uses the Smith-Waterman algorithm (Smith and Waterman, 1981).

Multiple sequence alignments (MSAs) are the base of most bioinformatic research. They are widely used to construct phylogenetic trees in order to determine evolutionary relationships among different species; to determine signature patterns; to characterize protein families; and finally, to identify evolutionarily conserved functional residues in protein families. As MSAs have become the prerequisite for many molecular analyses, the need for a method to quickly construct a well-aligned MSA has led to the development of many MSA tools. Two of the most frequently used multiple sequence

alignment methods, CLUSTALW (Thompson *et al.*, 1994) and MUSCLE (Edgar, 2004) were used to construct MSAs in all studies in this thesis; and all default parameters were utilized.

CLUSTALW is the most widely used multiple sequence alignment for DNA or proteins. It uses a progressive approach that starts by building a guide tree. Based on the guide tree, pairwise alignments are subsequently added to the growing MSA from the most similar sequence to the least similar sequence. As it takes a heuristic approach, the alignments cannot be guaranteed to be globally optimal. Nevertheless, alignments generated using CLUSTALW are often used to depict evolutionary relationships between species within a MSA. Another popular method to produce MSAs is the alignment by log-expectation method, MUSCLE. It takes a similar approach as CLUSTALW by adding sequences progressively from the most similar to the least similar. However, what sets MUSCLE apart from CLUSTALW is that after each addition, sequences are re-aligned to ensure better accuracy. It is thought to produce more accurate MSAs than CLUSTALW without compromising the speed. However, in the benchmarking study, as a preliminary analysis (data not shown), both methods were utilized to construct MSAs and little difference was observed. Hence, only MSAs generated using one method (CLUSTALW) are shown in Chapter 2.

1.7. Coevolution vs. Co-adaptation

The general concept of coevolution in biology is that the change of one entity is caused by another. This is broadly applied to many levels of biology, from different co-varying ecological traits to correlated mutations in molecular biology. It is widely accepted that two entities in a coevolutionary relationship apply selective pressures on each other. The first documented use of the term “coevolution” was found in a study carried out by Ehrlich and Raven (1964), which examined the patterns of interactions between organisms with close ecological relationships. This study showed that butterfly groups that were closely related had similar patterns of food plant utilization. Determining such a relationship between a well-studied butterfly group and a newly discovered butterfly group could have facilitated the identification of important biological characteristics of the new butterfly group if the two groups had a similar

feeding pattern. In the context of molecular biology, the phenomenon of functionally linked proteins evolving at similar rates has been observed in ligands and their endogenous receptors (Fryxell, 1996; Pages *et al.*, 1997).

Based on the assumption that interacting proteins evolve in a correlated fashion (Moyle *et al.*, 1994), interacting proteins are under some common evolutionary pressure (coevolution), and compensatory mutations must occur at or near the interface between the interacting partners in order to maintain the interaction (co-adaptation). These could be the result of direct evolutionary pressure to maintain interface structure, biological function in a pathway, or common components of a molecular "machine". Co-adaptation has been observed in both intra- and inter-molecular interactions (Choi *et al.*, 2005; Ferrer-Costa *et al.*, 2007). It has also been proposed to be a possible mechanism for preventing diseases in organisms (Kulathinal *et al.*, 2004). Essentially, when one protein is mutated, the other must also mutate to compensate for the change; otherwise, the physical interaction would be broken and diseases could be induced. As for coevolution, a direct relationship between interaction and evolution rates has been reported (Fraser *et al.*, 2002). Fraser *et al.* also suggested that proteins with more interactors tend to evolve slowly because a larger portion of the protein is involved in the function. Furthermore, some indirect factors have been found to support the association between interaction and coevolution. For instance, the mRNA expression levels for interacting proteins have been found to co-evolve (Fraser *et al.*, 2004), meaning that when the expression level for one protein changes, the expression level of its interacting partner also changes at a similar rate. A graphical representation of co-adaptation and coevolution is illustrated in Figure 1.4.

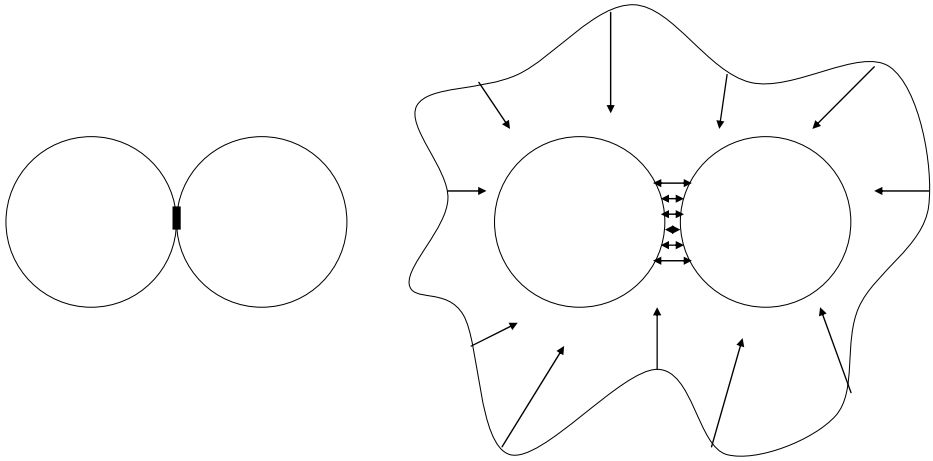


Figure 1.4. The difference between co-adaption and coevolution. Co-adaptation occurs when the interaction interface (vertical bar) is maintained by compensatory mutations between two interacting proteins (circles). For coevolution, interacting proteins are the result of indirect environmental influences which ultimately resulted in similar evolutionary rates.

The phenomenon of co-adaptation is widely recognized, but it remains controversial as to whether this signal can be detected in whole protein sequences, and whether it can be distinguished from the background speciation signal (Hakes *et al.*, 2007; Juan *et al.*, 2008a; Pazos and Valencia, 2008; Kann *et al.*, 2009). While some authors suggest that a coevolutionary signal can be detected (Juan *et al.*, 2008a; Pazos and Valencia, 2008), others have suggested that co-adaptation of physically interacting regions is not (or is only partly) responsible for this signal (Hakes *et al.*, 2007; Kann *et al.*, 2009). Indeed, one study concluded that *mirrortree* is less predictive than other co-functional signals, such as co-expression (Hakes *et al.*, 2007). Kann and colleagues refined this idea further and concluded that both binding neighbourhoods (co-adaptation) and common co-functional constraints (coevolution or correlated evolution) contribute to the *mirrortree* signal.

Although most of the coevolution-based methods are focused on whole protein sequences, many studies (Fodor and Aldrich, 2004; Shackelford and Karplus, 2007; Dunn *et al.*, 2008) have applied the coevolution concept to interaction predictions at the residue level. It is widely accepted that correlated mutations are often found within the

MSAs of interacting proteins. Earlier studies have also found an association between correlated mutations and spatial proximity (Gobel *et al.*, 1994; Olmea and Valencia, 1997). The authors suggested that the destabilizing effect caused by a mutation at one position could be evolutionarily rescued by another mutation occurring in a nearby location. Such a relationship is not only found between residues in the same protein, but also between residues from different proteins, especially interacting proteins (Pazos *et al.*, 1997; Yeang and Haussler, 2007; Burger and van Nimwegen, 2008). This relationship is examined in more detail in Chapter 4 using the mutual information (MI) statistic to measure interactions between two to three residues within a protein.

1.8. Overview

Prior to the prediction of PPIs, it was necessary to carry out a comprehensive benchmarking study (Chapter 2) for a full understanding of PPI methodologies. Hence, the frequently used PPI approach, *mirrortree*, was examined. This was chosen as there were more than 10 papers in the literature, and datasets were readily available or straightforward to generate. Moreover, there was growing controversy surrounding the utility of the method, whether it was suitable for the detection/prediction of co-adaptation (physical interaction) and coevolution (correlated evolution, perhaps owing to common function), and whether species correction genuinely improve the method. It should be noted that this is a whole sequence-based method. Again, it was necessary to know whether this was truly appropriate for predicting PPIs. Hence, a full benchmarking study was needed. A short supplied example of the *mirrortree* approach was used to demonstrate its utility on a relevant example in fibrillin-1 (Chapter 3). As fibrillin-1 is a multi-domain protein, it was decided to study the relationships between the domains. However, poor prediction results were obtained. One possible explanation for such results is that there is insufficient sequence diversity of the orthologues. Given that the benchmarking study had suggested that whole sequence-based methods were limited, a residue-based approach was selected. The mutual information statistic was utilized to compare co-variability between two protein positions, which were represented by two columns in the same MSA. Results are shown in Chapter 4, which shows that this has considerable promise, and indeed a novel methodology used in *mirrortree* was exploited at the residue level to show a significant improvement in

prediction performance. Finally, a more empirical approach, using well-characterized motifs in a system of pharmaceutical interest where there is plenty of data, was tested. Different G protein-couple receptor (GPCR) binding sites were compared to motifs in the PRINTS database, and significant associations between certain GPCR functions and different hierarchical level of PRINTS motifs were found.

2. Benchmarking of *Mirrortree* Based Computational Protein-Protein Interaction Methods

2.1. Aim

The objective of this study was to derive a robust protein-protein interaction prediction approach by benchmarking many existing coevolution based bioinformatic methods in both prokaryotic and eukaryotic systems. To obtain the best combination of techniques for accurate predictions, many aspects that are highly influential to the prediction results, such as sequence diversity, orthologue selection, negative dataset generation, entropy reduction and speciation signal correction were also analyzed. The outcome of this work should provide comprehensive insight into facilitating the prediction of putative interacting protein partners.

2.2. Introduction

Protein-protein interactions (PPIs) are vital for all living cells. To function properly in a biological process, proteins cluster together to form protein complexes. Therefore, identifying protein-protein interactions is the first step toward understanding basic cellular processes. From the earliest methods, such as the yeast two-hybrid system (Fields and Song, 1989; Uetz *et al.*, 2000), affinity purification and mass spectrometry assays (Ho *et al.*, 2002; Gavin *et al.*, 2006; Krogan *et al.*, 2006), to the more recent protein microarray (Kung and Snyder, 2006; Weinrich *et al.*, 2009; Yu *et al.*, 2010), many high throughput experimental methods have been developed to detect PPIs; however, analysis of these experimental datasets shows that they are prone to a high rate of false positives and false negatives. Many groups have compared datasets generated by utilizing these methods and confirmed that only a moderate amount of these so-called high quality datasets actually overlap. This clearly shows that experimental datasets require some quality control or assessment before being used for extensive

analysis. Nonetheless, PPI data generated using high throughput techniques or more focused low throughput studies can be obtained from many repositories, such as DIP (Xenarios *et al.*, 2000), BIND (Bader *et al.*, 2003), MINT (Chatr-aryamontri *et al.*, 2007), IntAct (Kerrien *et al.*, 2007) and the Yeast Interactome Database (Yu *et al.*, 2008).

In addition to experimental techniques, many computational methods were developed for both the prediction of putative PPIs, and as an alternative practice; data resulting from these methods can also be utilized for the validation of experimental results. These methods include text-mining (Hoffmann *et al.*, 2005), structural templates (Aloy *et al.*, 2004; Aloy and Russell, 2004), domain fusion (Enright *et al.*, 1999), mutual information (Gloor *et al.*, 2005; Dunn *et al.*, 2008), phylogenetic profiling (Marcotte *et al.*, 1999; Pellegrini *et al.*, 1999) and coevolutionary based methods (Pazos and Valencia, 2002; Tan *et al.*, 2004; Pazos *et al.*, 2005; Sato *et al.*, 2005). In particular, the similarity of phylogenetic trees, also known as the *mirrortree* approach (Goh *et al.*, 2000; Pazos and Valencia, 2001; Pazos *et al.*, 2005), has recently been drawing much attention.

2.2.1. Mirrortree Approach

The basis of the *mirrortree* approach is that interacting proteins are often under similar evolutionary pressures. It has been observed that many interacting proteins tend to evolve in a correlated fashion (Moyle *et al.*, 1994). That is, when mutation occurs in one protein, compensatory mutation must happen in the interacting partner in order to maintain the interaction. If this does not occur, the mutated sequence will be removed by natural selection due to reduced fitness, and the interaction relationship will be lost. Given that coevolution of interacting proteins is likely to occur, the interaction can be detected by quantifying the similarity between two phylogenetic trees from two protein families. The implementation of this approach starts by first identifying homologous protein sequences, or more specifically, orthologues (homologues that occur from a speciation event among closely related species). Once these orthologous sequences are identified, they can be aligned using various sequence alignment tools, such as ClustalW (Thompson *et al.*, 1994) and Muscle (Edgar, 2004), to build multiple

sequence alignments (MSA). Various distance methods can then be used to compute pairwise genetic distances between all the proteins in an alignment. However, direct comparisons of tree topologies are quite complex and are often omitted. As an alternative, tree branch lengths of two trees are estimated and represented in genetic distances. Finally, a linear correlation relationship between the two protein distance matrices is calculated using the Pearson correlation coefficient, r.

2.2.2. Speciation Signal Correction Methods

As proteins from the same set of species are included in phylogenetic tree constructions, there is a certain degree of inherent similarity between the two trees, regardless of whether the proteins interact. Without having the underlying speciation signal removed, it would be rather difficult to distinguish between interacting and non-interacting proteins as both could result in a positive correlation, owing to the branch lengths being dominated by speciation rather than coevolution. Conversely, if the signal attributed to the underlying speciation process is completely removed, no correlation would be detected between non-interacting proteins, while interacting proteins would still obtain a positive correlation. An example is shown in Figure 2.1, which demonstrates the influence of speciation pressure among the source species on the interaction prediction results.

An improved version of the *mirrortree* approach was developed by Pazos and colleagues (Pazos *et al.*, 2005) to eliminate the background signal that emerged from the underlying speciation events. RNA and protein trees representing the overall evolutionary histories of the species under test are respectively constructed using 16S subunit ribosomal RNAs and proteins that are highly similar to the RNAs. Subsequently, a normalization matrix that consists of protein-RNA ratios obtained from the molecular protein and the RNA trees is generated to convert all distances in the RNA matrix to equivalent protein distances prior to the speciation signal removal step. After the conversion of RNA distances to protein distances, the speciation signal can be removed by subtracting an RNA distance matrix from the equivalent protein matrix. Another approach, developed by Sato and colleagues (Sato *et al.*, 2005) with the intent of reducing the speciation signal between two protein sequences, claims that when using

a projection operator to exclude the speciation signal from distance matrices, the overall predictive power was improved. Additionally, instead of the rRNA distances, the overall average pairwise protein distances, calculated based on all proteins involved in the analysis, were used.

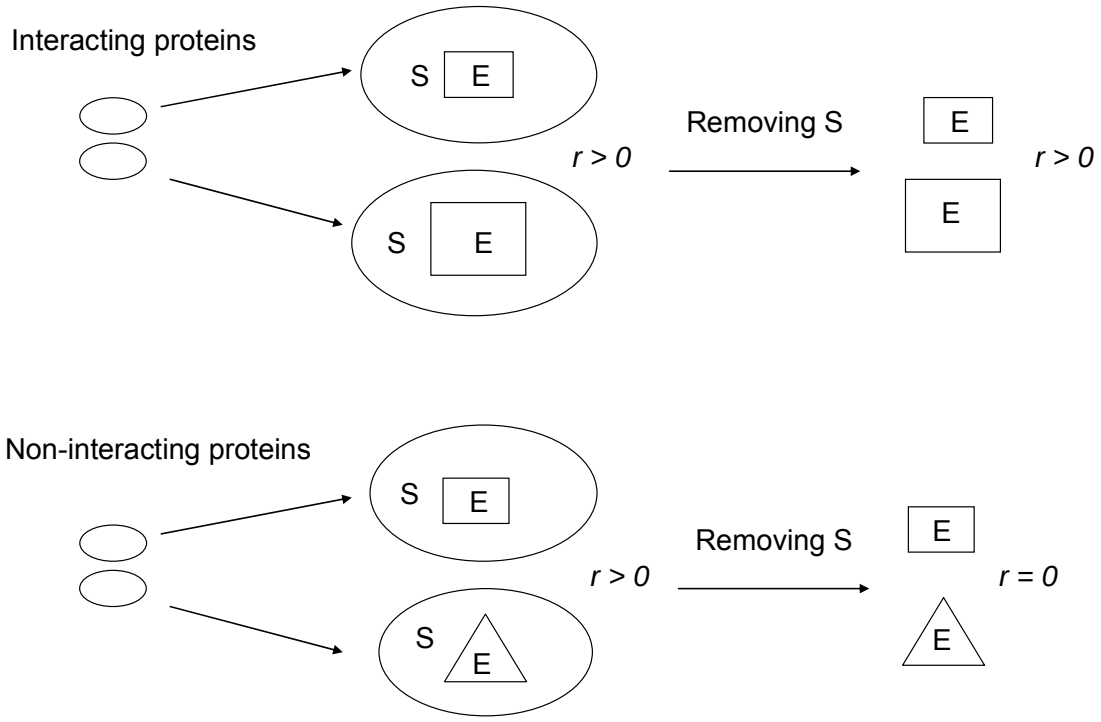


Figure 2.1. Influence of speciation signals on interaction prediction. The ovals represent proteins which contain evolutionary (E) and speciation (S) signals. When S and E are included for the calculation of the Pearson correlation coefficient, *r*, both interacting and non-interacting protein pairs would obtain a positive *r*-score. As shown in the figure, they retain the same shape. However, after removing the speciation signals, interacting protein pairs would still obtain a positive *r*-score, while non-interacting pairs would obtain an *r*-score of 0. As shown in the figure, the final non-interacting proteins are of different shapes.

2.2.3. Orthogonal vs. Non-orthogonal Approaches

Figure 2.2 illustrates a vector representation between the orthogonal approach (developed by Sato *et al.* (2005)) and the non-orthogonal approach (developed by Pazos *et al.* (2005)). Both approaches make use of the idea that raw evolutionary signals can be further separated into true evolutionary and speciation signals. This is represented in Figure 2.2 by the three vectors: V_R (raw evolutionary), V_E (true evolutionary) and V_S

(speciation). In the orthogonal approach, V_S is the projection of V_R such that V_E and V_S are perpendicular (orthogonal) to each other. As for the non-orthogonal approach, since no projection is made, the angle between V_S and V_R is not adjusted; instead, the length of V_S is adjusted. As suggested by Kann *et al.* (2007), the non-orthogonal approach is favoured when one protein mutates faster than the other protein, but still maintains the same tree branch length proportions as the other protein tree. In such a situation, the orthogonal approach would result in a perfect correlation between the two proteins, while different mutation rates would be penalized by the non-orthogonal approach. Another factor that could affect the prediction is that evolutionary pressure does not act uniformly across the whole protein sequence. Hence, some regions (highly conserved) could be under strong evolutionary pressure, while other regions (highly variable) are subject only to speciation influences. When the regions that are not under evolutionary pressure are much longer in one protein than the other protein, the predictive power of the non-orthogonal approach could be compromised due to the fact that speciation vector length is important for the non-orthogonal approach. In this case, the predictive power of the orthogonal approach is not affected. To compensate for the advantages and disadvantages of both approaches, Kann and colleagues implemented an entropy reduction step (ERS). High entropy regions are removed prior to the prediction analysis in an attempt to improve predictive power for both orthogonal and non-orthogonal approaches. After the ERS was applied, an improvement in predictive power was observed.

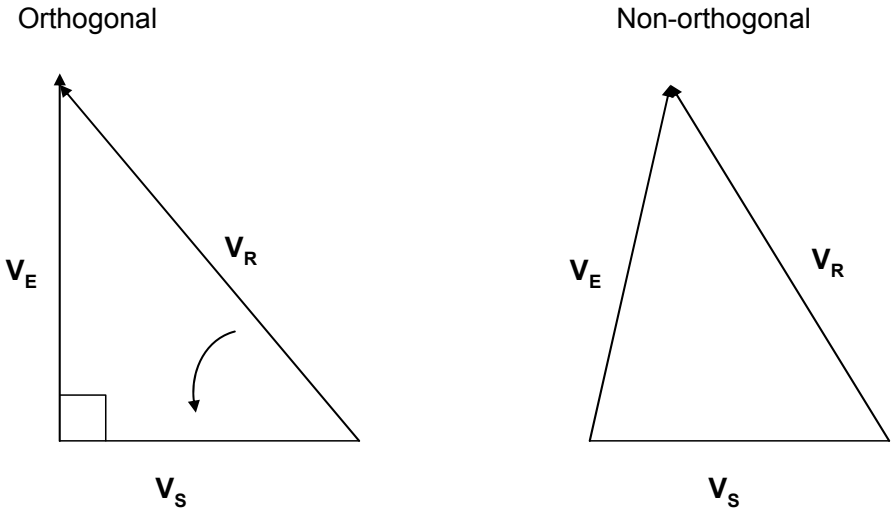


Figure 2.2. A schematic vector representation of orthogonal and non-orthogonal speciation correction methods. V_S is the vector representing the speciation relationships among all species in a MSA while V_R is the raw vector obtained directly from the distance matrix. The desired vector containing only evolutionary signals is signified as V_E.

2.2.4. Benchmarking Study Overview

To examine the *mirrortree* approach, and explore the possibility of identifying the best combination of approaches for PPI predictions, a benchmarking study was carried out in a very systematic and meticulous fashion designed to ensure the quality of the assessment. High quality positive datasets were collected based on crystal structures and multiple experimental evidence. In addition, putative negative datasets were generated using different approaches in order to assess the interaction prediction methodologies. To determine the evolutionary relationship between two proteins, different orthologue selection methods, or resources, were utilized to acquire orthologous sequences for multiple sequence alignment construction. These methods were then evaluated by measuring the degree of coevolution between proteins in different datasets using the *r*- and *Z*-score statistics. Finally, as an overall metric, the area under the Receiver Operating Characteristic (ROC) curve was calculated to assess the predictive power of each method. A schematic diagram showing the overall analysis is illustrated in Figure 2.3.

Although in the original analysis carried out by Pazos *et al.* (2005), the *mirrortree* approach was only tested using prokaryotes, to extend this method to a more general PPI prediction, it was important to also examine the likelihood of using such a method for eukaryotes. However, while most prokaryotes are unicellular and proteins from such organisms often contain only one functional unit (single domain), eukaryotes are generally more complex and contain multi-domain proteins. As conserved regions are often related to functional regions, and would be under more prominent evolutionary pressures, it is possible that restricting the analysis to these regions could increase the prediction accuracy. To examine this phenomenon, the entropy reduction step (ERS) developed by Kann and colleagues, was utilized to remove highly variable regions in MSAs. Indeed, improvement was observed for many datasets, especially when using prokaryotic data. Even without ERS, prokaryotic datasets have shown superior prediction performance than the eukaryotic sets. This suggests that the *mirrortree* approach relies on a similar proportion of important residues in each protein partner, and better predictive power is expected when using these regions. Consequently, this approach is likely more suitable for domain-domain interactions than for multi-domain protein interactions. Indeed, this was the focus of the Kann study, rather than whole protein sequences.

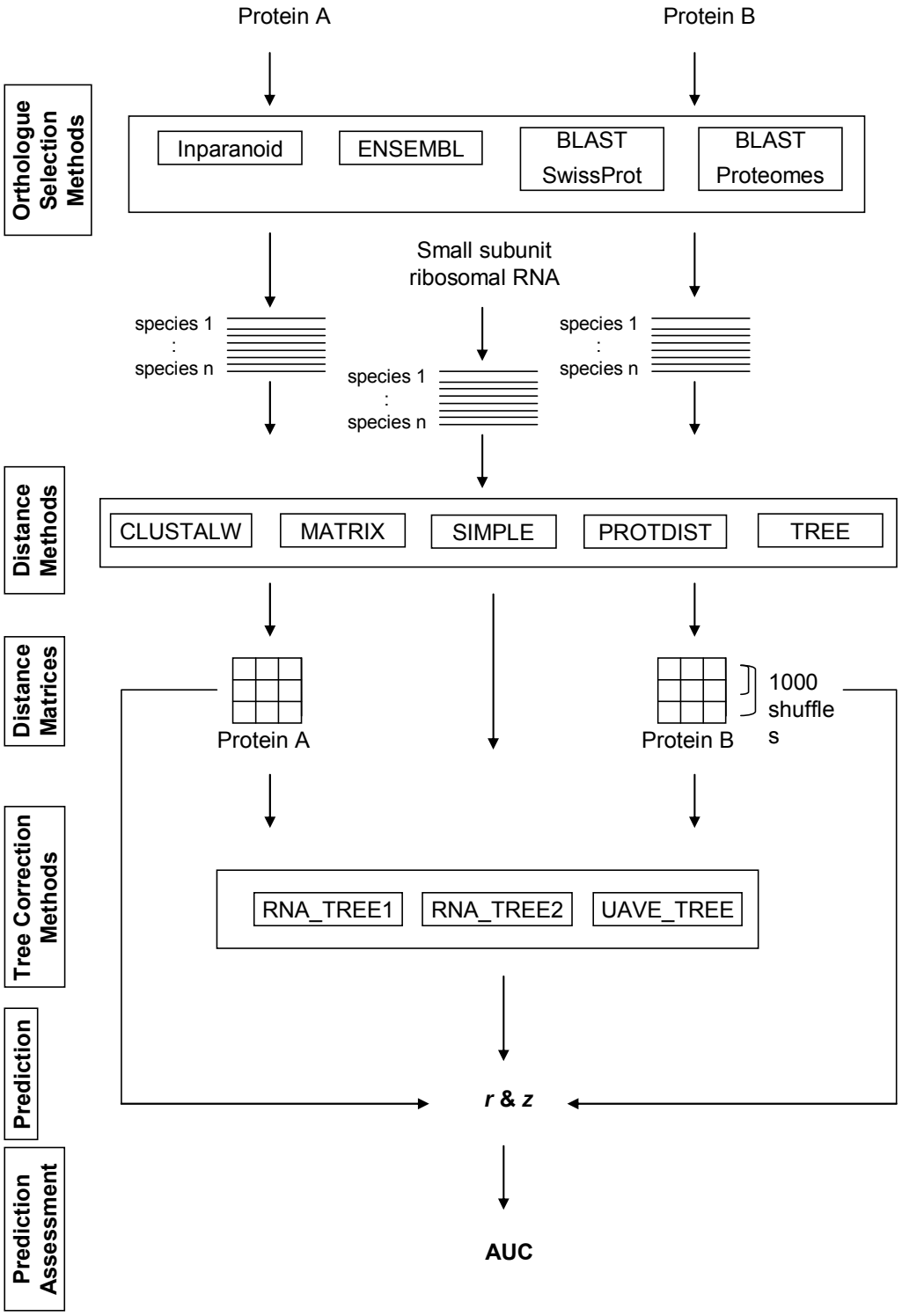


Figure 2. 3. Schematic diagram for the analysis of the *mirrortree* based protein-protein interaction prediction approach.

2.3. Methods

2.3.1. Datasets

In order to better reciprocate tests carried out by other groups, it was necessary to use the same or similar datasets as the original methods. Hence, wherever possible, positive and negative datasets were generated in the same way as they were in the original methods described by Pazos *et al.* (2005) and Tan *et al.* (2004). However, in order to test whether dataset origin affects prediction performance, additional datasets were generated from other sources. Due to the fact that many of the datasets that were obtained from the previous studies were quite small, an effort was also made to generate larger datasets to examine dataset size bias. In particular, Hakes+ and GFP- were generated to be three times the size of the other eukaryotic dataset, Tan+, which was obtained from the Advice website (http://advice.i2r.a-star.edu.sg/doc/sup.php). A description of all datasets used in this study is listed in Table 2.1.

2.3.1.1. Positive Datasets

The Pazos+ dataset was generated based on the *E. coli* interacting protein file, Ecoli20040203, which was obtained from the DIP database (http://dip.doe-mbi.ucla.edu/dip/Main.cgi) with 516 experimentally determined interacting protein pairs. However, many of these protein pairs were derived from self-self and hetero-species interactions. Using this type of interaction data could result in biased predictions; hence, these protein pairs and also proteins with obsolete IDs were removed prior to the analysis. After restricting protein pairs to those with a minimum of 10 orthologous species in common, the remaining 268 interacting protein pairs were further reduced to 154 pairs (consisting of 283 proteins). Multiple sequence alignments containing less than 10 species are commonly thought not to retain sufficient evolutionary signals, and therefore a similar minimum species cutoff has been imposed in other studies (Tan *et al.*, 2004; Pazos *et al.*, 2005).

					Number of protein pairs for each ortholoque selection method (average number of species per MSA)			
Dataset abbreviation	Interaction Status	Organism	Source	Validation Method (s)	Inparanoid	ENSEMBL	BLAST-SwissProt	BLAST-Proteomes
Pazos+	Interacting proteins (coevolution)	*Escherichia coli*	Pazos *et al., 2005*	Annotated in Database of Interacting Proteins (DIP)	154 (24)	-	114 (63)	151 (27)
Pazos-	Non-interacting proteins	*Escherichia coli*	This study	Random selection of two proteins from the Pazos+ set	154 (19)	-	108 (41)	152 (24)
Tan+	Interacting proteins (coevolution)	*Saccharomyces cerevisiae*	Tan *et al., 2004*	Minimum three experimental results	99 (25)	75 (28)	55 (30)	99 (28)
Hakes+	Interacting proteins (co-adaptation)	*Saccharomyces cerevisiae*	Hakes *et al., 2007*	Observed in crystal structures from Protein Quaternary Structure database	297 (23)	296 (29)	297 (24)	297 (29)
Tan-	Non-interacting proteins	*Saccharomyces cerevisiae*	Tan *et al., 2004*	One protein found in the mitochondrial membrane and the other protein found in the nuclear membrane	35 (21)	33 (28)	17 (22)	35 (27)
GFP-	Non-interacting proteins	*Saccharomyces cerevisiae*	This study	Random selection of two proteins from two non-adjacent cellular compartments from the Yeast GFP Localization Database	297 (20)	297 (29)	297 (16)	297 (27)

Table 2.1. Descriptions of positive and negative datasets generated using sequences from *Escherichia coli* and *Saccharomyces cerevisiae*.

A dataset of 110 interacting protein pairs for *Saccharomyces cerevisiae* was obtained from the Advice website (http://advice.i2r.a-star.edu.sg/doc/sup.php), where all interaction data was verified by multiple experiments. The same filtering criteria used for the Pazos+ dataset was applied in order to remove obsolete identifiers, as well as protein pairs with fewer than 10 common orthologous species. Self-self interacting pairs were also removed; consequently, 99 pairs of interacting proteins were retained for the Tan+ dataset.

A third positive dataset, Hakes+, was generated by randomly selecting 297 protein pairs consisting of three sets of 99 non-redundant protein pairs taken from a large collection of structure-based *Saccharomyces cerevisiae* interacting protein pairs (Hakes *et al.*, 2007). The three Hakes+ subsets were generated to ensure that any bias that might arise due to different dataset sizes would be minimized while being compared to the other eukaryotic dataset, Tan+. As our preliminary tests (not shown) revealed very little bias between different dataset sizes, these three subsets were subsequently combined to create a larger dataset consisting of 297 protein pairs. This data was derived, using the rule of 5 or more atomic contacts within 7.5Å of each other, from the Protein Quaternary Structure database (PQS, Henrick and Thornton, 1998) and the BioGRID database (Stark *et al.*, 2006).

2.3.1.2. Negative Datasets

Since there are no experimentally determined datasets for non-interacting protein pairs for *E. coli* or for other organisms, a set of 154 putative non-interacting protein pairs was generated. Following previous research (Pazos *et al.*, 2005), this was achieved by randomly selecting two proteins from the positive dataset, Pazos+, to produce a negative dataset, Pazos-. Protein pairs were included in this negative dataset only if they were not identified as interacting proteins in the positive dataset.

A eukaryotic negative dataset, Tan-, was also obtained from the Advice website (http://advice.i2r.a-star.edu.sg/doc/sup.php) in the same way as the Tan+ dataset. The criteria used by the authors (Tan *et al.*, 2004) to ensure that the proteins in this negative set were not interacting was based on the premise that one protein was found in the

mitochondrial membrane and the other protein was found in the nuclear membrane. Although all proteins in the Tan- set are membrane proteins, membrane proteins are probably under-represented in the other datasets. Due to the membrane protein bias, it is possible that proteins in Tan- might be under different evolutionary pressures when compared to the other datasets.

GFP-, an additional eukaryotic negative dataset made up of 297 putative non-interacting protein pairs, was created using data from the yeast GFP fusion localization database (Ghaemmaghami *et al.*, 2003). The *Saccharomyces cerevisiae* fusion library in the database was created by tagging each open reading frame with a high-affinity epitope, and subsequently monitoring expression via immunochemistry from its natural location. After obtaining the yeast expression database from the GFP fusion localization website (http://yeastgfp.yeastgenome.org/), proteins shown to express in two non-adjacent yeast subcellular compartments were selected to create the GFP- dataset.

2.3.2. Orthologue Selection Methods

As different orthologues lead to different compositions of MSAs, and result in different predictions, it is essential to identify a good method for selecting real orthologues in order to make accurate predictions. Orthologues in this study were generated from four different sources or methods for all eukaryotic proteins, and three for the prokaryotic proteins. This is because only eukaryotic data is available in the Ensembl Compara Homology database. These methods include: 1) Inparanoid, 2) ENSEMBL, 3) BLAST-SwissProt, and 4) BLAST-Proteomes. Example trees consisting of orthologues obtained via different orthologue selection methods are shown in Figure 2.4 for a representative *E. Coli* protein, and Figure 2.5 for a representative yeast protein.

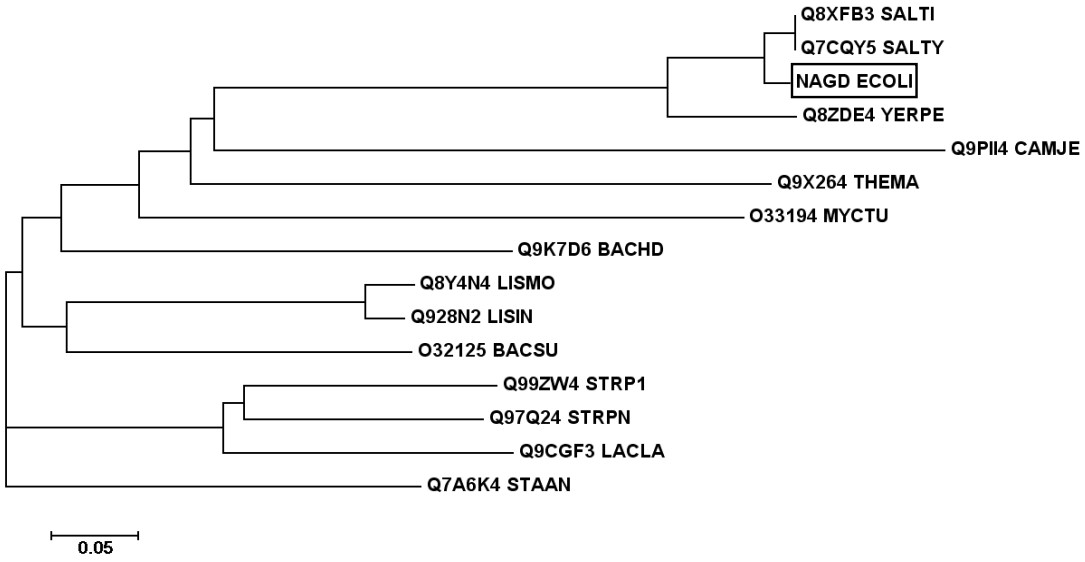
The Inparanoid program (O'Brien *et al.*, 2005), version 5.1, was obtained from http://inparanoid.sbc.su.se/cgi-bin/index.cgi to generate orthologues for the Inparanoid method. This program implements the reciprocal best BLAST hit (RBH) approach and works by first using BLAST to search a querying database comprised of multiple complete proteomes with an E-value cutoff of 10^{-5}. The subject database is then reverse-searched to obtain the top best-best hits, which are considered putative orthologues. The

E. coli proteome, as well as 43 fully-sequenced bacterial proteomes, were obtained from the integr8 database (http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do), and respectively used as the subject database and query databases. For the eukaryotic datasets, *Saccharomyces cerevisiae* and 34 complete eukaryotic proteomes were also obtained from the integr8 database, and were subsequently utilized as the subject and query databases to obtain orthologues.

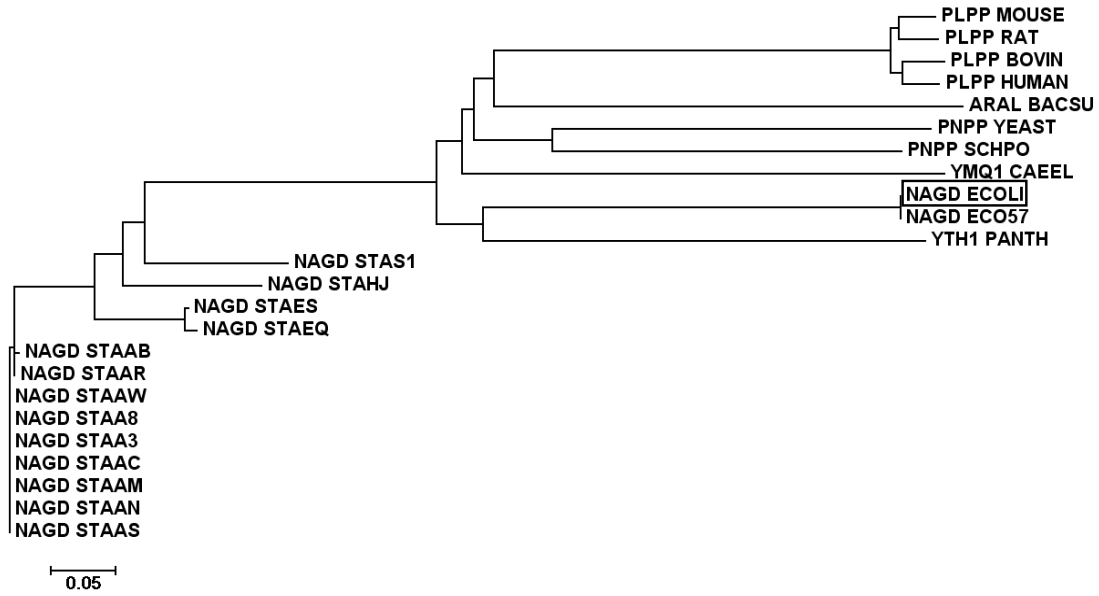
The ENSEMBL method obtained eukaryotic orthologous sequences from the Ensembl Compara Homology database (Flicek *et al.*, 2008), a very high quality resource for orthologues. Although it would be ideal to also obtain orthologues from the same species used for the Inparanoid calculations, the Ensembl Compara Homology database is unfortunately limited to only eukaryotic species. Furthermore, only 29 species in the Ensembl Compara Homology database overlapped with the 35 species used for the RBH approach, so orthologues were only extracted from these overlapping proteomes. When multiple putative orthologues emerged, the one with the highest percentage identity to the yeast protein was selected.

In addition to the more sophisticated methods described above, two simple top BLAST approaches, BLAST- SwissProt and BLAST-Proteomes, were implemented. The BLAST-Proteomes method used a database consisting of 43 complete eukaryotic proteomes (the same species as were used for the Inparanoid method), while the BLAST-SwissProt method used the SwissProt (release 10.0) database (http://www.ebi.ac.uk/swissprot/). Candidate proteins were searched using BLAST against each database. Each top search result with an E-value lower or equal to 10^{-5} was deemed an orthologue, and hence included for the analysis.

A) Inparanoid



B) BLAST-SwissProt



C) BLAST-Proteomes

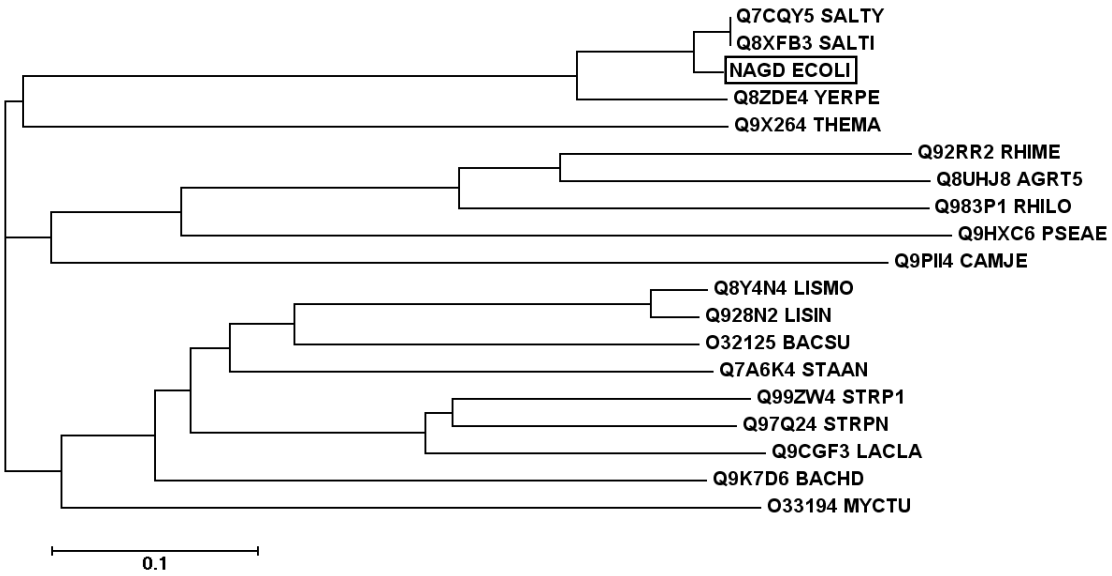
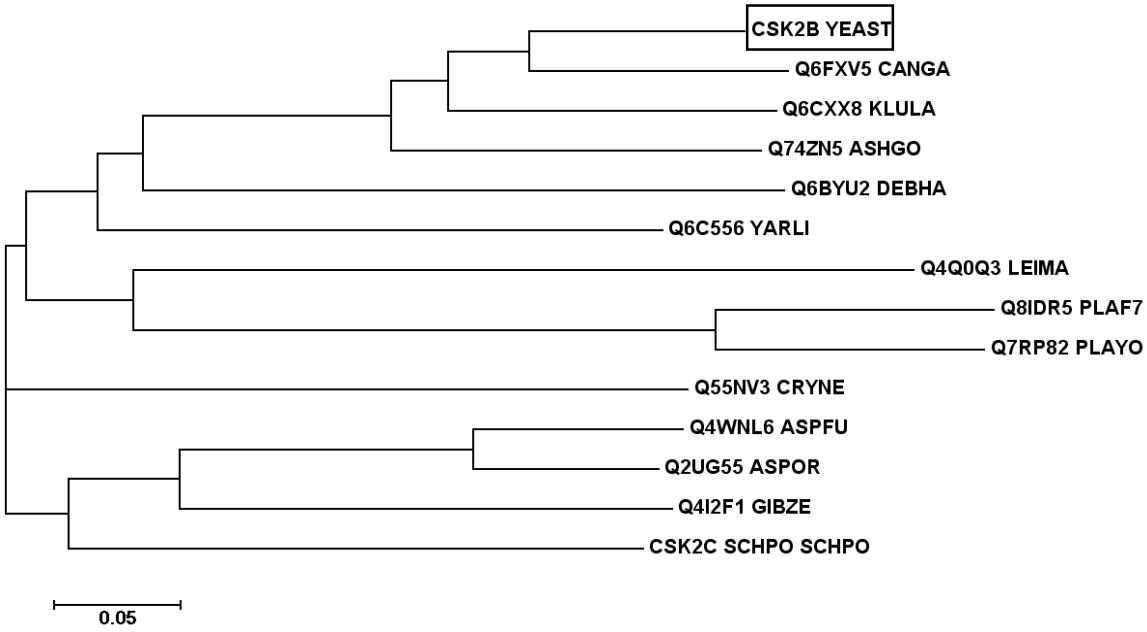
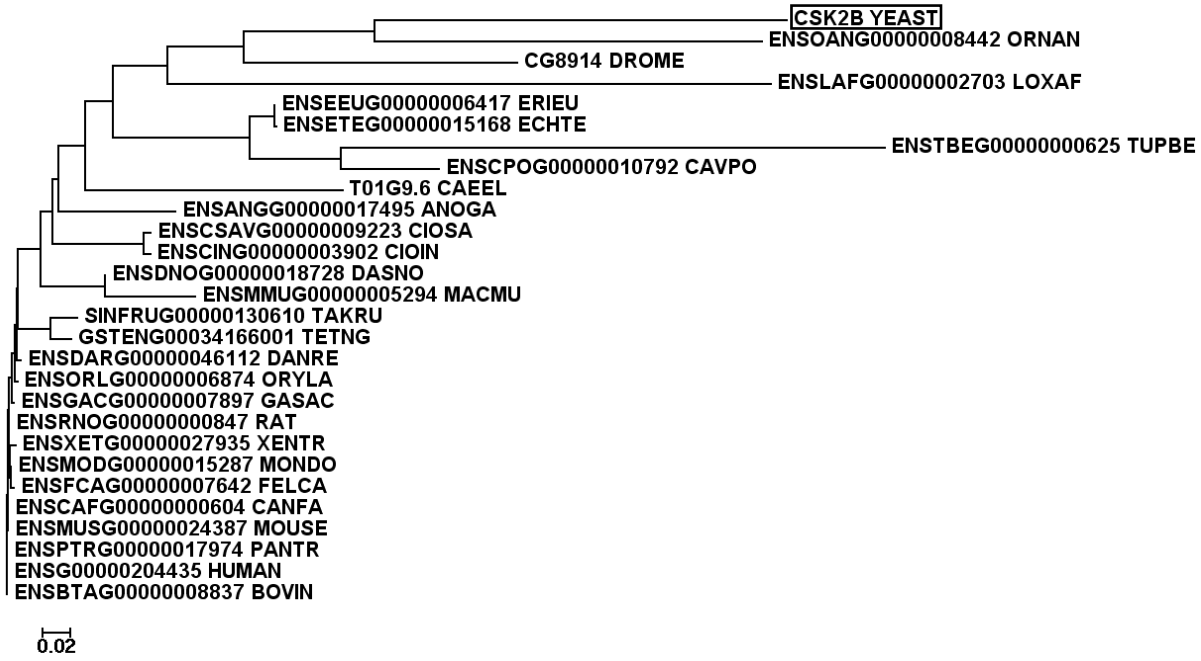


Figure 2.4. Phylogenetic trees for protein NAGD_ECOLI (outlined in boxes), showing orthologues that were obtained using 3 different methods. These methods were A) Inparanoid, B) BLAST-SwissProt and C) BLAST-Proteomes.

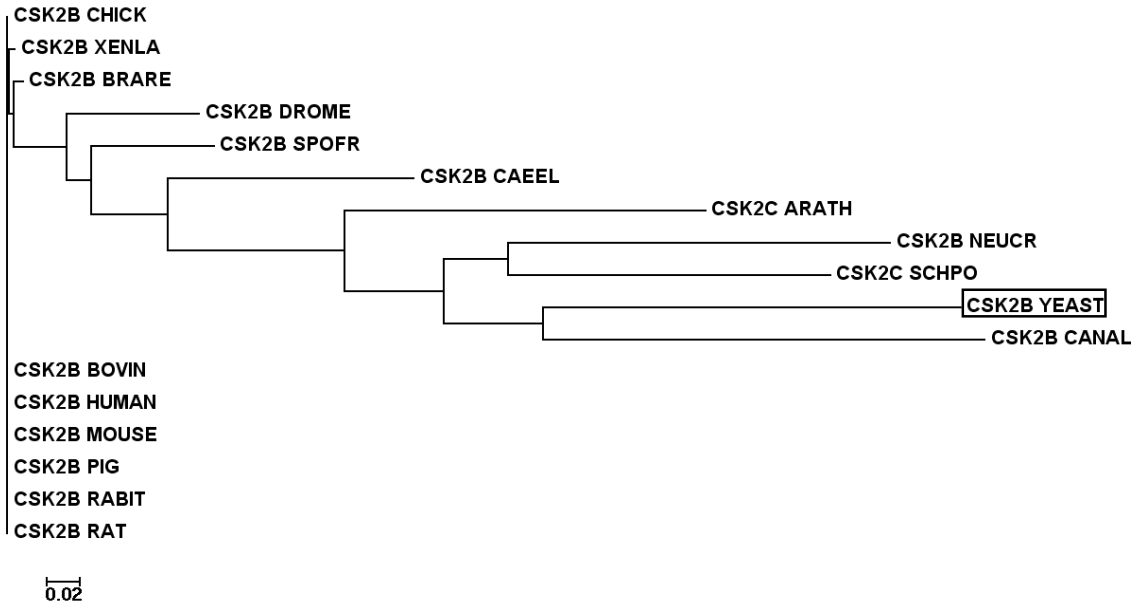
A) Inparanoid



B) ENSEMBL



C) BLAST-SwissProt



D) BLAST-Proteomes

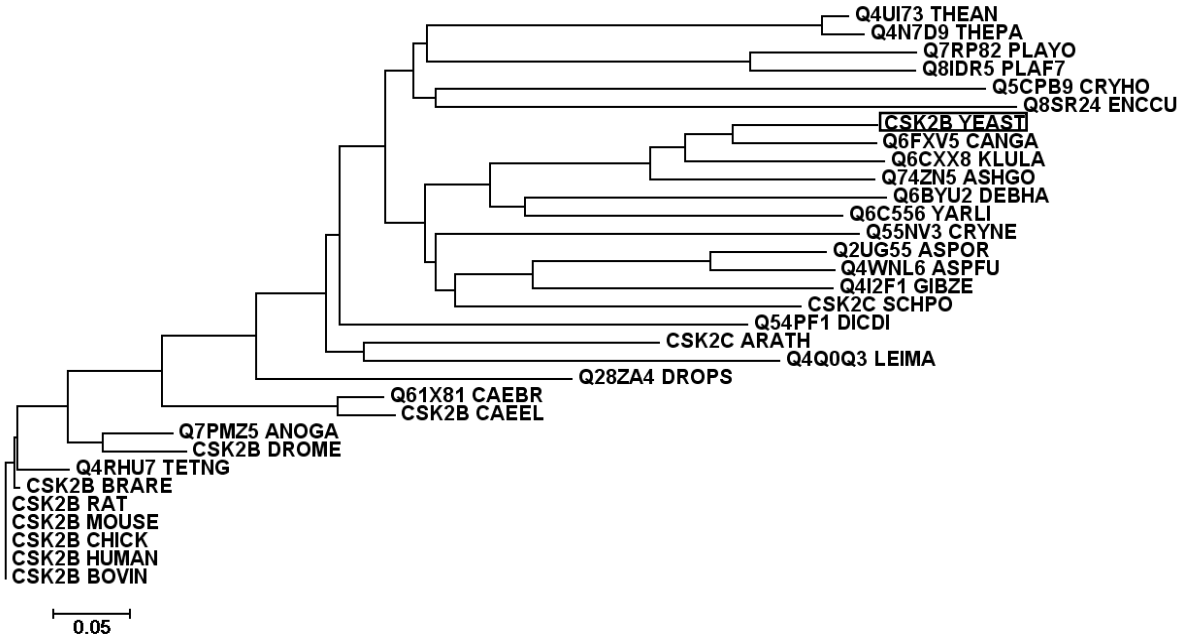


Figure 2.5. Phylogenetic trees for protein CSK2B_YEAST (outlined in boxes), showing orthologues that were obtained using 4 different methods. These methods were A) Inparanoid, B) ENSEMBL, C) BLAST-SwissProt and D) BLAST-Proteomes.

2.3.3. Genetic Distance Methods

Orthologues derived from common species for each protein pair partner were aligned using ClustalW (Thompson *et al.*, 1994) with default parameters. MSAs with nine or fewer sequences were removed from the analysis. Five genetic distance methods, 1) ClustalW, 2) PROTDIST, 3) SIMPLE, 4) MATRIX and 5) TREE, were then applied to the remaining MSAs, and the distance matrices were used to infer their coevolutionary relationships.

For the ClustalW method, pairwise alignment scores, which measure the degree of similarity between two sequences in a MSA, were obtained from the ClustalW output log file. These percentage identity scores were used as a measure for genetic distance between two sequences.

The Protdist program (part of the Phylip package) that implemented the Jones-

Taylor-Thornton model (JTT; Jones *et al.*, 1992) was utilized to calculate evolutionary distances between two sequences for each MSA in the PROTDIST method. The basis of the JTT model is the observation of amino acid exchanges in a protein sequence. This model generates distances in scaled units of expected fractions of amino acids changed, and is set as the default model in the Protdist program.

The SIMPLE method implemented a scoring system that measures the number of identical non-gapped positions between each two sequences in a MSA. For each two sequences, *i* and *j*, from the same MSA at each amino acid residue position, a score of 1 was assigned if there was a match between the two amino acid residues, given that neither one of the residues was a gap. A score of -10 was assigned when it was the opening of a continuous stretch of gap residues, and a score of -1 was assigned when it was not an opening gap. Another method, MATRIX, was derived from the method implemented by Kim and colleague (Kim *et al.*, 2004). This method assigned a penalty score of -10 for a residue if it was an opening of a continuous stretch of gap residues, and a score of -1 when it was not an opening gap. Unlike the SIMPLE method, BLOSUM62 was utilized as the amino acid substitution matrix in order to measure the degree of similarity between two amino acid residues at each position. Finally, SIMPLE and MATRIX distances between sequences *i* and *j*, d_{ij}, were calculated as follows:

$$d_{ij} = 1 - \left(\frac{2S_{ij}}{S_{ii}.S_{jj}} \right)$$

where S_{ii} and S_{jj} are the self-self alignment scores and S_{ij} represents the pairwise distances between sequence *i* and *j*.

Following the method used by Pazos and colleagues (Pazos *et al.*, 2005) to generate distance matrices, one neighbour-joining tree for each MSA was constructed using ClustalW. All tree branch lengths separating each species were summed up to generate a distance matrix. Although phylogenetic trees are often utilized to depict the evolutionary history of a set of species, it is not necessary that the constructed trees match the known species trees to represent true sequence divergence (Graur and Li, 2000; Nei and Kumar, 2000) as different protein trees might have different evolutionary

histories. When a protein tree is different from the true species tree, both the tree branch lengths and topology can be quite different from that of the species tree. As ClustalW constructs a guide tree before progressively aligning sequences to the hierarchy in the guide tree, it is possible that the accuracy of the phylogenetic tree could be improved by using more sophisticated methods (Nelesen *et al.*, 2008) to generate guide trees. Other phylogenetic tree construction methods, such as maximum parsimony and maximum likelihood, have been shown to produce trees that are closer to the true model of evolution, but are more computationally intensive to run.

2.3.4. Speciation Signal Correction Methods

Three methods, RNA_TREE1 (Pazos *et al.*, 2005), RNA_TREE2 (Sato *et al.*, 2005) and UAVE_TREE (Sato *et al.*, 2005), were implemented in order to remove the underlying speciation signal within each MSA.

The RNA_TREE1 approach follows the same approach described by Pazos *et al.* (2005). A tree of life was constructed using 16S small subunit ribosomal RNA sequences obtained from the European ribosomal RNA database (http://bioinformatics.psb.ugent.be/webtools/rRNA/). To construct the eukaryotic tree of life, 18S small subunit rRNA sequences were extracted from the NCBI website (http://www.ncbi.nlm.nih.gov/) instead. All phylogenetic trees were then constructed using the TREE method described in the previous section. To ensure compatibility of all the data, a normalization distance matrix was generated for each of the prokaryotic and eukaryotic sets, and used to convert RNA to protein distances. Proteins that are most similar to the RNA tree, determined by the Pearson correlation coefficient (r), were then used to construct the protein-RNA normalization matrix by computing a protein/RNA distance value for each inter-species comparison. All RNA distances were then converted to equivalent protein distances using these conversion ratios. Subsequently, they were subtracted from the protein distances generated by the TREE method.

The RNA_TREE2 and UAVE_TREE methods were applied using the same approaches developed by Sato and colleagues (Sato *et al.*, 2005). Projection operators were used in an attempt to eliminate the background signal that resulted from

phylogenetic relationships among the species used for the analysis. Essentially, two distance matrices were first converted to phylogenetic vectors by organizing all the values in each matrix in the same order, and were then transformed into the projection vectors ρ_{vi} and $\rho_{vj.}$ The computation of ρ_{vi} is shown below:

$$\rho_{vi} = V_i - V_{unit}\langle V_{unit} \mid V_i \rangle$$

where V_i is the phylogenetic vector for protein *i* and V_{unit} is a unit vector. $\langle V_{unit} \mid V_i \rangle$ represents the inner product of a unit vector and the phylogenetic vector *i*.

For the RNA_TREE2 method, the unit vector, V_{unit_16S}, was obtained using 16S rRNA sequences. However, in the UAVE_TREE method, the average of all phylogenetic vectors was used as a unit vector, V_{unit_AVE}, to determine projection operators.

$$V_{unit\_16S} = \frac{V_{unit\_16S}}{\sqrt{\langle V_{unit\_16S} \mid V_{unit\_16S} \rangle}}$$

$$V_{unit\_AVE} = \frac{1}{m}\sum_{1}^{m} \frac{V_{unit\_AVE}}{\sqrt{\langle V_{unit\_AVE} \mid V_{unit\_AVE} \rangle}}$$

The prediction about the two proteins, *i* and *j*, can then be made by calculating the inner product of the two projection vectors, ρ_{vi} and ρ_{vj}

$$\rho_{ij} = \langle \rho_{vi} \mid \rho_{vj} \rangle$$

2.3.5. Protein Interaction Prediction and Performance Assessment

The coevolutionary relationship between two proteins was measured by the Pearson correlation coefficient, *r*, as follows:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

where *x* and *y* are the corresponding distance values from the two matrices that are being compared. N represents the total number of *x* or *y* – in this case, the number of sequences in the MSAs of the proteins being compared. The higher the *r*-score, which must take a value between -1 and 1, the more likely that the two proteins are coevolving, although that does not necessarily imply a physical interaction.

To determine the significance of *r*-scores, the *Z*-score approach developed by Kim and colleagues (Kim *et al.*, 2004) was implemented. For the estimation of the background mean, r_{mean}, and standard deviation, σ_r, used for the *Z*-score calculation, the sequential order of one of the MSAs for each protein pair was randomly shuffled 1000 times. A Z-score was then computed using the following equation:

$$Z = \frac{r - r_{mean}}{\sigma_r}$$

Z-scores of 1.64, 2.33 and 3.09 correspond to *p*-values of 0.05, 0.01 and 0.001, respectively. In addition to *Z*-scores and the Z-score-derived *p-values*, the significance of *r*-scores can also be determined by comparing r_{mean} values from the randomizations and from the predictions to calculate *p-values* using standard t-tests.

The accuracy of the predictions was then measured by computing the sensitivity (SN) and specificity (SP) using the following equations:

$$SN = \frac{TP}{TP + FN} \qquad SP = \frac{TN}{TN + FP}$$

where TP, FN, TN and FP, respectively, denote true positives, false negatives, true negatives and false positives. Receiver operating characteristic (ROC) curves were computed to estimate the sensitivity and specificity for each method at different *r* score thresholds. The area under a ROC curve (AUC) was then generated using the trapezoid rule to assess the predictive power for each method. An AUC value of 0.5 suggests a random prediction.

2.3.6. Entropy Reduction

Although the *mirrortree* approach was originally designed to predict PPIs using whole sequences, it is possible that functionally or structurally important regions were isolated to be used for the predictions. It is widely believed that highly conserved positions among orthologous sequences derived from different species are more likely to be functionally important and therefore less likely to be affected by the underlying speciation signals that are embedded within MSAs (Kann *et al.*, 2007). Several attempts have been made to increase the accuracy of PPI predictions by utilizing residues from specific regions such as surface and interface (Hakes *et al.*, 2007), or those that are simply conserved (Kann *et al.*, 2007). The entropy reduction step (ERS), as described by Kann and colleagues (Kann *et al.*, 2007), was implemented in order to eliminate regions outside of these conserved domains in a MSA. In this method, the standard Shannon entropy *H* was computed via the use of the al2co program (Pei and Grishin, 2001). It should be noted that the entropies calculated by the al2co program are scaled, and therefore are not directly comparable to raw Shannon's entropies. An al2co entropy cutoff of 1.9 was used to remove all columns with entropy above the cutoff value. In addition to using the entropy cutoff, a second approach involving removing all of the gapped columns in a MSA was taken to eliminate highly variable columns. Subsequently, genetic distance matrices and the Pearson correlation coefficients were computed to determine the level of coevolution for all protein pairs in various datasets. Furthermore, AUC scores were also calculated to assess the overall effect of restricting the analysis to conserved domains.

It should be noted that structural domains could also be utilized as an alternative (or potentially as an addition) to the ERS. Evolutionary domains (which may also be structural) are fundamental evolutionary "units" of a protein. In the analysis of co-occurring domain sets in yeast proteins, Cohen-Gihon *et al.* (2007) found that co-occurring domains (whether sequence-based or structural-based) have similar functions and are likely to co-interact. This means that if the coevolutionary signal is there because domains interact, whole protein sequences, especially those related to eukaryotic systems, should not be used. However, while some groups have studied domain-domain interactions using the *mirrortree* approach and come to results that no enormous difference was observed (Hakes *et al.*, 2007), others (Kann *et al.*, 2009) found improvements. To further investigate domain-domain interactions, a multi-domain protein, fibrillin-1, was studied. The results are shown in Chapter 3.

2.3.7. Optimal Protein-RNA Ratio Experiment

An experiment was designed to determine the optimal protein-RNA ratio for the conversion of genetic distances from RNA to protein space. Essentially, an artificial protein-RNA normalization matrix filled with a single value, ranging from 0.01 to 5, was utilized for the interaction predictions made using the RNA_TREE1 method. The ratio was systematically increased by 0.1, and the Pearson correlation coefficient was re-generated based on the new artificial normalization matrix. The results shown in Figure 2.10 contain *r*-scores up to 3 as the trend is quite consistent after 3.

2.3.8. Sequence Diversity Experiment

In order to determine whether MSAs with different sequence diversities would affect interaction prediction results, a controlled experiment was designed to progressively either increase or reduce the overall sequence diversity by selectively removing species from the MSAs.

Ten interacting protein pairs and ten non-interacting protein pairs were selected from the Pazos+ and Pazos- datasets. All the selected MSAs contained 36 sequences initially. The number of sequences in each MSA was reduced one at a time until only 3

sequences were left. The reason for this is that this is the minimum number of sequences required to build a phylogenetic tree. For each protein pair, distance matrices were constructed by summing up all tree branch lengths in each tree, according to the TREE method. Next, all the genetic distance values associated with each species were summed up. For the decreasing sequence diversity test, the species with the highest overall average distance between two proteins was removed from the phylogenetic trees for both proteins. Alternatively, the sequence with the lowest average overall distance was removed for the increasing sequence diversity test. The *r*- and *Z*-scores were recalculated after each sequence removal from both MSAs. Using an average distance between protein pairs across two MSAs to measure the average sequence diversity ensures that a single species is removed each time, and guarantees that the same number of sequences is in each MSA. An illustration of the experiment for decreasing sequence diversity is shown in Figure 2.6.

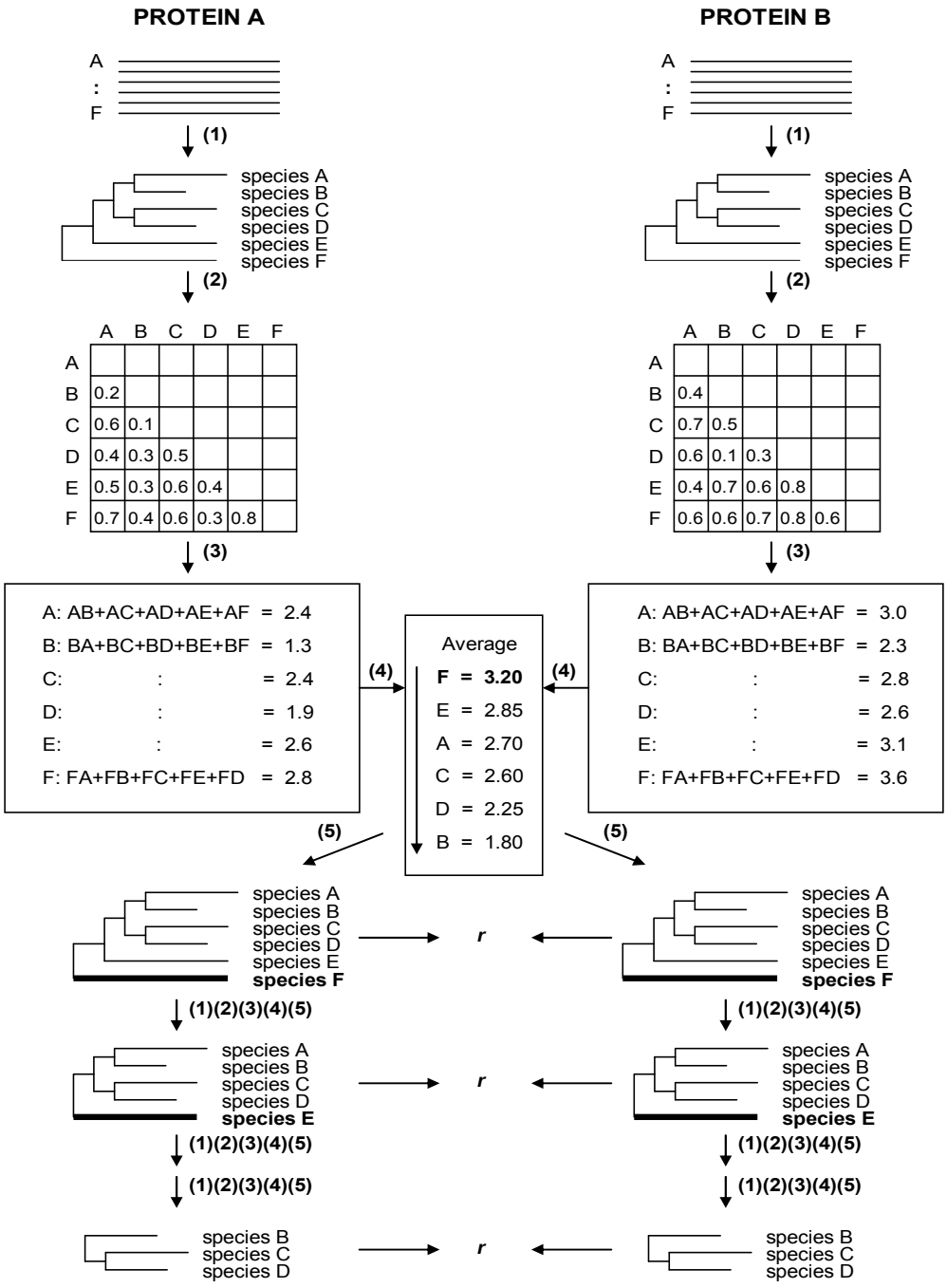


Figure 2.6. Example of a decreasing sequence diversity experiment. 1) A neighbour-joining tree is constructed for Protein A and B. 2) All pairwise tree branch lengths are determined for species A to F. 3) Sum of all tree branch lengths associated with each species is calculated. 4) An average sum of tree branch lengths between Protein A and B for each species is calculated. 5) The sequence associated with the longest average tree branch species is removed from protein A and B. Steps 1 to 5 were repeated to remove the sequence with the next highest tree branch length, one sequence at a time, until only three sequences were left in a MSA. The Pearson correlation coefficient is also calculated after one sequence is removed from each MSA to determine the degree of interaction.

It is possible that the source database for the generation of orthologue sets could have an effect on the experiment, although it probably does not change the general conclusion. In an extreme case, if the query species is very closely related to another species, it could be the first species to be removed by the increasing diversity test. Conversely, if the query species is particularly divergent from the rest of the species, it could also be removed first by the decreasing diversity test. For example, the query sequence CSK2B_YEAST in Figure 2.5B) has a relatively long overall tree branch length from root to the terminal node, and would probably be eliminated early in the case of removing the most divergent species. The main reason is likely due to the fact that many of the species in the Ensembl database are vertebrate, where yeast could be considered as an outlier. Although the sequence diversity calculations for the individual alignment could be affected, it does not change the overall trend. The influence could also be protein family specific, where some are affected and some are not. As this experiment was designed to test whether one protein family interacts with another, as well as the association with sequence diversity, it does not make much difference what species the removed sequence comes from. For example, as shown in Figure 2.8, the increasing diversity set consistently has better predictive power for interaction, even after removing just one sequence. Although the choice of database, method to align the sequences, and all the other parameters are important for the *mirrortree* approach, they are secondary to this experiment.

2.4. Results and Discussion

2.4.1. Datasets

Most studies often only utilize one dataset from one domain of life for their analysis; however, when tested by other groups using a different dataset, the results sometimes differ or even contradict the original findings. Hence, in order to comprehensively benchmark *mirrortree* based methods, several datasets (Table 2.1) were generated for both positive and negative examples from prokaryotic and eukaryotic proteomes. Where possible, the same number of protein pairs was used throughout the analysis. However, different orthologue selection methods resulted in different orthologues, and with the minimum requirement of 10 common species, it was

impossible for some protein pairs to meet the criteria. As such, they were excluded from the analysis. Nevertheless, the number of protein pairs varies slightly, but remains broadly the same across all orthologue selection methods.

Interacting protein pairs in the positive datasets were obtained either through multiple independent experimental results (i.e. Pazos+ and Tan+), or directly from crystal structures (Hakes+). All protein pairs in the Hakes+ dataset obtained from the Protein Quaternary Structure database (PQS; Henrick and Thornton, 1998) display five or more inter-molecular atomic contacts within 7.5Å of each other. These positive datasets are therefore all considered to be high quality. Although there are many high quality PPI datasets, the availability of high quality non-interacting datasets was still inadequate. Often, negative datasets were derived from randomly selecting unpaired proteins from a positive dataset. However, this method could face the risk of pairing non-identified real interacting protein pairs. Consequently, in addition to the negative dataset (Pazos-) generated using this method, two negative datasets consisting of proteins selected from non-adjacent subcelluar compartments were obtained (Tan-) and generated (GFP-) following reported good practice from previous studies (Tan *et al.*, 2004; Ben-Hur and Noble, 2006) .

The *mirrortree* approach was applied to all datasets in Table 2.1, using all appropriate orthologue selection methods and intergenic distance metrics. The Pearson correlation coefficient, *r*, for each protein pair was then computed to determine the degree of coevolution. The *r*-score essentially measures the similarity between the two distance matrices that were derived from the proteins under consideration, with *r* ranging from -1 to 1. Higher *r*-scores indicate a greater probability that the two proteins interact. These were also converted to *Z*-scores, a normalization step that has been proposed to assess the statistical significance of individual *r*-scores (Goh *et al.*, 2000). One of the pairwise distance matrices is shuffled 1000 times to mimic the background distribution so that a background mean and the associated standard deviation can be obtained to compute the *Z*-score. The prediction results are summarized in Table 2.2A for the prokaryotic, and 2.2B for the eukaryotic datasets.

A)

Dataset		Inter-protein interaction prediction in *r* (*Z*) Inparanoid	BLAST-SwissProt	BLAST-Proteomes	Average *r* (*Z*)
Pazos+	ClustalW	0.85 (8.91)	0.77 (14.67)	0.81 (9.41)	0.76 (9.56)
	PROTDIST	0.78 (6.89)	0.67 (10.99)	0.73 (6.94)	
	SIMPLE	0.73 (6.84)	0.64 (11.21)	0.67 (7.01)	
	MATRIX	0.83 (8.22)	0.73 (13.31)	0.78 (8.51)	
	TREE	0.78 (6.85)	0.77 (14.17)	0.82 (9.49)	
Pazos-	ClustalW	0.75 (6.58)	0.68 (10.98)	0.67 (7.27)	0.63 (6.85)
	PROTDIST	0.63 (4.77)	0.53 (6.96)	0.52 (4.58)	
	SIMPLE	0.63 (4.88)	0.56 (7.60)	0.50 (4.63)	
	MATRIX	0.73 (5.99)	0.65 (9.57)	0.63 (6.30)	
	TREE	0.64 (4.71)	0.68 (10.7)	0.68 (7.29)	
Average *r* (*Z*)		0.73 (6.46)	0.67 (11.02)	0.68 (7.14)	

B)

Dataset		Inter-protein interaction prediction in *r* (*Z*) Inparanoid	ENSEMBL	BLAST-SwissProt	BLAST-Proteomes	Average *r* (*Z*)
Tan+	ClustalW	0.83 (8.87)	0.63 (4.13)	0.64 (8.81)	0.71 (7.71)	0.63 (5.78)
	PROTDIST	0.76 (6.24)	0.66 (4.03)	0.56 (6.23)	0.60 (4.87)	
	SIMPLE	0.53 (4.56)	0.40 (2.51)	0.60 (7.11)	0.45 (4.02)	
	MATRIX	0.75 (7.39)	0.54 (3.45)	0.66 (8.76)	0.66 (6.62)	
	TREE	0.77 (6.06)	0.60 (3.55)	0.56 (5.91)	0.61 (4.72)	
Hakes+	ClustalW	0.74 (7.44)	0.56 (3.62)	0.46 (5.49)	0.59 (6.29)	0.51 (4.53)
	PROTDIST	0.67 (5.53)	0.61 (3.70)	0.36 (3.83)	0.48 (3.88)	
	SIMPLE	0.51 (4.56)	0.34 (2.14)	0.39 (4.14)	0.38 (3.48)	
	MATRIX	0.69 (6.70)	0.47 (3.06)	0.45 (5.22)	0.55 (5.61)	
	TREE	0.67 (5.38)	0.52 (3.13)	0.36 (3.73)	0.48 (3.72)	
Tan-	ClustalW	0.80 (7.77)	0.53 (3.39)	0.48 (6.79)	0.68 (8.51)	0.54 (5.19)
	PROTDIST	0.74 (5.85)	0.65 (3.92)	0.33 (4.10)	0.59 (6.30)	
	SIMPLE	0.55 (4.36)	0.27 (1.64)	0.40 (5.12)	0.39 (4.33)	
	MATRIX	0.73 (6.56)	0.43 (2.66)	0.46 (6.32)	0.62 (7.43)	
	TREE	0.75 (5.78)	0.52 (3.03)	0.34 (3.90)	0.59 (5.91)	
GFP-	ClustalW	0.72 (7.59)	0.56 (4.22)	0.27 (2.53)	0.51 (6.21)	0.44 (3.96)
	PROTDIST	0.64 (5.77)	0.62 (4.14)	0.17 (1.29)	0.38 (3.46)	
	SIMPLE	0.46 (4.21)	0.36 (2.64)	0.17 (1.52)	0.28 (2.91)	
	MATRIX	0.69 (7.08)	0.51 (3.83)	0.25 (2.31)	0.48 (5.63)	
	TREE	0.64 (5.53)	0.56 (3.69)	0.16 (1.21)	0.39 (3.37)	
Average *r* (*Z*)		0.68 (6.16)	0.52 (3.32)	0.40 (4.72)	0.52 (5.25)	

Table 2.2. Interaction prediction results for all orthologue selection and distance methods.

As shown in Table 2.2, most of the *Z*-scores obtained for both the positive and negative sets are higher than 1.64, which corresponds to a *p*-value of 0.05. This suggests that all protein pairs have above-background correlations, and that the *r*-scores are significant. Generally, the mean *r*-scores from the randomizations are close to 0, which is much lower than the lowest mean *r*-score (GFP-) of the negative set. *P*-values can also be calculated by directly comparing the randomization *r*-scores to the predicted *r*-scores. For example, the *r*-score for the protein pair from Pazos+, ATP6_ECOLI and ATPL_ECOLI , was calculated as 0.85 with a *Z*-score of 9.09, which corresponds to a *p*-value smaller than 0.001. Furthermore, the randomized *p*-value was determined to be 0 where all randomized *r*-scores were lower than the highest predicted *r*-score. A similar trend was also observed for the negative datasets. For instance, the negative protein pair from Pazos-, ACRB_ECOLI and ATPE_ECOLI, had an *r*-score of 0.53 and a Z-score of 2.85, which corresponds to a significant *p*-value of 0.002. The randomized *p*-value was computed as 3.25×10^{-12} where 7 randomized *r*-scores were found to be higher than the highest predicted *r*-score. As the results of *Z*-score-derived *p*-values have been shown to be in agreement with the randomized *p*-value results, the Z-score statistic appears to be a reasonable approach for determining the significance of *r*-scores, at least assuming the method used to shuffle sequence order in one protein MSA is a good Null model.

To evaluate the effectiveness of the prediction methods, ROC-style calculations were performed, computing AUC (Area Under the Curve) values for both *r*- and *Z*-score statistics, for matched pairs of positive and negative datasets. This is a more effective way to evaluate prediction performance than *r*-scores alone, with random prediction equivalent to an AUC of 0.5. The Hakes+/Tan- set performed the worst, with a mean AUC value of 0.46, while the best performance was found in the Tan+/GFP- set, which had a mean AUC value of 0.71 (Table 2.3). The poor performance of the Hakes+/Tan- set could be explained by the poor quality of the negative dataset, Tan-. When the Hakes+ positive dataset was paired with another negative dataset, GFP-, the mean AUC value increased to 0.58. The same trend was also reflected for the Tan+/Tan- and Tan+/GFP- sets, as the mean AUC value for Tan+/GFP- is much higher than the mean AUC value for the Tan+/Tan- set. Interestingly, both the *r*- and *Z*-scores for the Tan- negative set are higher than the same statistics found for the Hakes+ positive set (Table

2.2). For a negative dataset to result in higher correlation between the proteins under test than a positive dataset, it suggests that either the proteins in the positive set were mistakenly annotated as interacting or the non-interacting proteins in the negative set were actually yet to be discovered interacting proteins. The former is rather unlikely for the positive datasets utilized in this chapter, as experimental or structural evidence was taken into account when generating these datasets.

It has been suggested that restricting the selection of negative protein-protein interaction pairs to only a few specific cellular compartments could lead to substantial bias in results (Ben-Hur and Noble, 2006). Membrane proteins are under-represented in all the other datasets; however, all Tan- dataset proteins are membrane proteins which might be under different evolutionary pressure. Therefore, it is not surprising to obtain inferior prediction results when using the Tan- negative set. As for the positive datasets, Tan+ outperformed Hakes+ with higher mean AUC scores when paired with both the eukaryotic negative datasets, Tan- and GFP-. Higher *r*- (0.63) and *Z*-scores (5.78) were also obtained for the Tan+ set.

It is somewhat surprising to see that the structurally defined dataset (Hakes+) produced poor prediction performance, as such data is often considered to be a "Gold standard" for directly interacting proteins. Three possible explanations should be considered. First, perhaps 3D data is not as good as data obtained from other experimental sources, and does not capture the full repertoire of coevolutionary signals present in the sequences. Second, there might be some pre-existing bias in the Hakes+ data set which leads to lower quality predictions. Finally, it is possible that the *mirrortree* approach attempts to determine the coevolutionary relationship between two proteins (coevolution) rather than direct, physical contacts (co-adaptation) between them. It has been suggested that coevolving proteins are often not directly contacting in 3-dimensional space, but rather are located in the same protein complexes and/or have similar functions (Pazos *et al.*, 1997; Yeang and Haussler, 2007; Burger and van Nimwegen, 2008). As a result, it is more likely that the lower prediction performance of the Hakes+ set is due to the third possibility. Indeed, Yeang and Haussler observed similar findings using residue-based approaches, noting that coevolution is not necessarily well-correlated with physical interaction. The discrepancy among different

datasets strongly implies that multiple high quality datasets are essential for PPI analysis in order to obtain unbiased inferences.

		Methodology performance assessment: AUC based on *r* (AUC based on *z*)					
Datasets (positive/negative)		Inparanoid	ENSEMBL	BLAST-SwissProt	BLAST-Proteomes	Average AUC, excluding ENSEMBL	Average AUC, including ENSEMBL
Pazos+/Pazos-	ClustalW	0.68 (0.67)	-	0.61 (0.62)	0.76 (0.65)	0.69 (0.66)	-
	PROTDIST	0.71 (0.70)		0.64 (0.66)	0.77 (0.72)		
	SIMPLE	0.66 (0.69)		0.59 (0.62)	0.74 (0.72)		
	MATRIX	0.69 (0.68)		0.60 (0.63)	0.75 (0.67)		
	TREE	0.71 (0.69)		0.61 (0.62)	0.76 (0.65)		
Tan+/Tan-	ClustalW	0.62 (0.61)	0.64 (0.67)	0.68 (0.59)	0.57 (0.41)	0.61 (0.52)	0.61 (0.55)
	PROTDIST	0.58 (0.57)	0.52 (0.54)	0.70 (0.56)	0.55 (0.33)		
	SIMPLE	0.50 (0.54)	0.68 (0.68)	0.71 (0.62)	0.58 (0.49)		
	MATRIX	0.59 (0.58)	0.64 (0.67)	0.71 (0.61)	0.60 (0.41)		
	TREE	0.58 (0.56)	0.59 (0.60)	0.69 (0.57)	0.55 (0.35)		
Tan+/GFP-	ClustalW	0.73 (0.64)	0.61 (0.49)	0.84 (0.85)	0.78 (0.67)	0.76 (0.70)	0.71 (0.64)
	PROTDIST	0.73 (0.59)	0.57 (0.50)	0.81 (0.82)	0.76 (0.69)		
	SIMPLE	0.59 (0.54)	0.57 (0.48)	0.88 (0.87)	0.70 (0.64)		
	MATRIX	0.63 (0.52)	0.54 (0.44)	0.87 (0.87)	0.75 (0.61)		
	TREE	0.74 (0.61)	0.56 (0.48)	0.82 (0.82)	0.75 (0.69)		
Hakes+/Tan-	ClustalW	0.41 (0.46)	0.53 (0.56)	0.48 (0.45)	0.34 (0.23)	0.44 (0.40)	0.46 (0.44)
	PROTDIST	0.42 (0.46)	0.46 (0.46)	0.50 (0.43)	0.36 (0.21)		
	SIMPLE	0.47 (0.52)	0.59 (0.61)	0.48 (0.45)	0.47 (0.42)		
	MATRIX	0.48 (0.52)	0.56 (0.59)	0.48 (0.46)	0.38 (0.29)		
	TREE	0.42 (0.45)	0.50 (0.52)	0.50 (0.44)	0.36 (0.23)		
Hakes+/GFP-	ClustalW	0.54 (0.49)	0.49 (0.36)	0.69 (0.69)	0.61 (0.51)	0.62 (0.58)	0.58 (0.53)
	PROTDIST	0.57 (0.49)	0.51 (0.42)	0.68 (0.70)	0.62 (0.57)		
	SIMPLE	0.56 (0.53)	0.48 (0.41)	0.71 (0.71)	0.62 (0.57)		
	MATRIX	0.53 (0.46)	0.44 (0.33)	0.70 (0.70)	0.60 (0.48)		
	TREE	0.57 (0.50)	0.46 (0.38)	0.68 (0.70)	0.61 (0.56)		
Average AUC, excluding Pazos+/Pazos-		0.56 (0.53)	0.55 (0.51)	0.68 (0.65)	0.58 (0.47)		
Average AUC, including Pazos+/Pazos-		0.59 (0.56)	-	0.67 (0.64)	0.61 (0.51)		

Table 2.3. Interaction performance assessment for all orthologue selection and distance methods without the speciation signal correction.

2.4.2. Orthologue Selection Methods

In general, from different orthologue selection methods, there is a high degree of sensitivity in the *r*- and *Z*-score statistics. The data shows that different orthologue sets produced a wide variation in the reported *r*- and *Z*-scores. This is not surprising, as detection of orthology is not a trivial task, and different methods will find different proteins and also different numbers of species. The Inparanoid method, implemented the RBH approach, is widely held to be a more accurate approach to determining orthology than a simple BLAST search (O'Brien *et al.*, 2005). Indeed, it generated the highest average *r*- and *Z*-score statistics, while the BLAST-SwissProt method generated the lowest (Table 2.2). However, *r*- and *Z*-score statistics alone are not necessarily a good measure for determining the performance of a specific method for a single dataset; instead, positive and negative datasets should be compared for prediction purposes.

Shown in Tables 2.3 and 2.4, the highest mean AUC value was detected for the BLAST-SwissProt method, while ENSEMBL was shown to be the worst method. The performance could be ranked as follows: BLAST-SwissProt > BLAST-Proteomes > Inparanoid > ENSEMBL. Interestingly, the all-vs.-all mean sequence diversity for datasets generated using these methods follows exactly the same trend. The all-vs.-all sequence diversity was calculated by subtracting the mean percentage similarity from 1. The mean percentage similarity was derived by calculating the fraction of identical positions over the entire MSA length for each sequence pair in the MSA. Furthermore, an alternative approach was also taken to calculate sequence diversity by only comparing the query protein (the first protein in the MSA) and all non-query proteins instead of the all-vs.-all approach described above. Although the AUC rank for Inparanoid, BLAST-Proteomes and BLAST-SwissProt remains the same, the mean sequence diversity for ENSEMBL becomes the second highest from the lowest (Table 2.4). Essentially, this suggests that the sequences in the ENSEMBL dataset are quite divergent to the query protein, while many of the non-query sequences are closely related. Indeed, this is reflected in the poor performance (the lowest mean AUC score) of the data generated using the ENSEMBL method, as most ENSEMBL species are vertebrate while the query species is yeast. In contrast, the other orthologue methods

would obtain more fungi so that the species are more closely related, and show reduced sequence diversity from the query sequence.

As for the Inparanoid and BLAST-Proteomes methods, although the same 35 fully sequenced eukaryotic proteomes were utilized for their orthologue selections, BLAST-Proteomes had a slightly higher mean sequence diversity, which resulted in a higher mean AUC score. Since the BLAST-Proteomes method was a simplified version of the Inparanoid method, it was expected that the orthologues obtained using the BLAST-Proteomes method would not be as thoroughly evaluated. Consequently, some related but "false positive" sequences could be included. This is shown in the datasets, as the BLAST-Proteome generated datasets possess a slightly higher mean sequence diversity than the Inparanoid generated datasets. It is reasonable to assume that a set of true orthologous sequences should result in more accurate predictions; however it is evident that the sequence diversity of orthologous sequences with a major role could sometimes be a more imperative factor in achieving accurate predictions. The best orthology detection method was determined to be BLAST-SwissProt, which uses a query database that is comprised of sequences from over 10,000 species. Such a database selects sequences from a wide collection of species (either divergent or closely related) as opposed to the small number of species used for the other methods examined. Certainly, there is a positive correlation between predictive power and sequence diversity.

In addition to sequence diversity, the number of sequences per MSA for each different orthology selection method was also considered when to determining factors that might affect prediction results. Although the SwissProt database contains sequences derived from a large number of different organisms, many sequences are still missing due to the incorporation of incomplete proteomes. This means that for some species, more sequences are available than the others. Such variation in the completeness of different proteomes is revealed in the mean number of species. Essentially, when including Pazos+/Pazos- the set has the highest number of species for the BLAST-SwissProt method, but when excluding Pazos+/Pazos- it contains the lowest number of sequences per MSA (Table 2.4). Furthermore, the methodology difference between the RBH and simple BLAST approaches is reflected again in the mean number of

sequences, as more sequences were filtered out by the more sophisticated Inparanoid method than the BLAST-Proteomes method.

	Orthologue method	Inparanoid	ENSEMBL	BLAST-SwissProt	BLAST-Proteomes
Excluding Pazos+/Pazos-	Number of species per MSA	22	29	23	28
	Sequence diversity (all-vs.-all)	0.59	0.43	0.66	0.62
	Sequence diversity (query vs. non-query)	0.58	0.63	0.70	0.60
	AUC_r (AUC_z)	0.56 (0.53)	0.55 (0.51)	0.68 (0.65)	0.58 (0.47)
Including Pazos+/Pazos-	Number of species per MSA	22	-	32	27
	Sequence diversity (all-vs.-all)	0.58	-	0.62	0.61
	Sequence diversity (query vs. non-query)	0.54	-	0.63	0.57
	AUC_r (AUC_z)	0.59 (0.56)	-	0.67 (0.64)	0.61 (0.51)

Table 2.4. Average number of species per MSA, sequence diversity and AUC scores for different orthology detection methods.

Although the differences in the mean number of sequences per MSA for each orthologue method were small, they are broadly equivalent, averaging between 22 and 29 for the yeast datasets. The only exception was when incorporating the *E. coli* dataset using BLAST-SwissProt, an average of 32 sequences per MSA was produced. Overall, no evidence for performance improvement was observed with increasing numbers of species.

2.4.3. Distance Methods

Producing correct intergenic distances to measure the evolutionary relationships among all species in a MSA is critical for *mirrortree* based predictions, as this method compares the evolutionary relationship of two protein families. If the distance matrix of

one protein family is incorrectly generated, the degree of similarity between the two MSAs will not be measured correctly. In an attempt to determine the best method for measuring the intergenic distances among all species in a MSA, five distance methods, ranging from the more sophisticated PROTDIST method to the SIMPLE method, which uses a rather crude scoring system to measure the similarity among all sequences, were evaluated.

A similarly wide variation in *r*- and *Z*-scores was observed for different distance methods. Shown in both Table 2.2 and Figure 2.7, the ClustalW method generally produced the highest *r* and *Z*-scores, while the lowest mean *r* and *Z*-scores were observed for the SIMPLE method. Whether it was based on different datasets or orthologue methods, this trend was rather consistent. Since the SIMPLE distance method uses the crudest scoring system, which does not take different amino acid substitution rates into account when constructing distance matrices, the inability to detect full coevolutionary signal is to be expected. Consequently, this method produced the lowest average correlation scores. Several authors have favoured the use of PROTDIST and indeed this method does produce the highest correlation scores for orthologues selected via ENSEMBL (Table 2.2).

Nearly opposite trends of distance method performance were detected for the prokaryotic and eukaryotic datasets. The prokaryotic performance for different distance methods could be ranked as follows: PROTDIST > TREE > CLUSTALW > MATRIX > SIMPLE, while the performance rank for the eukaryotic set could be positioned as: SIMPLE > CLUSTALW > MATRIX > TREE > PROTDIST. PROTDIST performed the best, with the highest mean AUC score (0.707) when used to generate distance matrices for the prokaryotic set; however, it produced the lowest mean AUC score (0.584) for the eukaryotic set. In contrast, the SIMPLE distance method generated the highest mean AUC score (0.599) for the eukaryotic set and the lowest mean AUC score (0.663) for the prokaryotic set. The TREE method was found to be the second best method when evaluated using the prokaryotic data, but was determined to be the second worst method for the eukaryotic set.

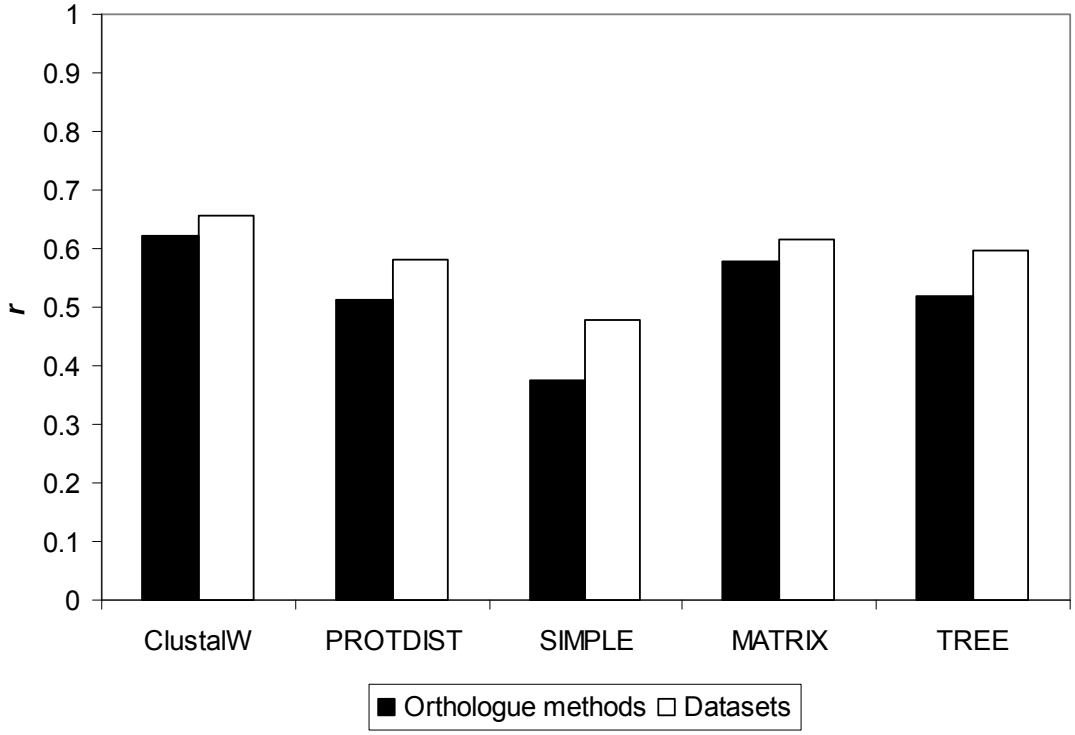


Figure 2.7. Average *r*-scores for different methods to estimate genetic distance between aligned protein sequences. The average *r*-scores for each distance method are presented when averaging over either all orthologue methods or over all datasets in this study (based on data in Table 2.3). For instance, the orthologue method based average *r*-score for the ClustalW distance method (first black bar) was calculated by averaging all of the ClustalW mean *r*-scores that were obtained for each of the four orthologue methods: Inparanoid, ENSEMBL, BLAST-SwissProt and BLAST-Proteomes, over all datasets. As for the dataset based ClustalW average *r*-score (first white bar), it was calculated by averaging all the ClustalW mean *r*-scores that were obtained for each of the 6 datasets: Pazos+, Pazos-, Tan+, Hakes+, Tan- and GFP- for all orthologue methods.

It has been suggested that different amino acid substitution matrices should be utilized for datasets with different sequence diversity in order to properly estimate evolution distances among all species in a MSA. Hence, in addition to the BLOSUM62 matrix used for all the predictions in Table 2.3, BLOSUM45 was also utilized for the Pazos+/Pazos- set, which was generated using the Inparanoid orthologue method. Since MATRIX performed worse when using the prokaryotic data, it was decided to use this dataset to evaluate whether an improvement would be obtained by using BLOSUM45. In general, higher numbered BLOSUM matrices are designed to be used for less divergent MSAs, while lower numbered BLOSUM matrices are designed for more divergent MSAs. Essentially, BLOSUM62 should be utilized to compare more closely related sequences, while BLOSUM45 should be used for more distantly related sequences. However, no significant improvement was found when using BLOSUM45,

as the resulting mean AUC scores for these two substitution matrices were very similar. The BLOSUM45 matrix produced a mean AUC score of 0.688 while the BLOSUM 62 generated a mean AUC score of 0.689. Overall, regardless of all the effort made to try to determine an overall best distance method, there is no clear pattern to support one method above another.

2.4.4. Sequence Diversity Experiment

According to earlier results (Table 2.3 and Table 2.4), it is quite apparent that sequence diversity is a key factor in detecting coevolutionary signals. In order to properly assess the association between higher sequence diversity and accurate predictions, a focused sequence diversity experiment was conducted on a controlled set of protein family pairs from *E. coli*. In an iterative process, either the most or least divergent species were removed from the MSAs of the protein pairs at each step. This allowed for the creation of a set of MSAs of either increasing or decreasing diversity, while controlling the number of species and ensuring sequences from the same species were being compared at each step.

The sequence diversity experiment results are shown in Figure 2.8, for both *r*- and *Z*-score statistics, and using the TREE distance method. In general, the *r* and *Z*-scores increase for cases where the sequence diversity decreases, as more closely related sequences are being compared. However, more importantly, the difference between the positive and negative sets is higher for the increasing diversity (orange lines) than for the decreasing diversity (purple lines). This difference was achieved almost immediately after removing one or two sequences and increasing the sequence diversity in the alignments. The pattern of the difference between the positive and negative sets was more consistent when it was measured using *r*-scores throughout the stepwise removal of sequences. When measured with the *Z*-scores, a larger difference for the first few sequences was shown, but it was only when 26 sequences were left that it decreased to a smaller difference. The results shown in Figure 2.8 offer further support to the observation that increasing diversity in multiple sequence alignments improves the predictive power of the approach. However, there also came a point where there were too few sequences remaining, which significantly reduced the overall predictive power.

In this case, a minimum of 20 sequences were retained to maintain decent performance, but the performance degraded when fewer than 10 sequences were present.

The sequence diversity for the existing protein pairs in the various datasets used throughout this study was also explored. These consisted of two positive datasets: one containing the prokaryotic Pazos+ set, and the other containing a combination of two eukaryotic datasets, Tan+ and Hakes+. Following the all-vs.-all approach utilized by Pazos *et al.*, negative datasets were generated by pairing all proteins in each positive set and all non-positive pairs were considered non-interacting. As shown in Figure 2.9, the Pearson correlation coefficient for each protein pair was then calculated and plotted against the average sequence diversity for the same protein pair.

Fitting a linear regression line to the positive and negative data points shows a trend with sequence diversity. Greater separation between the positive (red lines) and negative (black lines) datasets with increasing sequence diversity was observed. This trend is more clearly demonstrated in the prokaryotic datasets than the eukaryotic ones, but is present in both nonetheless. As has been noted before (Pazos *et al.*, 2005), *mirrortree* appears more successful when predicting prokaryotic species. The observation here is consistent with these results, as better separation of positive and negative datasets was seen for the Pazos+ set. Given that more prokaryotic genomes have been sequenced from a greater diversity of species, it is quite possible that the higher predictive power for prokaryotic species is associated with sequence diversity. This is also consistent with the proposed hypothesis that sequence diversity is an important factor in detecting any coevolutionary signal.

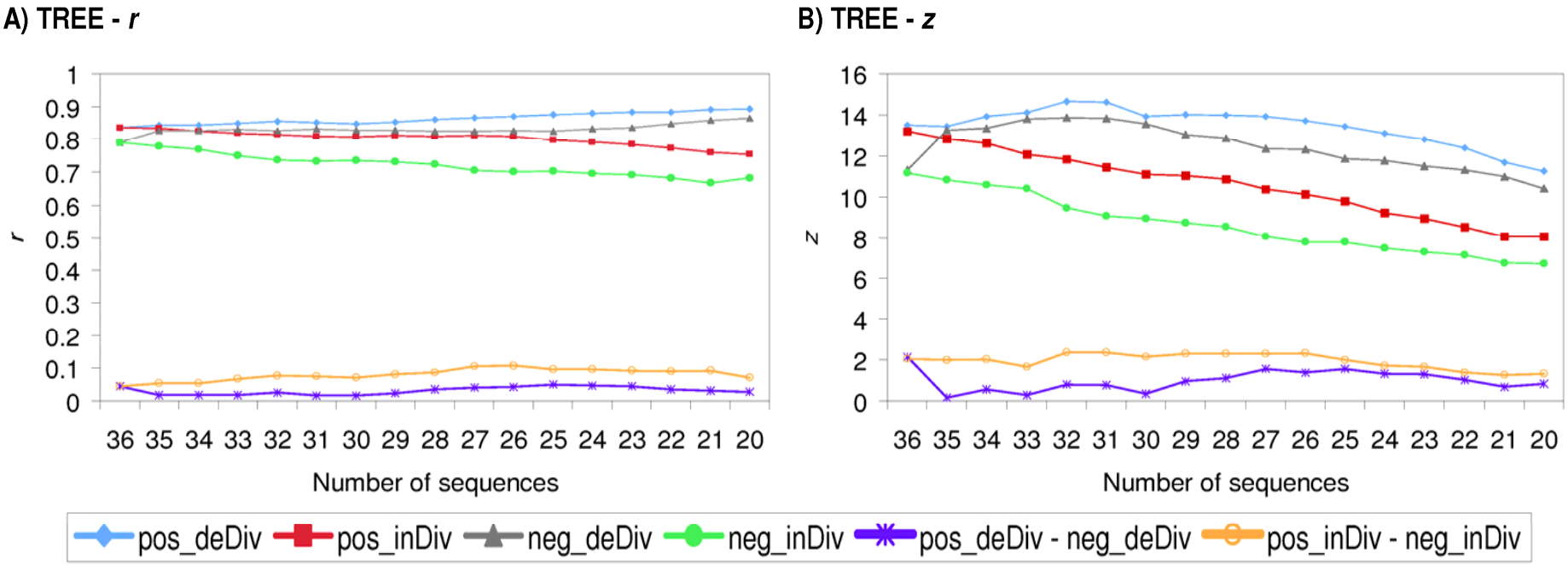


Figure 2.8. Effect of sequence diversity on prediction performance. Ten positive and ten negative protein pairs were selected from the Pazos+ and Pazos- datasets for this experiment. Based on the average sequence diversity, one sequence was removed at each step until only 20 sequences were left in each MSA. After the removal of the most or the least divergent sequence, *r* and *Z*-score statistics were computed using the standard TREE methods. The *pos_deDiv* and *neg_deDiv* curves represent the positive and the negative datasets with decreasing average sequence diversity, while the *pos_inDiv* and *neg_inDiv* curves represent the positive and the negative datasets with increasing average sequence diversity. The differences between the positive and the negative datasets with decreasing and increasing diversity are denoted as "*pos_deDiv – neg_deDiv*" and "*pos_inDiv – neg_inDiv*". Datasets with increasing sequence diversity appear to perform better (there are greater differences between the positives and the negatives).

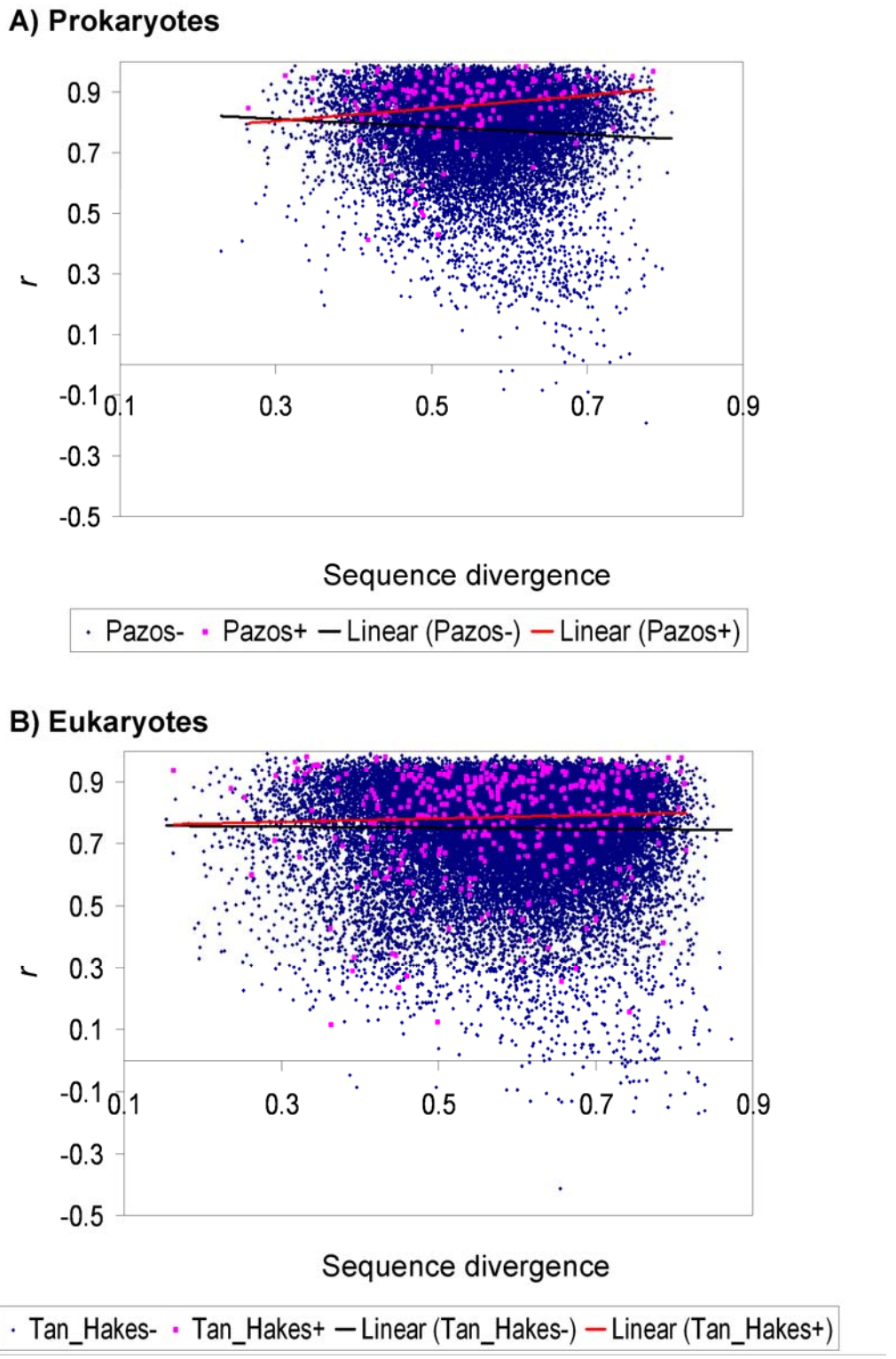


Figure 2.9. Investigation of the relationship between the Pearson correlation coefficient and sequence diversity. Shown in scatter plots for A) the prokaryotic set (Pazos+) and B) the eukaryotic set (Tan+ plus Hakes+). The linear correlation coefficients for all pairwise protein interaction tests are plotted against sequence diversity, averaged over the two protein MSAs. All negative dataset points were generated using positive datasets based on the all-vs.-all approach. Two linear regression lines are fitted to the data for the positive (red lines) and negative (black lines) examples in both plots.

It should be noted that sequence diversity from the query protein should also be considered when selecting appropriate orthologues for *mirrortree* style analysis. As shown in Table 2.4, the datasets generated using the ENSEMBL orthologue method resulted in the poorest prediction performance with the lowest mean sequence diversity when the all-vs.-all approach was applied to calculate sequence diversity. However, when the query vs. non-query relationships were considered, the ENSEMBL mean sequence diversity increased. Because as expected, the yeast query sequence is actually quite divergent from the other sequences in the same MSA. Although it was shown that higher sequence diversity generally results in more accurate predictions, too much distance from the query protein could potentially disrupt any coevolutionary signal, as the proteins being tested might be under quite different evolutionary pressures.

2.4.5. Speciation Signal Correction Methods

An earlier analysis (Section 2.4.3) evaluated five different distance methods and revealed no clear single best method for generating distance matrices for *mirrortree* style interaction predictions. Independent studies (Pazos *et al.*, 2005; Sato *et al.*, 2005) have suggested that there is a phylogenetic relationship among the species in a MSA and that such background signals could risk masking the true coevolutionary signal and eventually lead to incorrect predictions.

To address this issue, three approaches, all of which attempt to remove the underlying speciation signal, thereby maximizing any true coevolutionary signal produced by interaction or common functions, were evaluated. An overview is shown in Figure 2.1. The method, referred to here as RNA_TREE1 (Pazos *et al.*, 2005), uses small subunit ribosomal RNA sequences as sentinel sequences from which to estimate the species tree. It then scales RNA distances to protein distances, and subtracts the normalized RNA-based species signal from the protein trees. Pearson correlation coefficients, as well as *Z*-scores, are then calculated from the corrected distance matrices (see Methods). Conceptually similar methods were also developed by Sato and colleagues (Sato *et al.*, 2005).

The first method by Sato *et al.* uses RNA sequences again, but transforms distance matrices into unit vectors prior to the subtraction of the orthogonal component (referred to here as RNA_TREE2). A second method uses the average values across all proteins under consideration to generate a unit vector of average distances which is also subtracted (UAVE_TREE). Both methods were reported to significantly improve predictions for a test set of 26 *E. coli* proteins (Sato *et al.*, 2005), but were not tested on eukaryotic proteins. Next, prediction performance was assessed using separate positive and negative datasets (Pazos+/Pazos- and Hakes+/GFP-). Additionally, an all-vs.-all comparison of proteins in the positive test-set (Pazos+, Tan+ and Hakes+), which assumes all but positive-positive pairings are non-interacting, was carried out to minimize any possible bias created by using a small set of negative examples. All MSAs were generated using the Inparanoid method; the prediction performance for the separate positive and negative sets is summarized in Table 2.5, while the prediction performance for the all-vs.-all sets is shown in Table 2.6. In order to evaluate the performance of the three analyzed speciation signal correction methods, the TREE method (without any background signal correction) was also included for comparison purposes.

As shown in Tables 2.5 and 2.6, there is a general improvement in prediction performance in all cases where the two RNA-based correction methods, RNA_TREE1 and RNA_TREE2, have been applied. RNA_TREE1 appears to be a more effective method in reducing the speciation signal, as it outperformed RNA_TREE2 in most cases. However, consistently poor prediction performance was observed for all UAVE_TREE results, despite the fact that it was reported to significantly reduce the number of false positives and improve PPI predictions (Sato *et al.*, 2005).

	Prokaryotes		
	Methodology performance assessment: AUC		
Distance method	No ERS	ERS - 1.9	ERS - no gap
TREE (No correction)	0.689	0.693	0.704
RNA_TREE1	0.708	0.706	0.712
RNA_TREE2	0.698	0.704	0.712
UAVE_TREE	0.666	0.667	0.680

	Eukaryotes		
	Methodology performance assessment: AUC		
Distance method	No ERS	ERS - 1.9	ERS - no gap
TREE (No correction)	0.576	0.587	0.606
RNA_TREE1	0.651	0.678	0.671
RNA_TREE2	0.581	0.600	0.622
UAVE_TREE	0.491	0.500	0.513

Table 2.5. Performance assessment of tree-based distance methods for the separate positive/negative dataset approach. This approach compares separate positive and negative datasets (Pazos+/Pazos- and Hakes+/GFP-) for the calculations of AUC scores as exemplars for prokaryotic and eukaryotic species.

Again, there is dataset dependent variation in performance, with the poorest results obtained from the Hakes+ set (average AUC score of 0.544), and the best from the Tan+ dataset (average AUC score of 0.706). As pointed out by other studies (Pazos *et al.*, 1997; Yeang and Haussler, 2007; Burger and van Nimwegen, 2008), direct physical contact between coevolving proteins is not often observed; instead, many of the proteins located in the same protein complexes perform the same or similar functions. Hence, it is understandable that the 3D structure-based Hakes+ dataset would perform poorly when compared to the multiple-experimental-method based Tan+ set. In addition, it appears that sequence diversity was also a factor that contributed to such results, as Tan+ has a higher sequence diversity than Hakes+. Similar trends were observed for datasets that were generated using either the separate positive/negative or all-vs.-all approaches.

	Pazos+		
	Methodology performance assessment: AUC		
Distance method	No ERS	ERS - 1.9	ERS - no gap
TREE (No correction)	0.674	0.682	0.689
RNA_TREE1	0.687	0.679	0.669
RNA_TREE2	0.682	0.689	0.696
UAVE_TREE	0.648	0.654	0.660

	Tan+		
	Methodology performance assessment: AUC		
Distance method	No ERS	ERS - 1.9	ERS - no gap
TREE (No correction)	0.723	0.708	0.678
RNA_TREE1	0.744	0.747	0.740
RNA_TREE2	0.741	0.731	0.711
UAVE_TREE	0.670	0.655	0.622

	Hakes+		
	Methodology performance assessment: AUC		
Distance method	No ERS	ERS - 1.9	ERS - no gap
TREE (No correction)	0.540	0.536	0.548
RNA_TREE1	0.576	0.566	0.562
RNA_TREE2	0.543	0.539	0.552
UAVE_TREE	0.522	0.516	0.526

Table 2.6. Performance assessment of tree-based distance methods for the all-vs.-all approach. This approach compares three positive sets (Pazos+, Tan+ and Hakes+) with corresponding negative sets that were generated by pairing all proteins against all others in each positive set, and considering all non-positive pairings as negative examples.

Using a representative set of protein families to calculate the conversion, Pazos *et al.* reported an average intergenic distances ratio of 0.42 for protein:rRNA. However, in this benchmarking study, the ratio was typically estimated to be over 2.0. At very small values of p/r, only a very small correction will be made for any common speciation signal and the calculation produces effectively the standard *mirrortree* method. At large values of this conversion factor, the speciation signal will apparently dominate and lead to negative distances once the "speciation" signal is subtracted. The calculation, and hence prediction performance, is therefore extremely sensitive to the estimation of the correction factor p/r when converting RNA to protein distances.

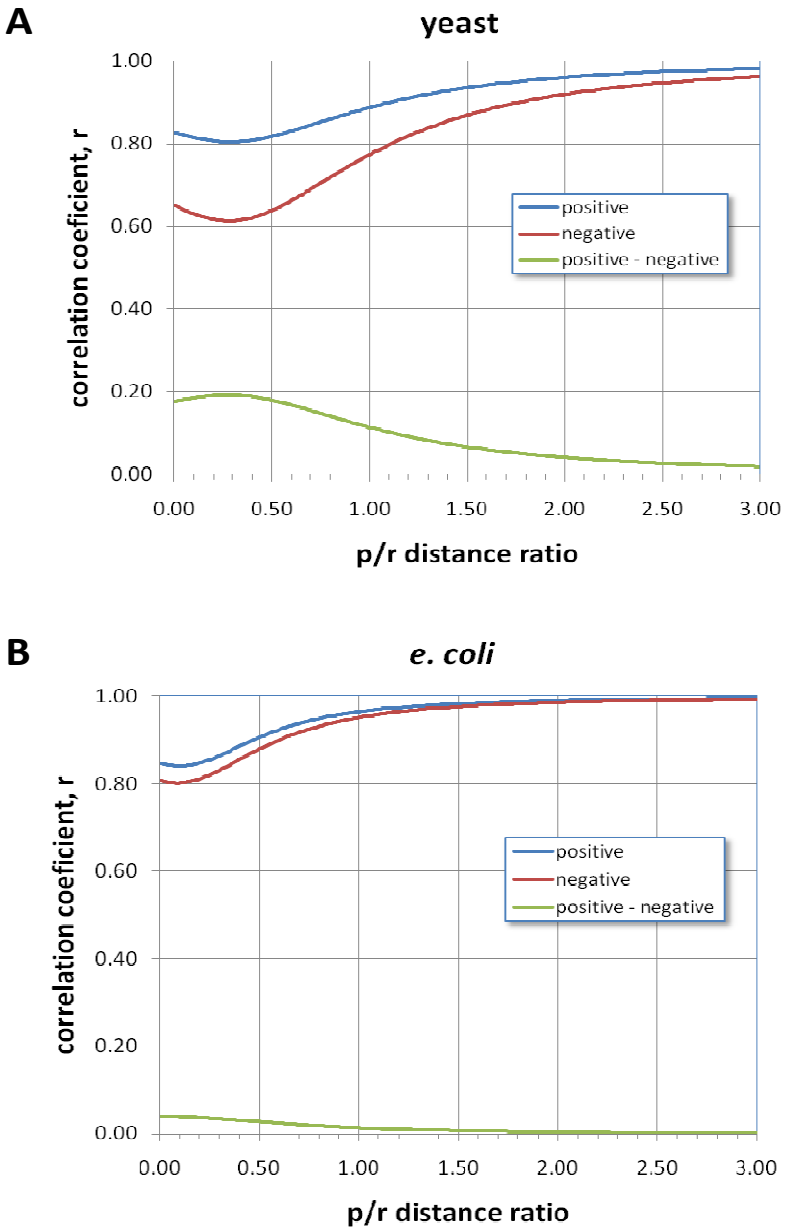


Figure 2.10. Correlation coefficient as a function of a protein/RNA conversion ratio when using the RNA_TREE1 correction method. A) Combined Tan and Hakes positives/negatives for yeast. B) The standard Pazos *E.coli* dataset for prokaryotic species.

To determine the optimal protein:RNA ratio, an experiment using paired positive and negative datasets for yeast and *E*. Coli were carried out to repeat the RNA_TREE1 correction step but varied the conversion ratio systematically rather than estimated it from a set of proteins that co-vary with ribosomal RNA matrices. As shown in Figure 2.10, this has an effect on the calculated correlation

coefficients averaged over the different datasets. The difference between the positive and negative datasets in *r* values is shown, and this peaks at around a p/r ratio of 0.25 for yeast, and slightly less for the *E. coli* dataset.

In addition to using the AUC statistic to evaluate the prediction performance, the percentage of false positives was also computed. Using the same method as described by Pazos and colleagues (Pazos *et al.*, 2005), *r*-scores for all protein pairs in an all-vs.-all calculation using positive sets were sorted in descending order for each individual protein, counting all non-interacting protein pairs ranked higher than the first true interacting protein pair as false positives. The percentage of false positives was then calculated by dividing the number of false positives by the total number of protein pairs in the dataset. For instance, a dataset with 50 protein pairs would produce a total of 100 *r*-scores for each individual protein *vs.* all others, only one of which is the true pair. After ranking the protein pairs based on their *r*-scores in descending order, if 10 protein pairs ranked higher than the true protein pair, the false positive rate would be 10%. Hence, an ideal predictor would universally rank the true interaction first above all false positives, and a random prediction would give an average of 50%. As shown in Table 2.7, the percentage of false positives varies when using different datasets, with values ranging from 18-35%. Nonetheless, the RNA_TREE1 appears to produce the lowest percentage of false positives, while the UAVE_TREE seems to produce the highest. The trend observed using percentage false positives is generally the same as the AUC score trend. Again, noticeably poorer performance results were obtained using the Hakes+ set compared to the other datasets.

Finally, also investigated was the relative entropy correction method (Kann *et al.*, 2007), where highly divergent and gapped positions in MSAs were removed via ERS, an entropy reduction step. This was done due to the fact that evolutionary pressures resulting from protein interactions are likely to act on only subsections of the sequence (Kann *et al.*, 2007). ERS uniformly improves the predictive performance for the all-vs.-all *E. coli* dataset (Table 2.6) and both the paired *E. coli* and yeast test sets (Table 2.5), but had almost no benefit in the all-vs.-all eukaryotic tests (Tan+ and Hakes+). Since eukaryotic proteins

generally contain multiple domains, many of which have been found to be independent proteins in prokaryotic proteomes (Davidson *et al.*, 1993), the process of identifying functional domains (supposedly more conserved and with a stronger coevolutionary signal than non-functional regions) in eukaryotic proteins is certainly more complex. Using a single entropy cutoff for the exclusion of columns from the alignments could lead to accidental removal of important functional positions; as such, a reduced coevolutionary signal as the optimal threshold is probably not general, and is likely to vary for alignment length, number and diversity of species.

	Pazos+		
	Average % False positives		
Distance method	No ERS	ERS - 1.9	ERS - no gap
TREE (No correction)	26.32	25.28	24.44
RNA_TREE1	20.82	21.53	21.50
RNA_TREE2	24.30	23.69	22.99
UAVE_TREE	28.61	27.53	26.96

	Tan+		
	Average % False positives		
Distance method	No ERS	ERS - 1.9	ERS - no gap
TREE (No correction)	20.45	20.44	23.58
RNA_TREE1	18.28	17.90	20.31
RNA_TREE2	19.39	19.56	21.58
UAVE_TREE	22.72	23.20	26.40

	Hakes+		
	Average % False positives		
Distance method	No ERS	ERS - 1.9	ERS - no gap
TREE (No correction)	33.81	34.70	33.75
RNA_TREE1	32.15	33.98	33.99
RNA_TREE2	33.78	34.90	33.76
UAVE_TREE	33.98	35.11	34.32

Table 2.7. Average false positive rates for tree-based distance methods for the all-vs.-all approach.

2.5. Summary

Protein-protein interactions have been extensively studied, as proteins often interact with other proteins to perform their biological functions; indeed,

PPIs are the initial step toward gaining a full understanding of cellular machinery. Many experimental and computational methods have been developed to detect or predict PPIs. However, large discrepancies in the data produced using many of these methods also exist. Prior to an in-depth analysis, a full understanding of the data and methods is necessary, as without it incorrect interpretations are possible. As such, an extensive benchmarking analysis was carried out to study a coevolution based PPI approach, *mirrortree*. Many other important factors that might influence the prediction results were also examined.

Different approaches were taken to obtain positive (multiple experimental methods and structural properties) and negative (different cellular compartments or non-positive pairs from all-vs.-all comparisons) datasets. It was revealed that different datasets could eventually lead to very different mean *r*- and *Z*-scores. In comparison to the structural-based dataset (Hakes+), better predictive performance on datasets was derived from high-throughout experiments (Tan+), such as yeast-two-hybrid and tandem affinity purifications. This is in strong agreement with the results from residue-based methods (Yeang and Haussler, 2007), which have revealed that although coevolving residue pairs are closer in space in the protein structures than would happen by chance, the converse was not true; physically interacting residue pairs were generally not observed to be coevolving. For the best global distance method, no clear consensus can be made, because the mean AUC values evidently cluster based on the datasets and orthologue selection methods rather than the distance methods. Within each dataset generated using a different orthologue selection method, the performance for all distance methods can be ranked accordingly by the mean AUC values. However, the ranks vary from dataset to dataset.

A few commonly used orthologue selection methods were utilized to generate orthologues for multiple sequence alignment construction. An apparent correlation between mean AUC scores and mean sequence diversity was discovered. Essentially, datasets constructed using more divergent species resulted in better prediction performance. In particular, the BLAST-SwissProt datasets have both the highest mean sequence diversity and the highest mean AUC score. There is no doubt that this method is likely to produce orthologues

from more diverse species, as the underlying database, SwissProt, consists of sequences from over 10,000 species. To further examine the association found between higher sequence diversity and better predictive performance, a sequence diversity experiment was carried out using both prokaryotic and eukaryotic sequences. Indeed, the increasing diversity test set performed better when there were greater differences in *r*-scores between the positive and the negative sets. Sequence diversity in MSAs is an important consideration, because if one considers an extreme example using closely related species, the pairwise intergenic distances derived from the MSA will all be small and very sensitive to any changes in the individual sequences, even when they have little bearing on the function. It may be particularly challenging to detect coevolution against such a background. As shown in the results in this study, this would not be expected to be the case in highly divergent families. Taken as a whole, these results suggest that sequence diversity should be considered when selecting orthologues.

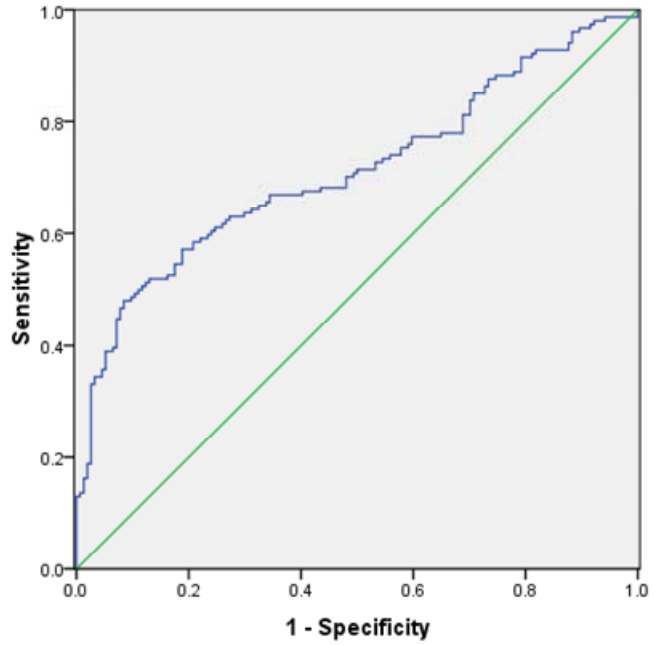
It has been suggested that by removing the underlying speciation signal, predictive power should dramatically improved. Indeed, both RNA_TREE1 and RNA_TREE2 produced higher mean AUC scores than the non-background-signal-corrected TREE method. In particular, RNA_TREE1 appears to be removing the speciation background signal more efficiently than RNA_TREE2, as it consistently produced higher mean AUC scores. However, the non-RNA based method, UAVE_TREE, produced poor performance and does not seem to offer any benefits for PPI prediction. Hakes and colleagues found that restricting alignments to surface or interface residues offers no improvement (Hakes *et al.*, 2007), although Kann *et al.* (2007) found that subtracting the binding regions from the alignments invariably reduced prediction performance (Kann *et al.*, 2009). In agreement with Hakes *et al.* but contrary to Kann *et al.*, no improvement was found after removing either highly variable or gapped columns for the all-vs.-all eukaryotic sets.

In summary, the *mirrortree* approach is highly sensitive to many dependent parameters. Nevertheless, improvement can be achieved by carefully selecting these parameters. As an example, ROC curves for the best performing prokaryotic and eukaryotic sets were plotted and shown in Figure 2.11. Previous

studies (Pazos and Valencia, 2001; Pazos *et al*., 2005) suggested a correlation coefficient of 0.80 as a reasonable cutoff to distinguish between interacting and non-interacting proteins. However, the corresponding sensitivity and specificity for such a cutoff in this study was found respectively to be 0.20 and 0.97 for the prokaryotic set. This indicates that only 20% of true interacting protein pairs would be correctly identified, and 97% of non-interacting protein pairs would be correctly identified. As for the eukaryotic set, the same cutoff of 0.80 would result in a sensitivity of 0.37 and a specificity of 0.91. Although the false negative rates are quite low for this cutoff, the probability of identifying true positives is arguably too low to be of practical use. To address this issue, a universal cutoff which usually gives a very high sensitivity at a modest specificity could be set. This would be tolerable in a genome wide screen, as some correct PPIs would be predicted with high accuracy. Indeed, Juan *et al*. shows some promise for this. However, it is not easy to derive a universal 'optimal' cutoff, and selecting a single *r*-score cutoff is not advisable as results will be highly variable; in some cases it would be over-predicting and in others under-predicting.

Due to the large number of factors that could influence prediction results, if parameters are not selected carefully, an application of the *mirrortree* approach for whole proteome studies could face a risk of erroneous results. For instance, due to different evolutionary selective constraints, comparing membrane and non-membrane proteins could result in poor predictions, as seen in the low mean AUC score for the Hakes+/Tan-. However, this could be used for more focused studies where the parameters are easier to adjust for, such as interactions between different ligands and protein families. It also shows more convincing performance when applied to protein datasets with higher sequence diversity, and for more generalized concepts of co-function/coevolution rather than direct, physical interactions, characterized as co-adaptation (Juan *et al.*, 2008b). For the prediction of physical interactions, residue-based coevolutionary methods appear better suited (Shackelford and Karplus, 2007; Yeang and Haussler, 2007; Hamacher, 2008). Lastly, the removal of the underlying speciation signal appears to be a critical step in improving prediction performance.

A) Best prokaryotic predictive performance



B) Best eukaryotic predictive performance

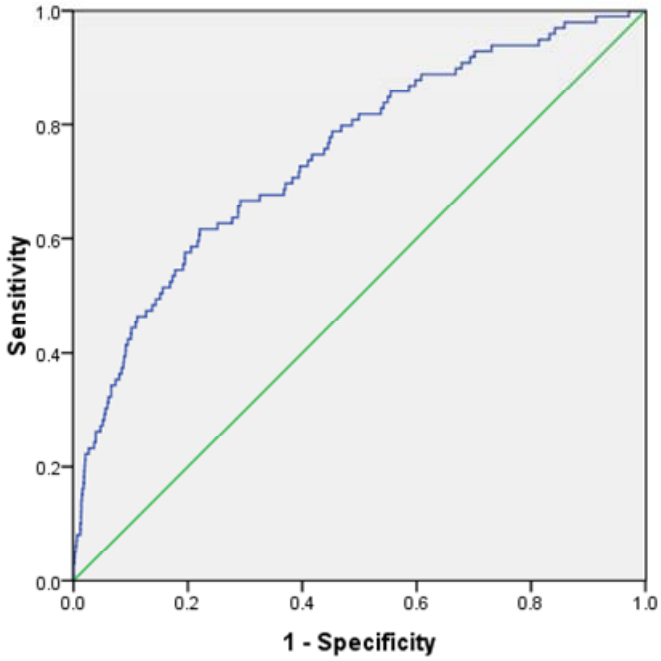


Figure 2.11. ROC curves for the A) prokaryotic and B) eukaryotic sets with the best predictive performance. The prokaryotic set with the highest predictive power was found to be the Pazos+/Pazos- set with all gaps removed and RNA_TREE1 or RNA_TREE2 (not shown here) methods applied. The best performing eukaryotic set was found to be the Tan+ all-vs.-all set when all columns with an entropy higher than 1.9 were removed. The distance method used for this set was RNA_TREE1.

3. Domain-Domain Interactions of the Fibrillin-1 Family

3.1. Aim

The aim of this study was to assess the utility of the *mirrortree* approach to the prediction of domain-domain interactions within a multi-domain protein that is of genuine experimental interest to structural biologists. Human fibrillin-1 was used as the test protein, and all predicted interdomain interactions were evaluated and compared with associated PDB structures and experimental results acquired from scientific literature.

3.2. Introduction

A protein domain is usually classified as a structurally, functionally and evolutionarily independent protein segment. During the course of evolution, domains may duplicate and shuffle to generate novel proteins with different functions (Bornberg-Bauer *et al.*, 2010). Hence, gaining insight into the evolutionary relationship between protein sequences is best achieved by comparing protein domains.

Many proteins, particularly eukaryotes, contain multiple domains. These domains often interact with each other, or with domains in other proteins, to maintain functional and structural integrity. For instance, in Figure 3.1A, the multi-domain human fibrillin-1 protein consists of 56 domains (Pereira *et al.*, 1993); 43 are calcium-binding epidermal growth factor-like (cbEGF) domains, four are epidermal growth factor (EGF) -like domains, seven are ‘eight-cysteine' or ‘TGFβ-binding protein-like’ (TB) domains and the remaining two are hybrid domains that are similar to cbEGF and TB domains. Early studies have shown that fibrillin-1 proteins form the backbone of microfibrils in a head-to-tail arrangement (Keene *et al.*, 1991; Sakai *et al.*, 1991; Reinhardt *et al.*, 1996). More

recent studies carried out by Lin *et al.* (2002), Marson *et al.* (2005) and Hubmacher *et al.* (2008) have also revealed interactions between the same terminal regions, i.e. N-terminus to N-terminus or C-terminus to C-terminus (Figure 3.1B). Fibrillin-1 has been found to be expressed in both developmental and adult stages, as it helps to provide structural support for microfibrils. It has also been identified in regions such as the heart, lungs, liver and central nervous system (Lin *et al.*, 2002). Many genetic disorders in connective tissue, such as Marfan syndrome and related conditions, are linked to fibrillin-1 (Lee *et al.*, 1991; Lemaire *et al.*, 2006; Lima *et al.*, 2010). As connective tissue is present throughout the body, deciphering the assembly mechanism of the fibrillin-1 protein family could facilitate an understanding of connective tissue pathogenesis.

The coevolution based *mirrortree* approach which was extensively studied in Chapter 2, has been proposed by several groups as a useful tool (Pazos *et al.*, 2005; Sato *et al.*, 2005; Juan *et al.*, 2008b; Kann *et al.*, 2009) for predicting protein-protein interaction using full protein sequences. Although this approach is highly sensitive to many dependent parameters, relatively accurate predictions can still be achieved by carefully selecting such parameters. As it is widely established that protein domains represent the core of evolution (Bagowski *et al.*, 2010), it would be intriguing to know whether such an approach could be utilized to predict domain-domain interactions using a current problem of local interest. To assess the feasibility of such application, various approaches were taken to obtain as many putative orthologues of the multi-domain protein, human fibrillin-1, as possible. Multiple sequence alignments (MSAs) representing each of the 56 domains and both terminal regions were extracted from the overall MSA and utilized as independent MSAs for the analysis. Predictions were then carried out using two *mirrortree* based methods: TREE and UAVE_TREE (see Chapter 2 Methods). Furthermore, the predictive power of these methods was evaluated using AUC scores, and 2-tailed t-tests were carried out to determine the significance of the results.

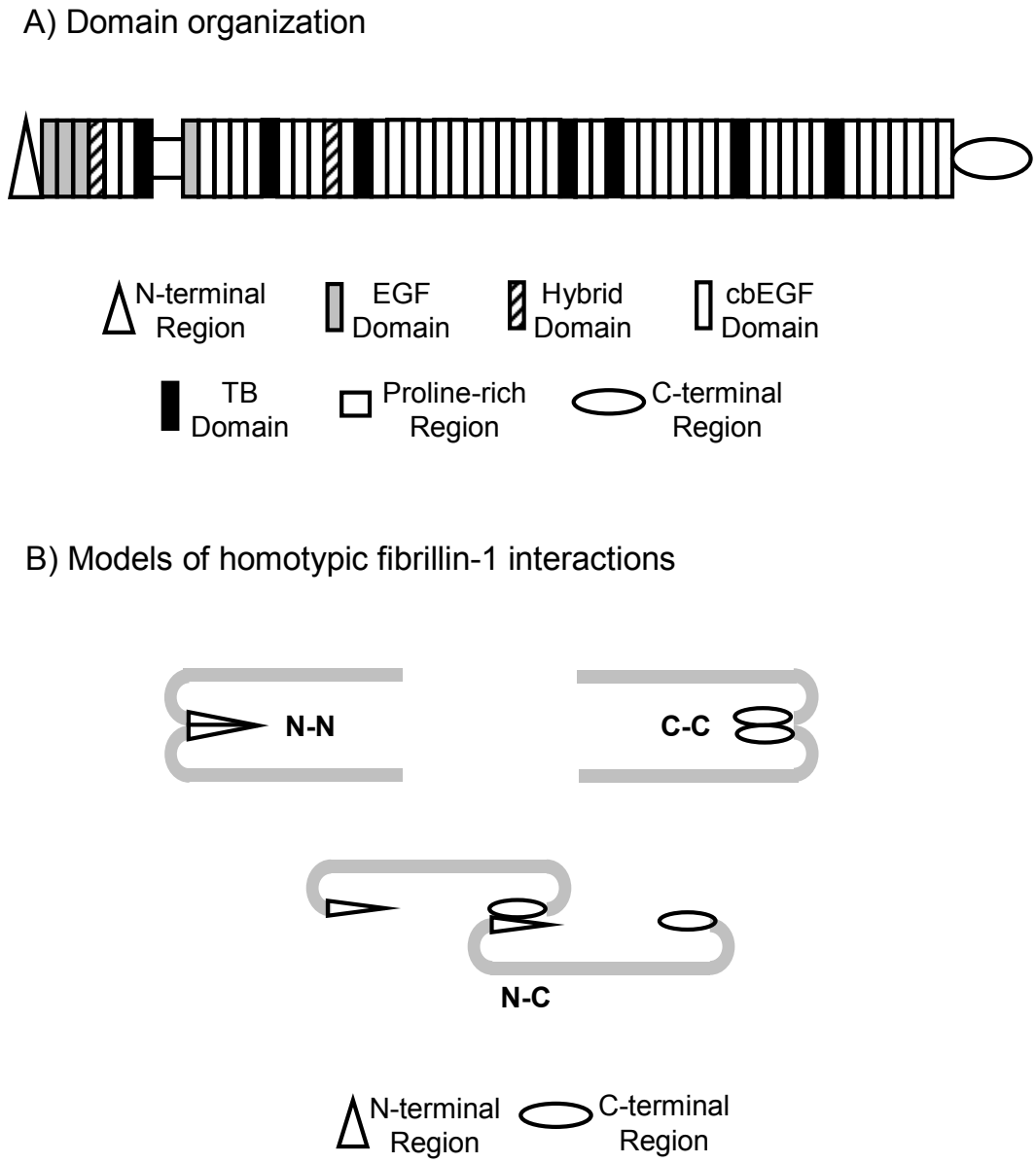


Figure 3.1. Schematic diagrams of human fibrillin-1. Fifty-six domains have been found in the human fibrillin-1 protein, and the domain organization is shown in A). Three current models of homotypic fibrillin-1 interactions are shown in B).

3.3. Methods

3.3.1. Datasets

Orthologous sequences for the human fibrillin-1 protein were obtained from Ensembl Genes 52 (Flicek *et al.*, 2008) for all available species in the database. In addition, to ensure the inclusion of all available orthologous sequences, the best reciprocal top-hit BLAST approach (Altschul *et al.*, 1990) was implemented, using default BLAST parameter values and an E-value cutoff of 10^{-5}. The query database for the BLAST searches consisted of all sequences from the SwissProt 56.7 and TrEMBL 56.7 datasets from the UniProt database (Magrane and Consortium, 2011). As shown in Table 3.1, the combination of the two orthologue methods led to the identification of 32 orthologues derived from a range of metazoan species.

The multiple sequence alignment program, ClustalW (Thompson *et al.*, 1994), was utilized with all default parameters to align all orthologous sequences acquired for the fibrillin-1 protein family. The N- and C-terminal regions and 56 (43 cbEGF, 4 EGF, 7 TB, and 2 hybrid) domain boundaries were then defined according to the domain positions described on the UniProt website (http://www.uniprot.org/uniprot/P35555).

Species	Taxon ID
Homo sapiens	9606
Erinaceus europaeus	9365
Sorex araneus	42254
Cavia porcellus	10141
Ochotona princeps	9978
Mus musculus	10090
Rattus norvegicus	10116
Dipodomys ordii	10020
Pan troglodytes	9598
Pongo pygmaeus	9600
Macaca mulatta	9544
Tarsius syrichta	9478
Tursiops truncates	9739
Equus caballus	9796
Bos taurus	9913
Sus scrofa	9823
Microcebus murinus	30608
Otolemur garnettii	30611
Myotis lucifugus	59463
Tupaia belangeri	37347
Spermophilus tridecemlineatus	43179
Lama pacos	30538
Oryctolagus cuniculus	9986
Procavia capensis	9813
Pteropus vampyrus	132908
Monodelphis domestica	13616
Gallus gallus	9031
Xenopus tropicalis	8364
Takifugu rubripes	31033
Tetraodon nigroviridis	99883
Oryzias latipes	8090
Xenopus laevis	8355
Podocoryna carnea	6096

Table 3.1. Species of fibrillin-1 orthologues.

3.3.2. Domain-Domain Interactions

Each domain in the fibrillin-1 MSA was treated as an independent MSA, and two *mirrortree* style methods, TREE and UAVE_TREE (see Chapter 2 Methods), were applied to obtain the Pearson correlation coefficient, *r*, and subsequently the Z-statistic for all putative domain-domain interactions. Due to the absence of certain regions in some orthologous sequences, the number of sequences for each domain MSA was not identical. Hence, only interactions between domains that contained a minimum of 10 sequences from common species were determined.

3.4. Results and Discussion

The benchmarking study in the previous chapter has shown promising results in predicting intermolecular protein interactions using the *mirrortree* approach. However, the study was mainly performed using full protein sequences, except in the entropy reduction section, where highly variable regions were removed in an attempt to reduce the background bias. To further extend such an approach by predicting protein interactions at the domain level, the potential coevolutionary relationships among all domains and terminal regions in fibrillin-1 were determined using the *mirrortree* approach. Additionally, experimental evidence and structural information were incorporated to evaluate the effectiveness of the application.

The coevolution analysis for interdomain interaction predictions was carried out using the TREE and UAVE_TREE methods. The other two protein-protein interaction prediction methods, RNA_TREE1 and RNA_TREE2, as described in Chapter 2, were not implemented for this study since an equivalent region of the RNA could not be obtained for each domain. Given that the aim of this study was to determine whether domain-domain interactions could be predicted by utilizing the *mirrortree* approach, the prediction results for the TREE and UAVE_TREE methods should be sufficient to demonstrate the functionality of this approach, in spite of the technical difficulties in implementing the RNA_TREE1 and RNA_TREE2 approaches.

A Pearson correlation coefficient, *r*-score, was calculated for each domain-domain or domain-terminus pair in order to predict any putative domain-domain interactions. One thousand six hundred and fifty-three pair combinations were produced utilizing the all-vs.-all approach, with the average *r*- and *Z*-scores calculated using the TREE and the UAVE_TREE methods. In order to assess the predictive power of the *mirrortree* approach for domain-domain interactions, values of the area under ROC curves (AUCs) were computed by comparing the predicted results to the known interacting regions in fibrillin-1 proteins. The true interacting domains include experimentally determined homotypic interactions (Ashworth *et al.*, 1999; Lin *et al.*, 2002; Marson *et al.*, 2005; Hubmacher *et al.*, 2008), structurally contacting positions (Berman *et al.*, 2000) and adjacent domains. In a study for mapping intramolecular and intermolecular protein family interactions, Park and colleagues (Park *et al.*, 2001) used a cutoff of 30 amino acids to denote interactions. Based on their analysis, all domains connected by less than 30 amino acids are likely to be interacting intramolecularly, and any sequence longer than 30 amino acids has a significant probability of containing a domain. As the average number of residues separating two adjacent domains was less than 30 amino acids in this study, adjacent domains were considered to be interacting domains. Of the 1653 domain pairs, 100 pairs were determined to be true interacting pairs while the rest were categorized as non-interacting domain pairs.

As shown in Table 3.2, the mean *r*-score that was computed using the TREE method for interacting domain pairs was 0.733, while a slightly smaller mean *r*-score (0.723) was obtained for the non-interacting set. The minuscule difference between the interacting and non-interacting sets was reflected in the close-to-random AUC score (0.516), as an AUC score of 0.500 signifies random predictions. After applying UAVE_TREE, a non-RNA based correction method, the mean *r*-score for the interacting set was calculated as 0.481, while the non-interacting set yielded a mean *r*-score of 0.516 using the same method. Unlike the TREE results, the negative set resulted in a higher AUC score than the positive set, and subsequently led to a 'below random' (AUC<0.500) AUC score. In addition, the *Z*-score statistic was utilized for the prediction calculations, with results similar to the *r*-score predictions obtained (Table 3.2). Although the AUC

scores for the TREE method are higher than those for the UAVE_TREE method, it is uncertain whether one method is superior to the other method, given that the overall performance is so poor, and that there is also uncertainty as to whether the negatives are truly negative and the positives are truly positive.

	r (Z)		
	Interacting	**Non-interacting**	**AUC**
TREE(no correction)	0.733 (4.30)	0.723 (4.26)	0.516 (0.503)
UAVE_TREE	0.481 (3.70)	0.516 (3.89)	0.482 (0.488)

Table 3.2. Fibrillin-1 domain-domain predictions using the TREE and UAVE_TREE methods.

The sequence diversity test carried out in Chapter 2 suggests that there is a positive association between predictive power and sequence diversity for *mirrortree* methods. Due to the poor prediction performance of the two *mirrortree* methods utilized for the fibrillin-1 domain-domain interaction predictions, sequence diversity was used to filter out domain pairs with low sequence diversity in an attempt to increase the overall mean sequence diversity. Two different methods were used to determine the sequence diversity for each MSA. The first method took an all-vs.-all approach by averaging all pairwise sequence identity values in the same MSA. The second method explored the relationship between the query sequence (the top sequence in the MSA) and the rest of the sequences in the MSA by calculating the average value of sequence identity values for all query vs. non-query pairs. Subsequently, the sequence diversity value could be calculated by subtracting the average sequence identity value from 1. Additionally, various sequence diversity cutoff values were applied, and the corresponding AUC values were determined after removing domain-domain pairs with lower mean sequence diversity than the cutoff value.

As is shown in Table 3.3, the unfiltered all-vs.-all mean sequence diversity is 0.109, which would suggest that most of the MSAs were highly conserved, with an average difference of 10% in pairwise sequence identity in each MSA. The mean sequence diversity cutoff was increased in increments of 0.01 from 0 to 0.19. A dramatic decrease in AUC scores was observed when the

mean sequence diversity reached 0.19, a result that was probably caused by the small number (3 positives and 38 negatives) of domain pairs retained after sequence diversity filtration. In general, small datasets are more prone to random predictions, as the uncertainties associated with the quality of datasets are expected to be greater in the case of small datasets. As expected, the AUC score increases when the mean sequence diversity increases, regardless of the type of *mirrortree* prediction methods used; furthermore, the highest AUC scores were obtained for datasets with a mean sequence diversity between 0.158 and 0.179. Nevertheless, datasets with these 'optimal' mean sequence diversity values were still highly conserved.

		AUC (*p*-value)			
SDIV Cutoff	**Mean SDIV**	**TREE_*r***	**TREE_*Z***	**UAVE_TREE_*r***	**UAVE_TREE_*Z***
0.00	0.109	0.516 (6.13E-01)	0.503 (8.00E-01)	0.482 (3.22E-01)	0.488 (5.42E-01)
0.01	0.109	0.516 (6.13E-01)	0.503 (8.00E-01)	0.482 (3.22E-01)	0.488 (5.42E-01)
0.02	0.109	0.516 (6.13E-01)	0.503 (8.00E-01)	0.482 (3.22E-01)	0.488 (5.42E-01)
0.03	0.109	0.516 (6.13E-01)	0.503 (8.00E-01)	0.482 (3.22E-01)	0.488 (5.42E-01)
0.04	0.110	0.516 (6.06E-01)	0.504 (7.93E-01)	0.482 (3.16E-01)	0.487 (5.32E-01)
0.05	0.110	0.519 (6.12E-01)	0.506 (7.91E-01)	0.487 (3.52E-01)	0.493 (6.01E-01)
0.06	0.114	0.541 (1.99E-01)	0.530 (2.84E-01)	0.510 (8.80E-01)	0.518 (8.51E-01)
0.07	0.118	0.553 (1.04E-01)	0.541 (1.51E-01)	0.525 (8.45E-01)	0.529 (6.74E-01)
0.08	0.124	0.550 (1.99E-01)	0.539 (2.52E-01)	0.531 (8.45E-01)	0.540 (6.02E-01)
0.09	0.129	0.559 (1.14E-01)	0.547 (1.72E-01)	0.535 (8.77E-01)	0.547 (5.77E-01)
0.10	0.136	0.574 (1.21E-01)	0.564 (1.63E-01)	0.546 (8.66E-01)	0.557 (5.87E-01)
0.11	0.143	0.609 (4.95E-03)	0.596 (1.61E-02)	0.570 (5.65E-01)	0.584 (3.86E-01)
0.12	0.151	0.601 (1.68E-02)	0.582 (6.29E-02)	0.551 (8.81E-01)	0.561 (6.84E-01)
0.13	0.158	0.617 (2.58E-02)	0.596 (9.11E-02)	0.552 (9.59E-01)	0.560 (7.51E-01)
0.14	0.164	0.615 (3.26E-02)	0.597 (8.97E-02)	0.581 (2.11E-01)	0.597 (6.27E-01)
0.15	0.172	0.608 (6.26E-02)	0.604 (1.11E-01)	0.590 (1.98E-01)	0.589 (1.25E-01)
0.16	0.179	0.580 (3.37E-01)	0.582 (3.38E-01)	0.590 (2.14E-01)	0.581 (2.30E-01)
0.17	0.186	0.588 (4.87E-01)	0.571 (5.09E-01)	0.587 (4.45E0-1)	0.584 (4.20E-01)
0.18	0.194	0.579 (8.87E-01)	0.556 (8.45E-01)	0.588 (7.93E-01)	0.551 (7.98E-01)
0.19	0.202	0.272 (3.97E-01)	0.202 (2.86E-01)	0.272 (3.97E-01)	0.281 (4.10E-01)

Table 3.3. Fibrillin-1 domain-domain predictions for the TREE and UAVE_TREE methods, filtered based on mean all-vs.-all mean sequence diversity.

In order to verify whether the AUC scores obtained were significant, a two-tailed t-test was carried out to compare the mean *r*- and *Z*-score for each positive and negative set, and a *p*-value was computed (shown in Table 3.3). When a *p*-value cutoff of 0.01 was applied, only one AUC score (0.609), which had a *p*-value of 4.95E-03 for the TREE_r method, and a mean sequence diversity of 0.143, was found to be significant. This result indicates that the mean *r*-score for the positive set is significantly higher than it is for the negative set. However, the AUC scores calculated based on Z-scores (Table 3.3) for the same set and for the rest of the dataset, did not result in significant *p*-values. This signifies that there is no difference between the mean *r*-score for the positive and negative sets. As shown in Chapter 2, the *mirrortree* method is highly sensitive to sequence diversity. Hence, it is highly likely that such poor prediction results could be caused by the low sequence diversity of the dataset. Even after filtering out MSAs with lower sequence diversity values, the highest average sequence diversity value obtained for the dataset is still lower than 0.20. As these sequences are highly conserved, little evolutionary signals among these species could be extracted for comparison; consequently, this resulted in poor predictions as the *mirrortree* approach is dependant on evolutionary relationships among the species being tested

The AUC scores for the second approach used to compute sequence diversity are shown in Table 3.4. This method compares the query and non-query sequences in the same MSA. As different sequence diversity computation approaches were implemented, the mean sequence diversity values in Tables 3.3 and 3.4 were not directly comparable, and different ranges of cutoff values were utilized to ensure that only datasets with reasonable sizes were utilized for the analysis. Generally, higher sequence diversity values were obtained when computed using the all-vs.-all approach. For the original, non-filtered MSAs, the mean sequence diversity value for the query vs. non-query was computed as 0.070 while the sequence diversity value for the all-vs.-all approach was determined to be 0.109. Such a trend was expected, as more sequence diversity values were used in the calculation when using the all-vs.-all approach. Although some might believe that the sequence diversity values that were obtained for the query vs. non-query set would better reflect the divergence of the species, and

that using these values to filter out low sequence diversity pairs should result in better predictions, similar trends were observed in Tables 3.3 and 3.4 for the all-vs.-all and query vs. non-query results. The mean AUC values in Table 3.4 appear to improve as the mean sequence diversity value increases. However, when using a *p*-value cutoff of 0.01, none of the mean AUC values obtained for this set were significant. In conclusion, regardless of the approach that was used to measure sequence diversity, the inherent low sequence diversity found in the fibrillin-1 family appears to be causing the poor predictions.

		AUC (*p*-value)			
SDIV Cutoff	**Mean SDIV**	**TREE_*r***	**TREE_*Z***	**UAVE_TREE_*r***	**UAVE_TREE_*Z***
0.00	0.070	0.516 (6.13E-01)	0.503 (8.00E-01)	0.482 (3.22E-01)	0.488 (5.42E-01)
0.01	0.070	0.516 (6.13E-01)	0.503 (8.00E-01)	0.482 (3.22E-01)	0.488 (5.42E-01)
0.02	0.070	0.516 (6.13E-01)	0.503 (8.00E-01)	0.482 (3.22E-01)	0.488 (5.42E-01)
0.03	0.071	0.523 (4.91E-01)	0.511 (6.61E-01)	0.488 (3.94E-01)	0.495 (6.52E-01)
0.04	0.074	0.526 (4.55E-01)	0.514 (6.00E-01)	0.492 (4.00E-01)	0.497 (6.57E-01)
0.05	0.079	0.553 (1.34E-01)	0.541 (1.93E-01)	0.527 (8.54E-01)	0.534 (6.50E-01)
0.06	0.084	0.577 (3.90E-02)	0.562 (6.44E-02)	0.533 (8.90E-01)	0.544 (6.03E-01)
0.07	0.091	0.578 (5.84E-02)	0.567 (9.69E-02)	0.536 (9.86E-01)	0.548 (7.00E-01)
0.08	0.098	0.607 (1.43E-02)	0.588 (5.85E-02)	0.551 (9.10E-01)	0.555 (7.32E-01)
0.09	0.104	0.580 (7.03E-02)	0.562 (1.71E-01)	0.554 (7.74E-01)	0.559 (6.25E-01)
0.10	0.112	0.620 (7.59E-02)	0.609 (1.25E-01)	0.611 (1.18E-01)	0.623 (5.42E-01)
0.11	0.120	0.584 (5.59E-01)	0.571 (6.04E-01)	0.585 (4.75E-01)	0.574 (4.81E-01)
0.12	0.127	0.435 (6.76E-01)	0.324 (4.89E-01)	0.399 (6.06E-01)	0.455 (8.35E-01)

Table 3.4. Fibrillin-1 domain-domain predictions for the TREE and UAVE_TREE methods, filtered based on mean query vs. non-query sequence diversity.

3.5. Summary

Due to the fact that the majority of the mean AUC scores for fibrillin-1 domain-domain interaction predictions do not have a significant *p*-value, no definitive conclusion could be made for the effectiveness of the *mirrortree* approach for domain-domain interaction predictions. However, regardless of the approach that is used to determine sequence diversity, it is evident that the MSAs used for this study have very low sequence diversity, which is what produced non-significant prediction results. Although different approaches were taken to ensure the inclusion of fibrillin-1 protein sequences from divergent species, it was not possible to increase the mean sequence diversity to a satisfactory level for meaningful predictions. As shown in Table 3.1, the fibrillin-1 sequences included in this study were derived from divergent species such as human (*Homo sapiens*), amphibians (*Xenopus tropicalis* and *Xenopus laevis*), jellyfish (*Podocoryna carnea*) and other types of animals. However, the mean sequence diversity values were still determined as 0.109 and 0.070 when using the all-vs.-all and query vs. non-query approaches respectively. Evidently, the fibrillin-1 family is highly conserved throughout all species and is not suitable for evolution based analysis. In summary, sequence diversity appears to be a critical factor when applying the *mirrortree* approach for domain-domain interactions.

4. Intramolecular Protein Interaction Predictions Using Mutual Information and Partial Correlation

4.1. Aim

The purpose of this chapter was to derive a predictive methodology for determining intramolecular protein interactions using mutual information and partial correlation. The effectiveness of incorporating information from a third position to facilitate the prediction for two contacting positions was evaluated. To address issues that were caused by various types of background signals, the study examined important factors, such as minimum entropy required per column in a multiple sequence alignment (MSA), ways of handling gaps, reduced amino acid alphabet groupings and the APC method. The conclusion of this study provides a useful approach for determining putative contacting residues.

4.2. Introduction

As stated in chapter 2, interacting proteins are likely to co-evolve in order to compensate for the changes brought on by evolutionary pressures. The same coevolutionary pressure is not likely to act on the whole protein in equal measure, but specifically, at the residue level involved in the interaction. To maintain structural stability, residues in close proximity in three-dimensional structures must be able to co-evolve; otherwise, the whole structure would become unstable and, in more critical regions (i.e. binding regions or active sites), even lead to a loss of function. In essence, in order to retain the interaction and to make up for the changes when one or more residues is mutated, it is presumed that other mutation(s) must occur in the corresponding interacting partner(s).

Mutual information (MI) has become a popular method for detecting coevolution among protein residues, and many approaches (Atchley *et al.*, 2000; Dunn *et al.*, 2008; Little and Chen, 2009; Brown and Brown, 2010) and tools (Yip *et al.*, 2008; Gouveia-Oliveira *et al.*, 2009; Bremm *et al.*, 2010) have been developed for such a purpose. MI measures the magnitude of co-variation between two random variables. The application of MI for the prediction of interacting residues starts with building a multiple sequence alignment (MSA), which ideally would include a large number of orthologous sequences. Each column in the MSA represents a random variable, and entropy for each column is determined to show the degree of variability in each MSA position. Following that, MI is calculated to quantify the co-variation between two positions. However, a column with 100% residue conservation will result in an entropy score of zero and, subsequently, a MI score of zero, too. This does not provide any insight toward the coevolution prediction of two residues. Hence, such columns were excluded from all analyses in this chapter. In general, MI scores range between 0 and 1; a MI score of 1 represents higher degree of coevolution between two positions, and zero signifies that no coevolution is found.

As pointed out by several authors (Fitch and Markowitz, 1970; Fitch, 1971; Atchley *et al.*, 2000), most amino acid sites are not completely independent of each other in terms of the (co)evolutionary pressures acting on them; rather they could contain signals attributable to structural and functional constraints or phylogenetic relationships among the species in a MSA. It is extremely important to remove these background signals prior to any in depth analysis, to ensure the resulting predictions reveal true coevolutionary signals. A recent study, carried out by Dunn *et al.* (2008), has shown a successful reduction of these biases by subtracting the average product of all column MI scores from each raw MI score of the same MSA. Substantial improvement for detecting residues that are in close proximity in folded protein structures was observed after the Average Product Correction (APC) method was applied. In their study, different interaction prediction methods were compared, and the MIp (the Mutual Information method employing APC) method was found to detect substantially more contacting pairs than other methods with similar accuracy. For instance, a homo-dimeric enzyme (triosphosphate isomerase) was utilized to examine the effectiveness of the APC approach, and several methods (Mir, OMES, McBASC and MIp) were compared. Indeed, a superior prediction performance was observed when using the MIp method, as

11 structurally contacting residue pairs were identified, while the least effective method, OMES, identified only 4 structurally contacting residue pairs. Furthermore, one MIp-predicted interacting residue pair was found to be in contact across the dimerization interface. Although it is possible to use the MIp method for intermolecular homo-dimerization interaction predictions, it would be rather difficult to distinguish the difference between dimerization and intra-subunit contacts without known structures.

The correlation-based approach epitomized by *mirrortree* (Pazos *et al.*, 2005) and its variants (Jothi *et al.*, 2005; Noivirt *et al.*, 2005; Sato *et al.*, 2005; Craig and Liao, 2007) is another commonly used measure to detect coevolution in protein sequences. However, this method does not work at the individual residue level. Typically, phylogenetic profiles for two protein families are determined, and the Pearson correlation coefficient is calculated based on the two profiles to assess the level of coevolution. However, high false positives rates have been reported to be a serious problem for this method (Sato *et al.*, 2005), and our own results (Chapter 2) also highlight its limitations. In an attempt to improve prediction quality by addressing this issue, Juan and colleagues (Juan *et al.*, 2008b) have developed a method that compares phylogenetic profiles from three protein families, instead of the conventional approach of using two. Highly specific coevolutionary signals (interactions between only two proteins) were determined by computing partial correlation scores, and more relaxed coevolution scores (interactions that form small protein complexes) were extracted from the ranked lists of partial correlation scores. According to Juan *et al.*, the partial correlation approach gave a considerable improvement in prediction results, with roughly double the accuracy of other coevolution-based prediction methods. Again, this approach claims to factor out confounding issues in the detection of protein coevolution and hence promises to be of use for prediction purposes.

Due to the suggested effectiveness of independent applications of both MI and partial correlation approaches for protein interaction predictions, for the current study, a strategy to combine the two methods was developed to facilitate the identification of intramolecular contacting positions. Following the comprehensive evaluation of the *mirrortree* approach, as described in Chapter 2 for the prediction of protein-protein interactions, it would be ideal to examine the above methods for intermolecular protein interactions (i.e. not homo-multimers and homo-dimers). However, such datasets are

hard to find and generate. In particular, there are considerable technical challenges in establishing orthology of proteins between species and in generating alignments, which are not required for intramolecular interactions. Therefore, since the primary aim of this chapter was to evaluate methods, rather than specific datasets, the same datasets used by Dunn *et al.* (2008) for evaluating the MI method were also applied for the current study to predict intramolecular protein interactions. Moreover, this allows any methodological improvements to be judged with consistency when compared to existing, published studies performed by Dunn and colleagues. Initially, MI scores were calculated for all position pairs in a protein, to assess whether the two positions in each pair were co-evolving. Subsequently, in a similar fashion to Juan *et al.* (2008b), three positions were compared, using partial correlation coefficients to assess whether the extra information of the third position would improve the interaction prediction. To minimize the background bias, all raw MI scores were corrected using the APC method before the subsequent calculations. In the end, scores for MI, correlation of MIs (MI-Correlatin) and partial correlation of MIs (MI-Partial correlation) were all compared, to establish the best approach for possible future applications. Furthermore, position combinations with the highest significant partial correlation coefficients were also evaluated to determine the effectiveness of the partial correlation level approach for predicting clusters consisting of three contacting positions. Additionally, this study investigated how different gap handling approaches, minimum entropy cutoffs and reduced alphabet schemes would affect the prediction results.

High prediction performance was achieved using a combination of MI and partial correlation. Of the three evaluated approaches for predicting intramolecular protein interactions, the MI-Partial correlation approach obtained the highest mean area under a ROC curve scores (AUC). The prediction performance was even further improved by the following: filtering highly conserved MSA regions with an entropy cutoff of 0.3; retaining all gaps in a MSA for the prediction; and removing the background signals using the APC method. The observed substantial improvement in prediction performance for the described methods suggests that this is indeed a useful approach for intramolecular interaction predictions.

4.3. Methods

4.3.1. Data

For comparison purposes, the same dataset containing 83 MSAs that was used previously in the study carried out by Dunn *et al.* (2008) was utilized for all analyses in this chapter. All MSAs consist of a minimum of 125 sequences, which was determined to be the minimum number of sequences one MSA should have in order to produce meaningful results (Martin *et al.*, 2005) for residue-residue contacting predictions. In parallel, the corresponding tertiary structures for all proteins were taken from the Protein Data Bank (PDB, Rose *et al.*, 2011).

Using tertiary structures, residues were defined as contacting if there was at least one contact between the non-hydrogen atoms within 12 angstroms (Å). Distance cutoff between two potentially contacting residues was relaxed from 6Å (the cutoff used by Dunn *et al.*) to 12Å, to ensure that there were still enough columns per MSA for the analysis after applying the rigorous *p*-value filtering criteria for the correlation (*p*-value cutoff $< 10^{-5}$) and the partial correlation (*p*-value cutoff $< 10^{-6}$) part of the analysis. This deliberately generous cutoff also can account for medium-range and indirect effects. For example, substitution of a small residue for a large residue (e.g. ALA to PHE) could cause an effect over such a distance. Equally, a broad definition such as this provides a robust definition for residue pairs deemed not to be interacting. The distances were determined using an in-house Fortran program, CONTA (Hubbard, personal communication).

4.3.2. Reduced Amino Acid Alphabet Schemes

As suggested by previous studies (Pollock *et al.*, 1999; Bacardit *et al.*, 2009), properly defined reduced amino acid alphabet groupings not only simplify the composition of the tested proteins in order to speed up the computation but may also improve coevolution predictions by enriching the signals above noise. Shown in Table 4.1, three reduced amino acid alphabet schemes were obtained from earlier studies (Attwood *et al.*, 1994; Rogov and Nekrasov, 2001; Tsai and Gerstein, 2002) and examined in this study. They were grouped based on stereochemical properties

(ALPHABET_1), volumes (ALPHABET_2) and amino acid residues similarity in natural protein sequences (ALPHABET_3), and consist of 7, 4, and 9 substitution characters, respectively.

ALPHABET_1	(Attwood *et al.*, 1994)	
Amino Acids	Substitution	Grouping criteria
A, I, L, M, V	A	Aliphatic
G, P	B	Special structure
C	C	Cysteine
F, W, Y	D	Aromatic
D, E	E	Polar negatively charged
H, K, R	F	Polar positively charged
N, Q, S, T	G	Polar neutral

ALPHABET_2	(Tsai and Gerstein, 2002)	
Amino Acids	Substitution	Grouping criteria
A, C, G, S	A	Volume $\leq$ 110 Å^3
D, N, P, T, V	B	110 Å^3 < Volume $\leq$ 140 Å^3
E, H, I, K, L, M, Q	C	140 Å^3 < Volume $\leq$ 170 Å^3
F, R, W, Y	D	170 Å^3 < Volume

ALPHABET_3	(Rogov and Nekrasov, 2001)	
Amino Acids	Substitution	Grouping criteria
E, K, Q, R	A	Charged
D, N	B	Polar
C	C	Cysteine
A, I, L, S, T, V	D	Hydrophobic (non-polar)
G, P	E	Special structure
M	F	Methionine
F, Y	G	Aromatic
W	H	Large
H	I	Histidine

Table 4.1. Three reduced alphabet groupings generated based on their stereochemical properties, volumes and amino acid residues similarity in natural protein sequences.

4.3.3. MSA Gap Handling Options

Gaps in MSAs can occur quite readily, and in different positions, depending on the multiple sequence alignment methods (which can align sequences differently and hence produce gaps at different positions). Similarly, gaps can confound conservations and coevolution metrics since there is no single "best" way to handle them. To reduce the complexity of the computation, many methods simply omit columns that contain gaps. However, it is to be expected that more distantly related orthologous sequences will produce more gapped MSAs, since they are less conserved. For example, when a MSA consisting of a large number of distantly related species is used for the analysis, removing all gapped columns can lead to a large portion of the MSA being removed and can possibly eliminate important signals. Thus, it is quite important to properly treat gaps in MSAs, where possible and appropriate.

Three different approaches were examined to account for missing residues. Shown in Figure 4.1, it is an example of a small artificial MSA consisting of four columns (C1 to C4) and thirteen rows (R1 to R13), with three gaps denoted as dashes at positions C1-R1, C2-R2 and C2-R9. In the example, each column represents a residue position, while each row is a different orthologue to the query sequence (first row). Upon application of the first gap handling approach, NO_GAPPED_COLUMNS, C1 and C2 were removed from the original MSA, resulting in only two columns, C3 and C4. Unlike the complete removal of gapped columns prior to the subsequent analysis in the NO_GAPPED_COLUMNS approach, the second gap handling method, NO_GAPPED_ROWS, performs the removal on a pairwise basis. Essentially, when two columns were compared, all gapped rows were excluded from the calculations of both entropy and MI scores. However, the same rows could be included for computing the statistics if gaps were no longer present in the same rows for a different column pair. For instance, R1, R2 and R9 would not be used toward the C1-C2 calculations, while only R1 would be excluded for the calculations for C1-C3. The last gap handling approach, 21_AMINO_ACID, simply assigns all gaps as the 21st amino acid, represented by Z; thus all columns would be retained.

1) Original MSA

	C1	C2	C3	C4
R1	-	V	V	L
R2	A	-	L	L
R3	T	N	L	L
R4	S	V	L	L
R5	S	L	V	L
R6	I	L	V	L
R7	S	I	V	L
R8	L	M	V	L
R9	N	-	L	L
R10	S	V	M	L
R11	D	Q	L	L
R12	I	L	L	L
R13	D	E	L	L

2) Gap handling approach I: NO_GAPPED_COLUMNS

	C3	C4
R1	V	L
R2	L	L
R3	L	L
R4	L	L
R5	V	L
R6	V	L
R7	V	L
R8	V	L
R9	L	L
R10	M	L
R11	L	L
R12	L	L
R13	L	L

3) Gap handling approach II: NO_GAPPED_ROWS

	C1	C2
R3	T	N
R4	S	V

	C1	C3
R2	A	L
R3	T	L

	C1	C4
R2	A	L
R3	T	L

R5	S	L
R6	I	L
R7	S	I
R8	L	M
R10	S	V
R11	D	Q
R12	I	L
R13	D	E

R4	S	L
R5	S	V
R6	I	V
R7	S	V
R8	L	V
R9	N	L
R10	S	M
R11	D	L
R12	I	L
R13	D	L

R4	S	L
R5	S	L
R6	I	L
R7	S	L
R8	L	L
R9	N	L
R10	S	L
R11	D	L
R12	I	L
R13	D	L

	C2	**C3**
R1	V	V
R3	N	L
R4	V	L
R5	L	V
R6	L	V
R7	I	V
R8	M	V
R10	V	M
R11	Q	L
R12	L	L
R13	E	L

	C2	**C4**
R1	V	L
R3	N	L
R4	V	L
R5	L	L
R6	L	L
R7	I	L
R8	M	L
R10	V	L
R11	Q	L
R12	L	L
R13	E	L

	C3	**C4**
R1	V	L
R2	L	L
R3	L	L
R4	L	L
R5	V	L
R6	V	L
R7	V	L
R8	V	L
R9	L	L
R10	M	L
R11	L	L
R12	L	L
R13	L	L

4) Gap handling approach III: 21_AMINO_ACID

	C1	**C2**	**C3**	**C4**
R1	**Z**	V	V	L
R2	A	**Z**	L	L
R3	T	N	L	L
R4	S	V	L	L
R5	S	L	V	L
R6	I	L	V	L
R7	S	I	V	L
R8	L	M	V	L
R9	N	**Z**	L	L

R10	S	V	M	L
R11	D	Q	L	L
R12	I	L	L	L
R13	D	E	L	L

Figure 4.1. Three approaches for handling gaps in a MSA. The original MSA, consisting of 4 columns and 13 rows is shown in 1). Modified MSAs resulting from the application of gap handling method: NO_GAPPED_COLUMNS, NO_GAPPED_ROWS and 21_AMINO_ACID are shown in 2), 3) and 4), respectively.

4.3.4. Mutual Information

Mutual information is a measure for determining the dependency between two random variables. It is based on Shannon's entropy information. In this study, a random variable is represented as a column in a MSA. MI determines whether the variability of one column is correlated to the variability of the other column. The mutual information statistic was calculated as follows:

$$MI(X:Y) = H(X) + H(Y) - H(X,Y)$$

where *H(X)* and *H(Y)* are marginal entropy for columns *X* and *Y*, and *H(X,Y)* represents the joint entropy for both columns.

The Shannon's entropy information theory was applied as:

$$H(X) = \sum_i p(x_i) \log p(x_i)$$

$$H(Y) = \sum_j p(y_j) \log p(y_j)$$

$$H(X,Y) = \sum_{i,j} p(x_i, y_j) \log p(x_i, y_j)$$

where $p(x_i)$ and $p(y_i)$ are the probability distributions of residue type *i* in columns *X* and *Y*, respectively. The joint probability distribution of residue type *i* and *j* in column *X* and *Y* is represented as $p(x_i, y_j)$.

4.3.5. Average Product Correction (APC)

Background bias, such as phylogenetic relationships among the species used in multiple sequence alignments, can significantly mask coevolutionary signals and reduce the deducibility of correct interactions. Thus, the average product correction (APC) developed by Dunn *et al.* (2008) was adapted in an attempt to remove background bias and, therefore, facilitate the identification of true interacting residue pairs. The equation for computing the background-bias corrected mutual information, MI_C, is shown below:

$$MI_C = MI_R - MI_{APC}$$

where MI_R is the raw MI statistic calculated using the equation in 4.3.4 and MI_{APC}, representing the background signal, is denoted as:

$$MI_{APC} = \frac{MI(\overline{X})MI(\overline{Y})}{\overline{MI}}$$

Here, the product of mean MIs for column *X* and column *Y* is divided by the overall mean ($\overline{MI}$) to yield MI_{APC}. It should be noted that $\overline{MI}$ is calculated based on all columns in the MSA, which also included columns *X* and *Y*.

4.3.6. Correlation and Partial Correlation of Mutual Information

To further improve the accuracy of interaction predictions, a partial correlation approach was implemented. This is the first time this type of approach has been applied to mutual information scores, as Juan and colleagues used it when applied to *mirrortree* whole sequence based approaches. Although each MI score quantifies the co-variability between two positions in a MSA, it is also intriguing to know whether similar patterns among these scores can be detected to help improve the predictions, as such patterns might be dictated or influenced by a third position. In principle, when two residues from a predicted interacting residue pair are found to have similar relationships with all the other residues in the same MSA, the likelihood of the two residues interacting is

potentially higher. This is intuitive when one considers real proteins, as unlike complementary base pairing in DNA and RNA, pairwise interaction between amino acids is too simplistic a general model, and typically most residues interact with more than one partner. As shown in Figure 4.2, the correlation between MI scores for positions *i* and *j* can be calculated using all the corresponding MI scores, listed in the same order, versus all other positions. The example shows that all MI scores for positions 5 (*i*) and 2 (*j*) were sorted based on a consistent third position order (i.e. 1, 3, 4 and 6) to produce comparable MI patterns for calculating a correlation coefficient, $r_{5,2}$. Furthermore, to calculate partial correlation of the same *i* and *j* positions based on a third position *k* of 1, $r_{i,j}$ ($r_{5,2}$), $r_{i,k}$ ($r_{5,1}$) and $r_{j,k}$ ($r_{2,1}$) were first calculated following the correlation approach described above. Subsequently, the three r scores can be used to calculated a partial correlation coefficient ($r_{5,2,1}$).

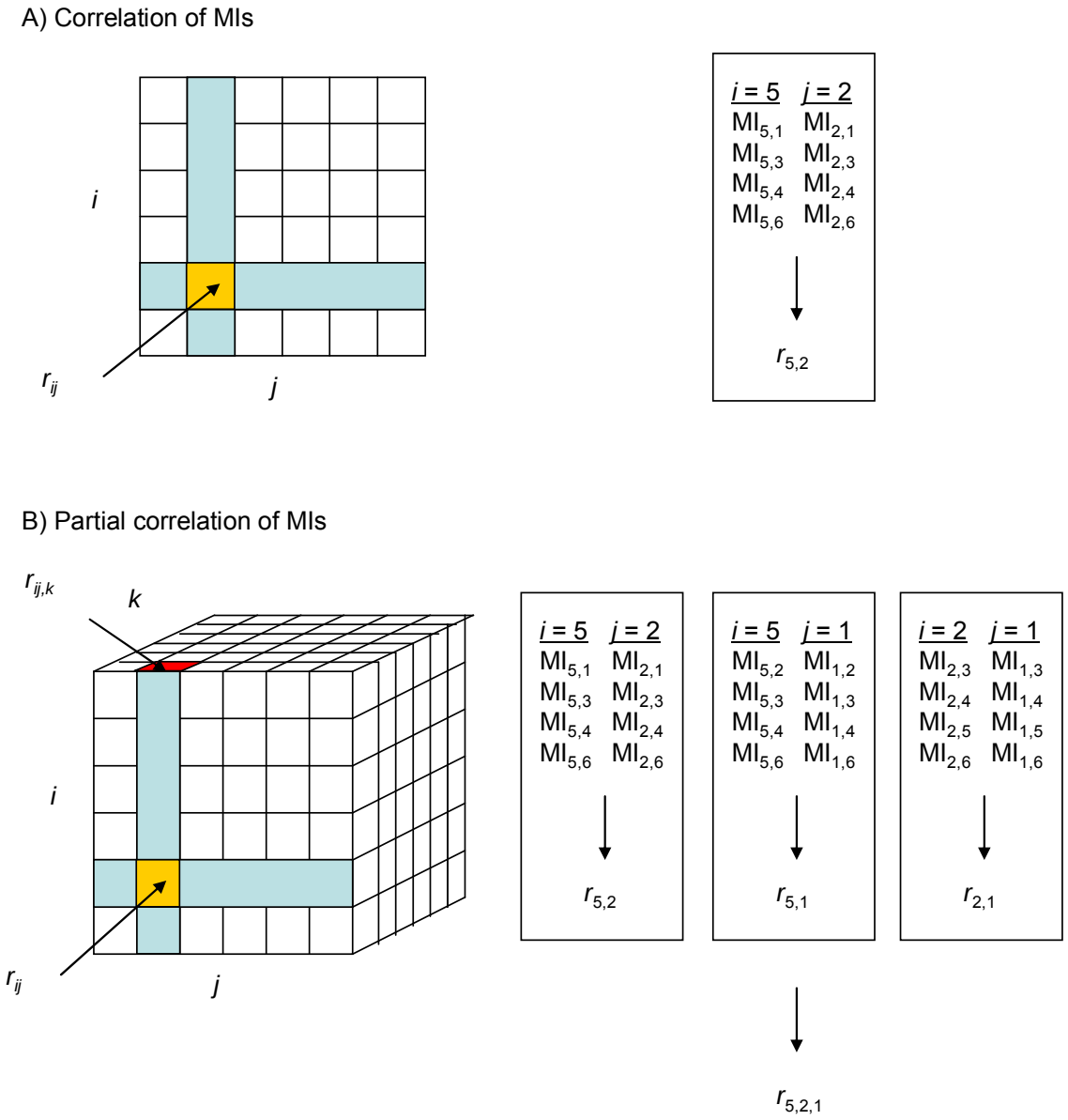


Figure 4.2. Example for calculating correlation and partial correlation of MI scores. Each square represents a MI value for two positions. A) For the correlation of MIs approach, only two positions (*i* and *j*) are considered. Correlation coefficient ($r_{i,j}$), shown in yellow, can be calculated using all $MI_{i,j}$ values from the corresponding columns and rows. B) For the partial correlation of MIs approach, in addition to positions *i* and *j*, a third position, *k*, is also utilized. Partial correlation, $r_{i,j,k}$, can then be calculated based on $r_{i,j}$, $r_{i,k}$ and $r_{j,k}$.

Correlation of MI scores for MSA positions *i* and *j*, $r_{i,j}$, was computed using the Pearson correlation coefficient as follows:

$$r_{ij} = \frac{\sum_{x=1}^{n}(MI_{ix} - \overline{MI}_{ix})(MI_{jx} - \overline{MI}_{jx})}{\sqrt{\sum_{x=1}^{n}(MI_{ix} - \overline{MI}_{ix})^2}\sqrt{\sum_{x=1}^{n}(MI_{jx} - \overline{MI}_{jx})^2}}$$

Partial correlation of MI scores for MSA positions *i* and *j* with respect to position *k*, $r_{i,j,k}$, was calculated as follows:

$$r_{i,j,k} = \frac{r_{ij} - r_{i,k}r_{j,k}}{\sqrt{(1 - r_{i,k}^2)(1 - r_{j,k}^2)}}$$

Following the same practice carried out by Juan *et al.*, all position pairs with a *p*-value equal to or larger than 10^{-5} (for the correlation analysis) and 10^{-6} (for the partial correlation analysis) were removed from the analysis to ensure that all correlation and partial correlation results were not generated by random chance. Such low *p*-values should greatly reduce the likelihood of obtaining false positive correlations obtained by chance.

4.3.7. Partial Correlation Level

In general, one can consider two types of intramolecular interactions. One is specific, where two residues interact exclusively with each other, while the other type involves multiple residues interacting to form clusters. Most intramolecular residue interactions in folded proteins are expected to fall into the latter type, as they must form direct contact with other residues to maintain the integrity of their 3-dimensional structure. The partial correlation level approach proposed by Juan *et al.* (2008b) has already demonstrated the capability of the method in determining interacting protein clusters, based on putative interactions predicted to occur between whole protein sequences. It is because of this potential that the approach is utilized here as an extension of the mutual information approach for detecting clustered residues.

Although it is possible to consider multiple levels of partial correlation, owing to the large computational power required, only partial correlation level 1 was implemented here. After removing all non-significant partial correlation results, filtered using a *p*-value of 10^{-6}, all remaining partial correlation scores for each position *i* and *j* pair were ranked for each variable position, *k*. From each ranked list, the position combination with the largest significant partial correlation (1st partial correlation level) was extracted to form the partial correlation level list. All combinations in the partial correlation level list represent clusters consisting of three contacting residues. An illustration of the process is shown in Figure 4.3.

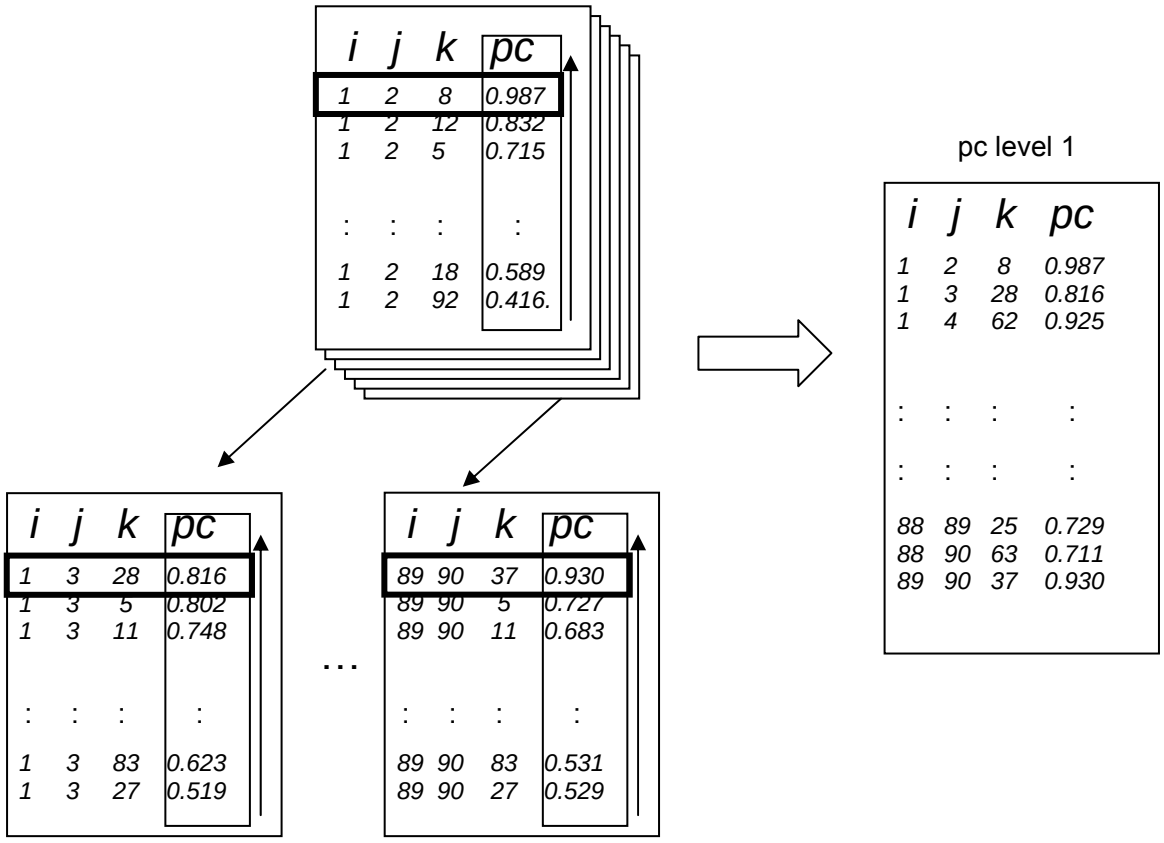


Figure 4.3. Depiction of the process for ranking partial correlation scores to obtain partial correlation level 1 results. On the left side of the figure, all partial correlation (pc) scores with the same *i* and *j* were placed in the same list and ranked in descending order. Then the top combination from each ranked list was extracted and added to the partial correlation level 1 list.

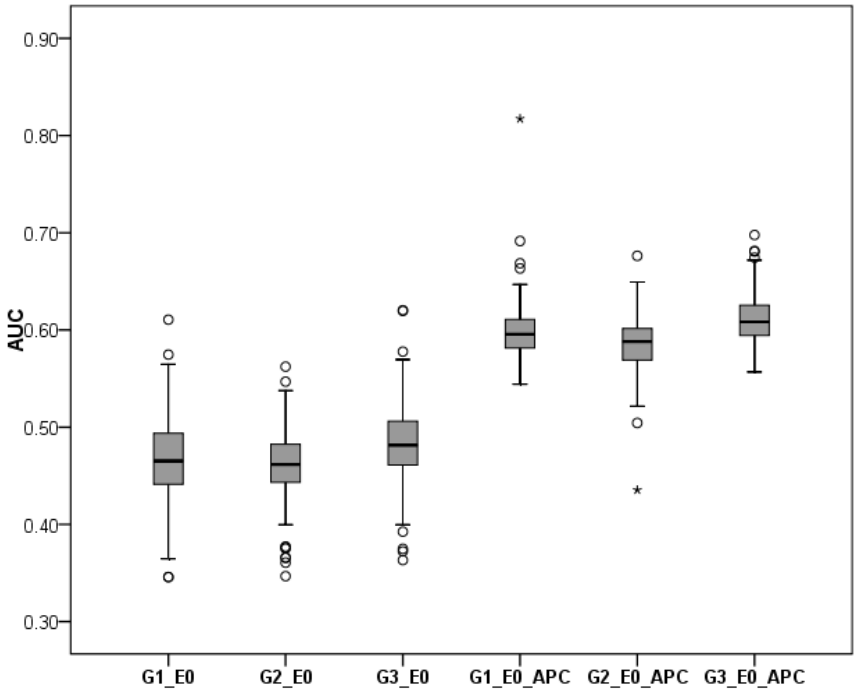
4.4. Results and Discussion

4.4.1. Mutual Information

4.4.1.1. Effect of the APC Method

In order to evaluate the APC method (Dunn *et al.*, 2008) properly, the same dataset, consisting of 83 MSAs with a minimum of 125 sequences each, used by Dunn *et al.*, was utilized for all analyses in this study. MI scores were calculated for all possible position pair combinations for each MSA and, to enhance the detection of true coevolutionary signal, the APC method was applied to facilitate the removal of background signals. Dunn and colleagues have shown that, when MI scores were used alone, without any effort of background bias removal, the method performed relatively poorly and detected the least number of contacting residues. By contrast, after applying the APC method, an increase of up to three or four times in the number of contacting residues was identified, without apparently compromising the accuracy. Essentially, they claim that APC is an effective method for improving coevolutionary signals derived from protein MSAs for residues in contact. Following their method, the MI analysis was carried out, and the atomic distance between two positions was determined for each residue pair. In principle, contacting residues should have relatively higher MI values when compared to non-contacting residues. The effectiveness of this approach was evaluated using the AUC statistic, with higher AUC values signifying better performance in predicting putative contacting residues.

A) Entropy cutoff of 0 (no entropy filtering)



B) Entropy cutoff of 0.3

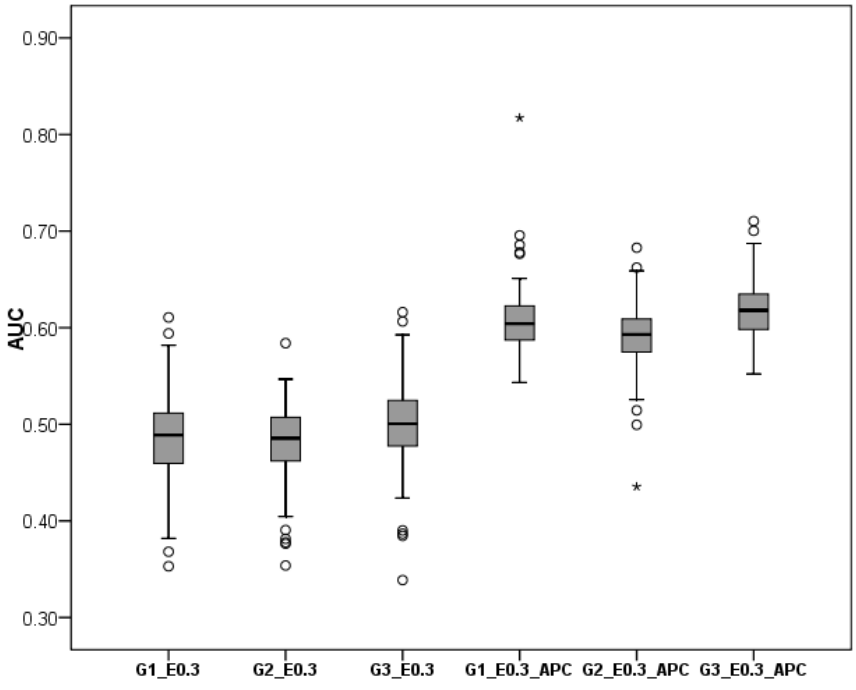


Figure 4.4. Box plots showing the differences of AUC scores for interaction predictions made based on different gap handling approaches and entropy cutoffs. The three gap handling approaches are NO_GAPPED_COLUMNS (G1), NO_GAPPED_ROWS (G2) and 21_AMINO_ACID (G3). The two entropy cutoffs are 0 (E0) and 0.3 (E0.3). Furthermore, the APC background signal correction method was also utilized. The results were grouped and plotted based on A) entropy cutoff of 0 (no entropy filtering) and B) entropy cutoff of 0.3.

In the current study, the APC method does appear to be removing background signals quite efficiently, as all mean AUC scores that were determined employing the APC approach were consistently higher than the AUC scores for the non-APC results for the same categories. The analysis results are shown in Figure 4.4. It should be noted that the whiskers in the box plot represent the interquartile range between the 25th and 75th percentiles, and the same definition applies for all box plots shown in this chapter. As shown in Figure 4.4A, the mean AUC scores for G1_E0_APC, G2_E0_APC and G3_E0_APC are respectively 0.601, 0.584 and 0.612, while the mean AUC scores for the equivalent non-APC sets are 0.468 (G1_E0), 0.460 (G2_E0) and 0.484 (G3_E0). The differences between all equivalent APC and non-APC sets are significant, as all the *p*-values are smaller than 0.01 (calculated using t-tests). Furthermore, the same trend with significant AUC differences between equivalent APC and non-APC sets can also be found in Figure 4.4.B, where an entropy cutoff of 0.3 was applied to the same datasets. Moreover, AUC scores for the non-APC sets were all roughly 0.5, which suggests that predictions generated using this approach can be achieved by randomly selecting residue pairs. Indeed, APC improves the predictions to a level above random, as AUC scores for all APC sets in Figure 4.4 are above 0.5, with the highest score being 0.619 for the G3_E0.3_APC set.

4.4.1.2. Gap Handling Methods

As suggested by a number of studies (Martin *et al.*, 2005; Buslje *et al.*, 2009), the minimum number of orthologous sequences required to build a MSA for meaningful coevolution predictions ranges from 125 to 400. With the constant improvement of sequencing technology, protein sequences for more and more species are readily becoming available. Inclusion of sequences derived from a large number of species should potentially reveal a more accurate and complete evolutionary history. However, somewhat perversely, since many orthologous sequences are possibly derived from distantly related species, the more that are included, the harder a MSA can be to accurately align, and, consequently, this results in more gapped positions. Therefore, if not handled properly, evolutionary signals in MSAs can be reduced or, in more extreme cases, even be abolished, due to improper handling of gapped columns.

In this study, after evaluating three gap handling methods, it is evident that 21_AMINO_ACID is superior to the other two methods, at least for the purposes of MI, as illustrated in Figure 4.4. The order of the methods from the best to worst performance is: 21_AMINO_ACID > NO_GAPPED_COLUMNS > NO_GAPPED_ROWS. Independent of whether the APC was used, or which entropy cutoff was applied, the mean AUC score for 21_AMINO_ACID was always the highest within each group. This result, however, is perhaps not surprising because 21_AMINO_ACID probably generated MSAs with the highest sequence diversity. The sequence diversity experiment in the *mirrortree* benchmarking study (refer to Chapter 2) demonstrated that MSAs with higher sequence diversity tend to result in more accurate predictions. Moreover, unlike NO_GAPPED_COLUMNS and NO_GAPPED_ROWS, no residues or sequences were removed when 21_AMINO_ACID was applied, so no accidental removal of coevolutionary signals would have taken place. Indeed, this would also be the method with the most theoretical "power", since it retained the most data. Finally, the least amount of improvement after applying the APC method was observed for the NO_GAPPED_ROWS results. One possible elucidation of this finding is that NO_GAPPED_ROWS tends to generate different lengths of columns for different position pairs and, perhaps, this inconsistency leads to the poor performance of the method. After applying t-tests to compare the mean AUC scores between different sets, NO_GAPPED_COLUMNS and NO_GAPPED_ROWS do not appear to be significantly different, as all *p*-values between sets that were treated using these two gap handling methods are all larger than 0.01. However, consistently, significant differences between NO_GAPPED_ROWS and 21_AMINO_ACID sets were observed. This indicates that different gap handling methods could result in significantly different prediction results.

4.4.1.3. Entropy Filtering

When MSA columns are highly or completely conserved, low entropy is present, and very little to no coevolutionary signals can be detected between these columns using MI. In the case of complete residue conservation, entropy for such columns is zero. To assess how residue conservation would affect prediction results, an entropy cutoff value of 0.3 was implemented as a comparison to the non-entropy-filtered MSA

results. This was done by removing all columns with entropy values lower than the cutoff.

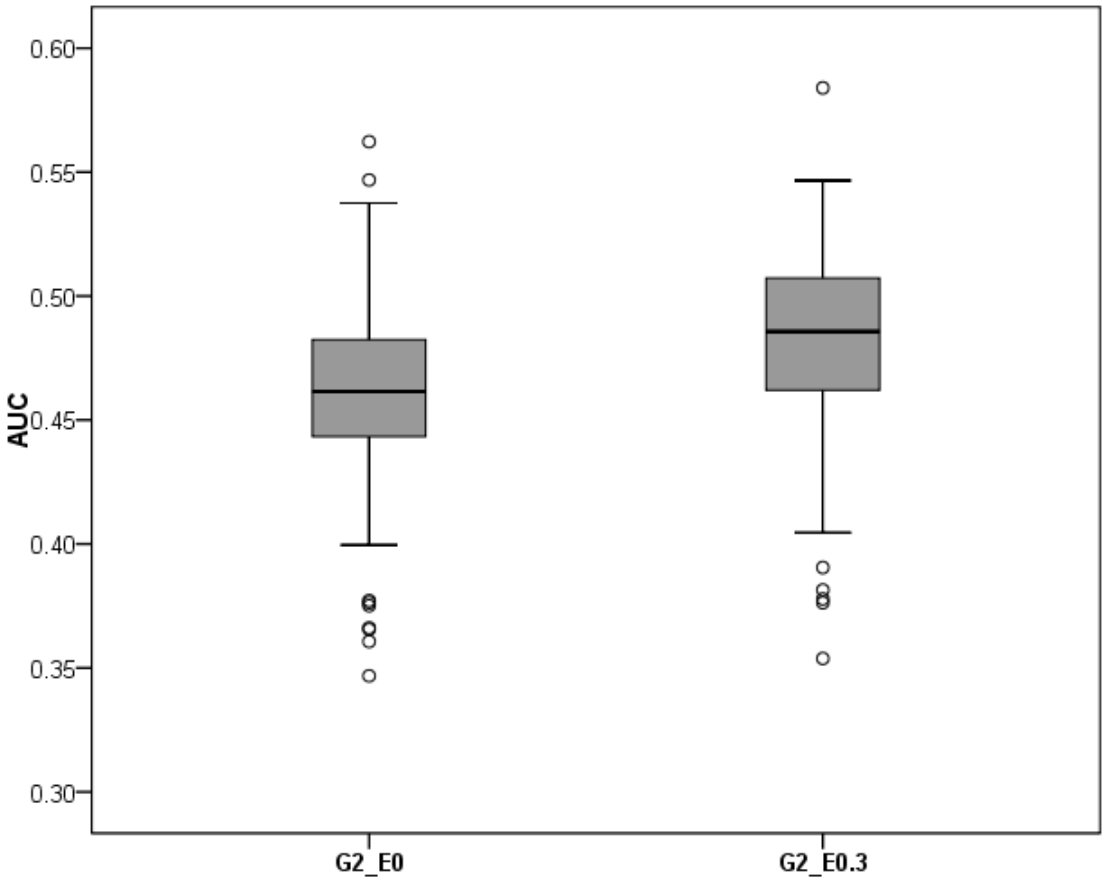


Figure 4.5. Box plot showing the differences of mean AUC scores for interaction predictions utilizing the NO_GAPPED_ROWS gap handling method and entropy cutoffs of 0 and 0.3.

The application of the entropy cutoff of 0.3 has a positive effect on the predictions, since improved AUC scores were observed for all entropy-filtered sets. It was suggested by Dunn *et al.* that entropy filtering was not necessary for the APC approach, as the APC method alone should sufficiently remove the influence of entropy. They showed that the MIp values for all position pairs in a MSA remained constant, while the entropies varied from 0.6 to 1.6, indicating the independence of the MIp values and entropies. However, in this study, a small improvement was still observed for the APC results. Certainly, more dramatic improvements were seen in the non-APC sets, with the largest difference being between G2_E0 and G2_E0.3 sets, shown in Figure 4.5. However, the small differences between the entropy filtered and non-entropy filtered prediction results are not statistically significant. Nevertheless, the consistently

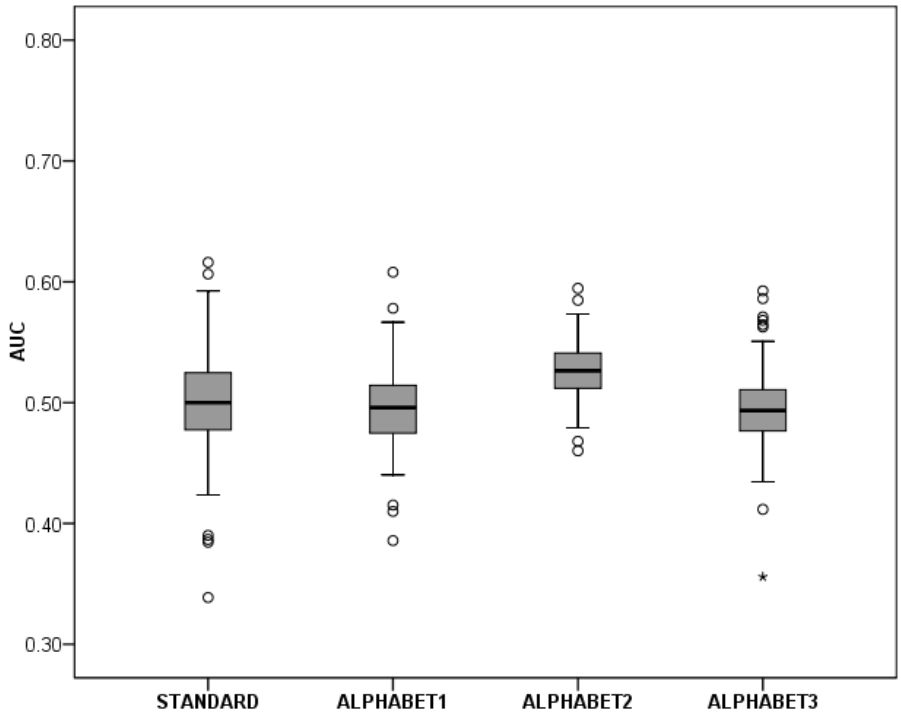
higher AUC scores for the entropy-filtered results would still suggest that the minimum sequence variability should be dealt with as a type of background bias or noise, which, when properly accounted for, can allow prediction accuracy to be improved.

4.4.1.4. Alphabet Reduction

Also investigated was the effect of different amino acid alphabet groupings on the predictions, based on the assumption that reduced alphabets might capture essential physicochemical features of the amino acids and therefore enhance coevolutionary signals. The best performing amino acid grouping for the non-APC results was ALPHABET_2, with approximately a 5.86 percent improvement from the STANDARD amino acid grouping (Figure 4.6A). ALPHABET_2 was also found to be significantly different from all the other alphabet groupings, as it is the only alphabet grouping in the non-APC set to have an AUC score above 0.5.

In contrast to the non-APC results, ALPHABET_2, based on residue volume, resulted in the lowest mean AUC score (0.565) when APC was applied (Figure 4.6B). This finding was not unexpected, as ALPHABET_2 contained the lowest number of characters (4). A study carried out by Bacardit *et al.* (2009) generated reduced alphabet sets consisting of 2 to 5 characters, with the results evaluated based on the number of nearest neighbours of a residue and on the solvent accessibility of residues. In their study, the results obtained using the 5 character alphabet showed accuracies similar to those achieved using the standard twenty amino acid alphabet. Since the number of reduced alphabets defined in ALPHABET_2 is below the theoretical number of characters required for reasonable prediction accuracy proposed by Bacardit *et al.*, important coevolutionary signals may be lost from the over simplification.

A) No correction



B) APC

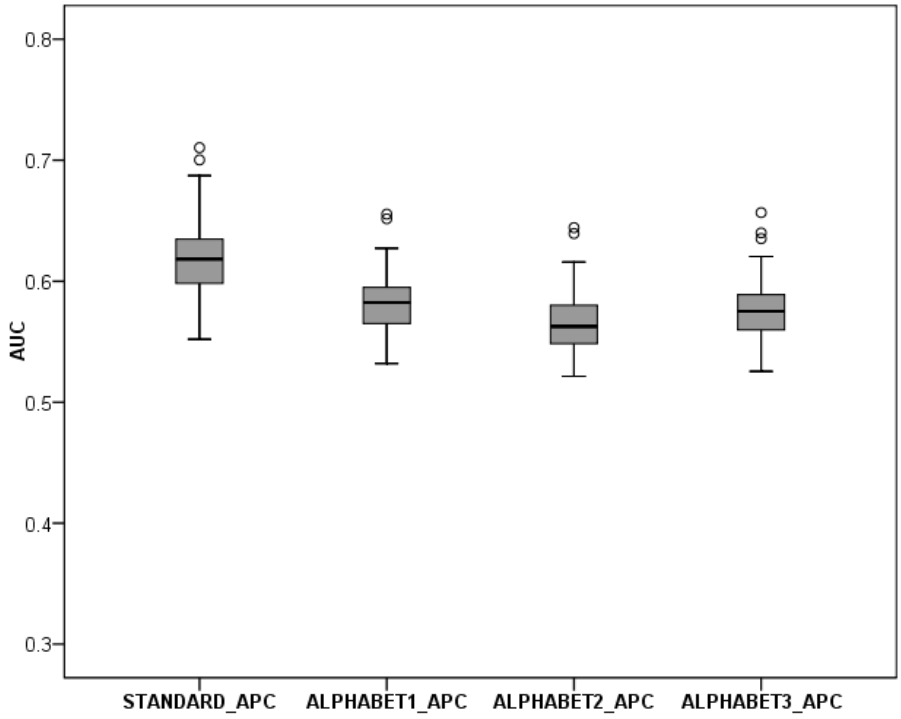


Figure 4. 6. Box plots showing the differences of AUC scores for assessing the effect of the standard 20 amino acid alphabet and the three amino acid reduction alphabet groupings. These results were based on the 21_AMINO_ACID dataset and filtered using an entropy cutoff of 0.3. The APC method was also applied, and the results for the non-APC-treated results are shown in A), with the APC results shown in B).

Consistent with earlier analysis results, all AUC scores for the APC set (Figure 4.6B) are higher than the non-APC results (Figure 4.6A). In the APC set, the standard twenty amino acid alphabet grouping performed the best, as the mean AUC score obtained for the STANDARD set was the highest (0.619) when compared to the three tested reduced alphabet groupings. Furthermore, all mean AUC scores in this set were found to be significantly different. Although no improvement was observed for the three reduced alphabet groupings assessed here (which were determined based on: stereochemical properties (ALPHABET_1), volumes (ALPHABET_2) and amino acid residues similarity in natural protein sequences (ALPHABET_3)), this finding does not necessarily suggest that no reduced alphabets should ever be applied. Instead, an optimal set of reduced alphabets should perhaps be generated using different features, such as size and charge characteristics (Pollock *et al.*, 1999) or information theory (Bacardit *et al.*, 2009), since analyses using these features have shown similar or slightly improved performance between their reduced alphabet and the standard twenty amino acid alphabet results.

Because the best prediction was achieved when using: the APC, 21_AMINO_ACID gap handling approach, entropy cutoff of 0.3 and the standard 20 amino acid alphabet, all onward analyses presented in this chapter were completed by applying these specifications.

4.4.2. Coevolution Detection Using Correlation and Partial Correlation of Mutual Information

After MI scores are computed for all position pairs in a protein MSA, it is in principle relatively easy to identify contacting residue pairs, based on the assumption that they usually have higher MI scores than non-contacting pairs in the same protein. However, the predictions are not error-free, as the best mean AUC score was determined to be 0.619, instead of 1, in this analysis, which is far from perfect. Thus, two correlation-based methods were implemented to attempt to improve the predictions. The general principle on which this strategy is based is that truly interacting positions must show common mutual information with other, contacting positions, especially with respect to all others in the aligned protein family. MI-Correlation and MI-Partial

correlation, respectively, were applied to calculate the correlation and partial correlation statistics for comparing all MI scores. For the MI-Correlation method, when two positions have similar relationships with other positions in terms of MI scores, they are more likely to be interacting. The similarity of MI patterns can be measured by using the Pearson correlation coefficient. In addition to comparing only two positions for the correlation approach, the MI pattern of a third position can also compared to facilitate the prediction for the MI-Partial correlation method.

The analysis results, shown in Figure 4.7, indicate that both the MI-Correlation and MI-Partial correlation methods have higher predictive power than the MI method, as the mean AUC scores for MI-Correlation (0.713) and MI-Partial correlation (0.731) were significantly higher than the MI AUC score (0.621), with *p*-values of 3.25×10^{-12} and 1.11×10^{-16}, respectively. Moreover, a small improvement was also found for the MI-Partial correlation method when compared to MI-Correlation. However, larger variations in the prediction results were also observed for the methods with higher AUC scores. As shown in Figure 4.7, both the MI-Correlation and MI-Partial correlation methods resulted in much larger standard deviation values than the MI method. Nevertheless, it appears that by comparing patterns of MI scores, whether between two or three positions, a substantial improvement in prediction performance could be achieved.

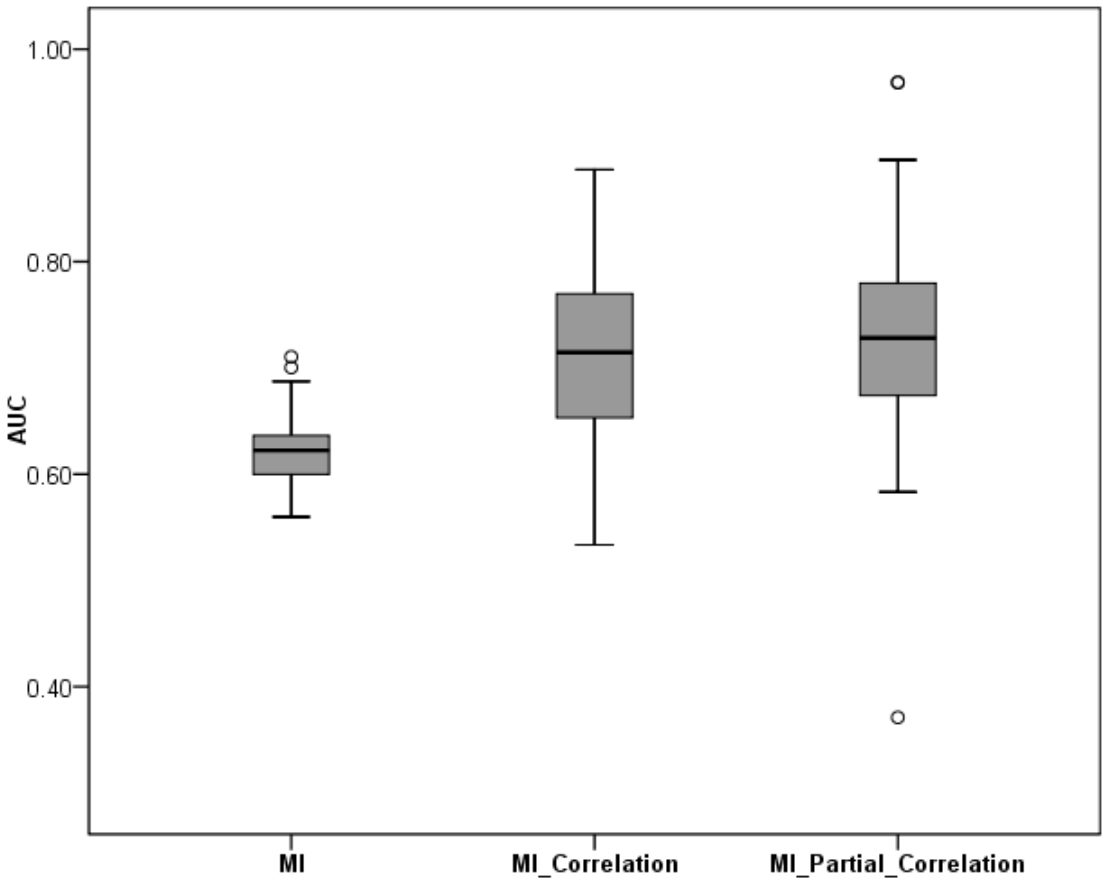


Figure 4.7. Box plot for the evaluation of intramolecular interactions generated using methods: MI, MI-Correlation and MI-Partial correlation. Proteins with number of contacting pairs <= 5 or number of non-contacting pairs <=5 were excluded from the mean AUC calculations.

In comparison to the more substantial improvements observed for the APC-treated results, smaller improvements were observed without the application of the APC method (Figure 4.8). Although the non-APC set follows the same trend as the APC set, the predictions generated by the non-APC methods appear to be quite random, as the mean AUC scores are close to 0.5. In contrast, the significantly higher mean AUC scores for the APC set resulted in more substantial improvements. As shown in Figure 4.8, a 23% improvement was observed for the MI set when the APC method was applied. Such a difference between the equivalent non-APC and APC results was increased to 31% for the MI-Correlation set. Furthermore, an even larger difference of 36% was found for the MI-Partial correlation set. Having been shown repeatedly to yield more accurate predictions, the APC method indeed appears to be removing background signals quite effectively. The results demonstrate that the background signal

removal step is extremely important for coevolution-based analysis. If such error is not removed from raw MI scores and carried forward to the correlation and partial correlation analyses, the level of bias is going to increase, since the MI-Correlation and MI-Partial correlation methods are based on MI scores.

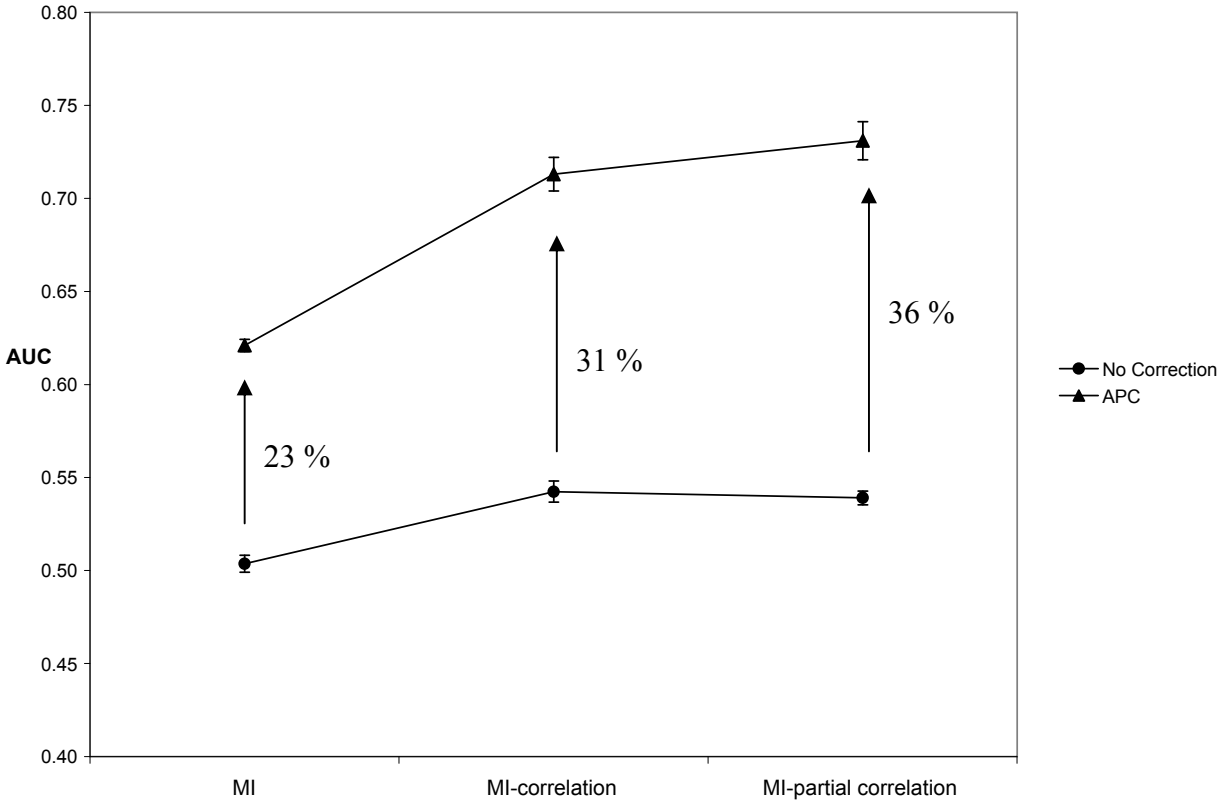


Figure 4. 8. Comparison of prediction improvement trends for MI, MI-correlation and MI-Partial correlation. Error bars are plotted to indicate standard error of the mean for each set.

4.4.3. Residue Cluster Prediction via the Partial Correlation Level Approach

Although the APC and MI-Partial correlation methods show promise, ultimately one would wish to be able to identify more than paired residues across an interacting protein interface. Instead, one would aim to find clusters or groups, which are expected to form the interface. To this end, the partial correlation level approach to identify potential clusters of interacting positions involving three residues was investigated. However, only partial correlation level 1 was implemented, because of the extensive

computational power required. This decision should lead to little impact on the predictions since, in theory, positions with the highest significant partial correlation scores should represent contacting clusters. To test the proposed theory, the highest significant partial correlation score was extracted from each ranked list and compared to the average distance of the three corresponding residues. In principle, all clusters extracted should be contacting, since they would all have significantly high partial correlations. However, similar to many existing interaction prediction methods, false positives would be expected, and indeed do exist. The accuracy of the method was measured by calculating the fraction of true contacting clusters. When the average atomic distance among the three residues in a cluster is 12Å or less, they are considered to be a true contacting cluster. If the average atomic distance among the three residues in a cluster is larger than 12Å, the residues of the cluster are considered to be non-contacting. Figure 4.9 shows the average accuracy for this analysis was 0.619. This outcome could be explained by two possible causes. First, a large proportion of clusters with low partial correlation scores but significant correlation *p*-values were included. To correct for this inclusion of false positives, perhaps, another filtering step should be incorporated to remove clusters with low significant partial correlation scores. Or, conversely, these residues could be truly co-evolving, but, rather than being caused by structural constraints, they were the results of functional constraints. It has been reported by a few groups (Pritchard and Dufton, 2000; Gloor *et al.*, 2005) that clusters of residues could co-evolve without being in direct contact. Furthermore, Gloor and colleagues also suggested that these residues tend to be near binding regions or active sites.

Nevertheless, when only considering combinations contributed to the top 50 highest partial correlation scores for each protein MSA, the average accuracy increased to 0.806 (Figure 4.9). When considering combinations with slightly lower partial correlation scores, the average accuracies are still quite high. The average accuracy scores for the top 100, 200 and 300 are 0.779, 0.743 and 0.719, respectively. Since it is not an easy task to separate structurally coevolving residues from false positive predictions or functionally coevolving residues, utilizing predictions with the highest partial correlation scores (i.e. the top 50 or top 100 scores for each protein MSA) ensures the quality of predictions, since larger partial correlation scores are often associated with real contacting positions.

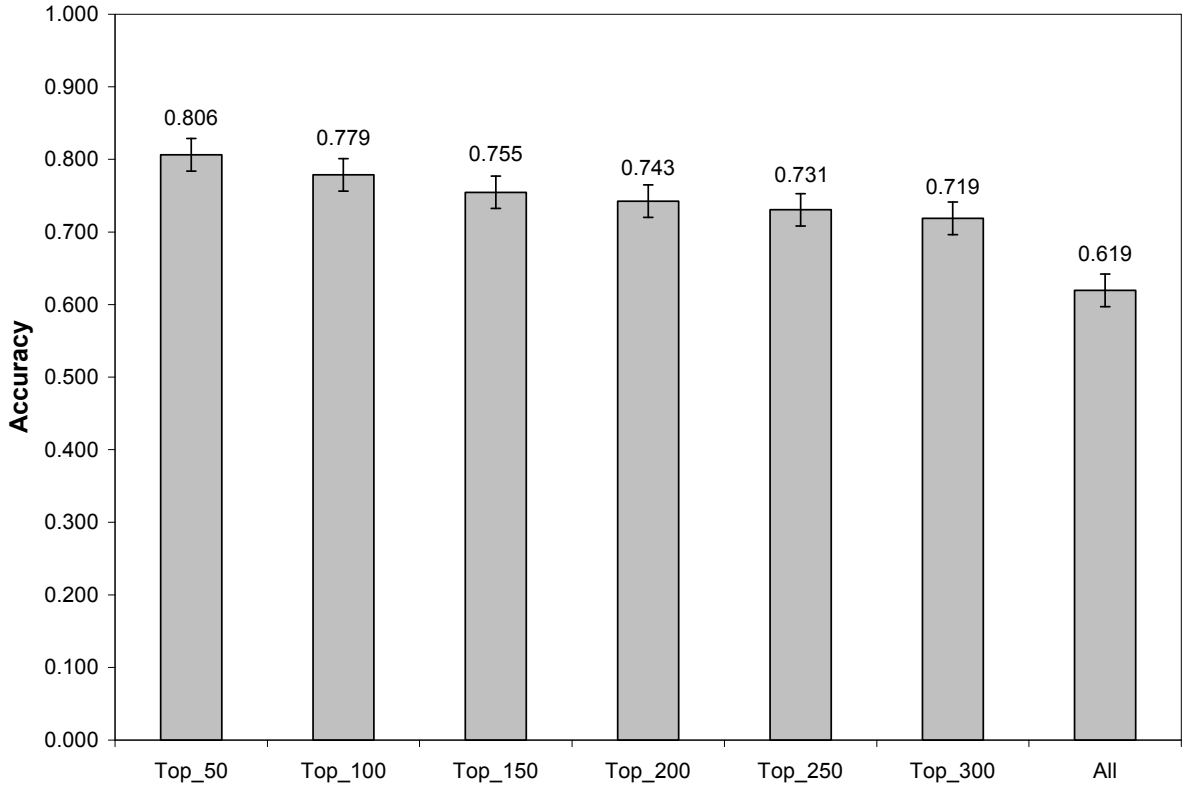


Figure 4. 9. Mean accuracies for the partial correlation level analysis. Accuracies were determined by counting the number of position combinations with an average atomic distance equal to or less than 12 Å and then dividing it by the total number of position combinations. Error bars are plotted to indicate standard error of the mean for each set.

As an example to demonstrate the effectiveness of the partial correlation and partial correlation level approaches in identifying contacting residues, a representative protein with a relatively large number of corresponding contacting residues in an associated protein structure was further analyzed. The prediction results for mouse 3' and 5'-cyclic phosphodiesterase 2A (PDE2A_MOUSE) are discussed below in detail. The MSA (1MC0_noXpfam01590) utilized in this study corresponds to positions 402-541 in the PDB structure of 1MC0. After applying gap handling option 21_AMINO_ACID and filtering out highly conserved columns with an entropy less than 0.3, partial correlation scores were generated for the remaining positions, resulting in 1,236,699 combinations. Following that, combinations with non-significant partial correlation scores were removed, and 108,331 combinations were retained for the subsequent analysis. Subsequently, atomic distances for all remaining combinations were determined using CONTA.

In this example, all contacting positions were further categorized into "definitely contacting", for positions with a separating distance less than 6Å, and "maybe contacting", for positions with a separating distance less than 12Å but equal to or greater than 6Å. All position pairs with distance of 12Å and larger were considered "not contacting". As shown in Figure 4.10 the mean partial correlation coefficient for the definitely contacting positions was 0.637, and the "maybe contacting" position set had an average partial correlation coefficient of 0.515. Both values are significantly higher than the mean partial correlation coefficient for the non-contacting position set (-0.199). The large partial correlation coefficient differences between the contacting and non-contacting results suggest that the partial correlation approach is indeed very effective in distinguishing contacting and non-contacting positions. It should also be noted that the standard deviation for the non-contacting positions was substantially larger than standard deviations for definitely contacting positions. Also, this remained true, to a lesser extent, for the "maybe contacting" positions, showing higher variations in the non-contacting predictions.

For analyses prior to the partial correlation method, all distances were determined based on only two positions, since binary interacting relationships were the main focus. In the partial correlation level approach, average distances were computed using three positions to estimate the closeness of the residues in each predicted cluster. After extracting combinations with the highest significant partial correlation coefficient from each ranked list, 767 three-residue clusters were identified.

An interesting observation was recognized. The average distances for three-residue clusters (positions *i*, *j* and *k*) were substantially larger than the average distances for two contacting positions (positions *i* and *j*), especially for the predictions with higher partial correlation coefficients (Figure 4.11). For the top 50 contacting results, the average distance calculated based on *i, j and k* positions was approximately double the average distance for *i* and *j* positions. This suggests that partial correlation level 1 analysis mostly identifies clusters with two very close positions and a third position being slightly further away. As shown in the study carried out by Juan *et al.* (2008b), partial correlation level 1 might not necessarily result in the best predictions in all cases. Instead, the best predictions were observed for the 10^{th} level in their analysis. Perhaps, one approach to improve the partial correlation level predictions is to determine results

based on few partial correlation levels, rather than just level 1. Similar to the trend shown in Figure 4.9, predictions for the highest partial correlations were more accurate, as the top 50 and top 100 predictions were shown to have relatively shorter average distances.

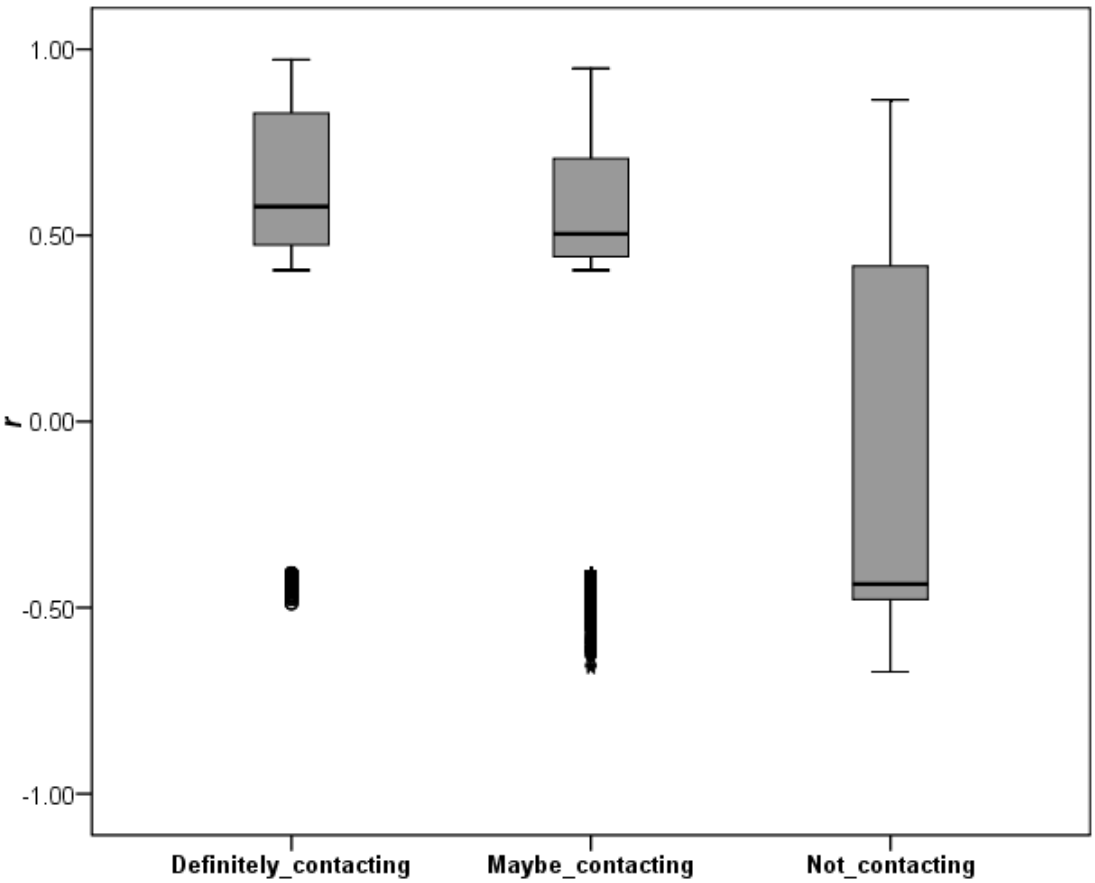


Figure 4.10. Box plot showing partial correlation analysis results for 1MC0_noXpfam01590.

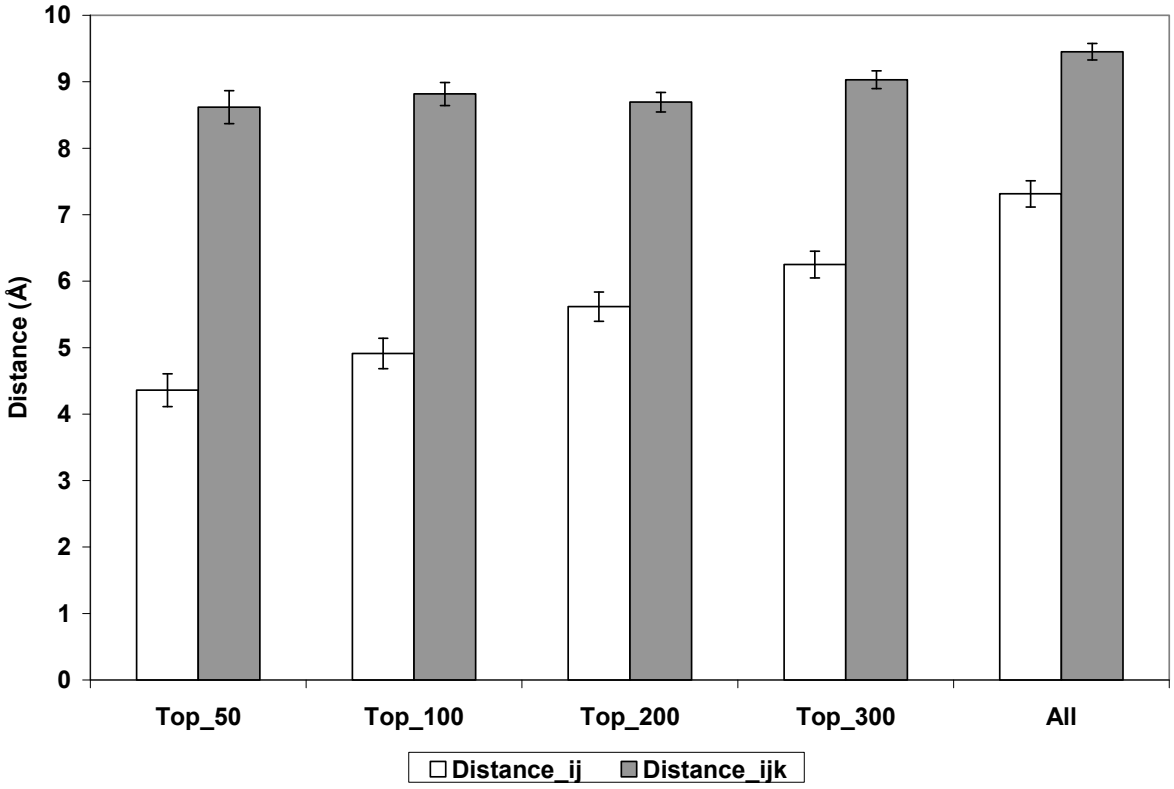


Figure 4.11. Mean distances for the partial correlation level analysis for 1MC0_noXpfam01590. Mean distances computed based on positions *i* and *j* are represented by the white bars, and other distances based on positions *i*, *j* and *k* are represented by the grey bars. The average ij and ijk distances for all 767 three-residue clusters were utilized for the 'All' category in the graph.

4.5. Summary

In this chapter, the effectiveness of using the mutual information, correlation and partial correlation statistics for predicting contacting positions was assessed. In addition, many important factors (i.e. gap, entropy and reduced alphabet) associated with MSAs were evaluated. To determine whether there is an association between entropy and prediction performance, an entropy cutoff of 0.3 was applied to remove highly conserved regions in a MSA prior to the prediction of contacting positions in a protein. Compared to the non-entropy filtered prediction results, certainly, higher mean AUC scores were obtained when using the entropy cutoff of 0.3. As the overall mean entropy for all columns in a MSA is related to the sequence diversity of the MSA (sequences with higher diversity would generally produce a MSA with a higher overall mean

entropy), similar to the sequence diversity results in Chapter 2, higher entropies improve interaction predictions.

The best method for treating gaps in a MSA was evaluated to be the 21_AMINO_ACID method. When using this method, no removal of gapped columns or rows would be made; and as a result, more columns with higher entropies were likely to be included for the prediction calculations, which would, consequently, increase the overall mean entropy. As stated above, higher entropies would result in higher prediction performance, so it is understandable that 21_AMINO_ACID would outperform the other two gap handling methods examined. None of the three reduced alphabet groupings (based on stereochemical properties, volumes and amino acid residues similarity in natural protein sequences) were observed to reach performance comparable to the standard twenty amino acid alphabet grouping. Few groups (Pollock *et al.*, 1999; Bacardit *et al.*, 2009) have reported successful generation of optimal reduced alphabet groupings, but the processes were either quite complex or required extensive computational power. Furthermore, a previously reported method for removing background signals, APC, (Dunn *et al.*, 2008) was implemented, and its effectiveness was indeed observed throughout this chapter. Higher AUC scores were obtained for all APC predictions when compared with the predictions obtained without the APC step.

After examining the AUC scores for the mutual information approach and the two correlation-based approaches, MI-Partial correlation was determined to be the most effective method for predicting intramolecular protein interactions. It efficiently identified contacting positions with the highest accuracy. As an extension to the MI-Partial correlation approach, partial correlation level analysis was carried out by extracting the highest significant partial correlation combinations to determine clusters consisting of three contacting residues. However, this method tends to identify clusters consisting of two close positions and a slightly farther position. This is not surprising, since a third position is almost certainly going to be further away than two close ones, so the mean distance must go up. However, whether the mean distance should go up as much as it does has not been tested explicitly. In order to identify tightly packed clusters, the top 50 or 100 partial correlation level results should be utilized, as the accuracies for the top results were much higher than the overall predictions. Moreover,

as reported by Juan *et al.* (2008b), further comparisons implementing multiple partial correlation levels might be required to identify the optimal partial correlation level for the predictions. However, the computational power required to analyze large proteins could be significant.

Finally, although it was not tested here, it is also possible that intermolecular interactions could be predicted by applying the same approaches described here to two different proteins. Essentially, MSAs of two different proteins would be concatenated first and position pairs (belong to different proteins) with high correlation values would likely be part of the binding interface between the two proteins. However, as discussed, there are methodological challenges to generating such datasets. One would need to generate two MSAs from, ideally, orthologous protein pairs. Inevitably, not all species will possess the domain/protein in question, leading to the reduction in size of the MSAs. However, if sufficient numbers of sequences are present, this approach has promise for selecting candidate interacting residue positions, which could be used as constraints or additional evidence in tandem with other experimental or computational approaches. This strategy might be dependent on restricting the predictions to low sensitivity/high precision to ensure high likelihood of good quality prediction, but this seems possible (see Figure 4.9). For example, docking studies could benefit by restricting the search space to those consistent with the MI predicted contacts.

Further testing would be required for intermolecular prediction studies, for instance, using the Hakes+ dataset (Chapter 2). Since interacting proteins in Hakes+ are determined based on their 3D structures, the partial correlation values for physically contacting positions could be compared to the partial correlation values for non-contacting positions. Given that, in the intramolecular analysis shown in this chapter, contacting positions tend to have higher partial correlation values, it is therefore reasonable to suggest that positions that are part of the binding interface are also likely to have higher partial correlation values. As a preliminary step, this would be especially useful for identifying putative binding regions prior to carrying out experimental verification methods, such as site-directed mutagenesis.

5. Characteristics of Functional Binding Sites for GPCRs in Relation to PRINTS Motifs

5.1. Aim

The main aim of this study was to characterize the following G protein-coupled receptor (GPCR) interactions with PRINTS (Attwood *et al.*, 1994) motifs: ligand-binding, G protein-coupling, oligomerization and general protein-protein interaction binding sites. This study was carried out for GPCR families of adrenergic, chemokine, interleukin-8, dopamine, histamine, muscarinic and serotonin receptors; and the bovine rhodopsin structure (1F88) was utilized as the template for structural mapping. If a correlation was found between any of the functional binding sites analyzed and GPCR fingerprint motifs, PRINTS fingerprints could then be used as a predictive tool to provide valuable functional insights, which could eventually lead to GPCR-targeted drug development.

5.2. Introduction

G protein-coupled receptors are the largest family of membrane proteins and are responsible for the majority of transmembrane signal transduction. Many GPCRs have been found to contribute to various types of disorders and to link to multiple forms of cancer (Dorsam and Gutkind, 2007). For instance, chemokine receptor type 2 (CXCR2) was identified to contribute to pancreatic (Matsuo *et al.*, 2009), lung (Keane *et al.*, 2004), prostate (Reiland *et al.*, 1999), ovarian (Yang *et al.*, 2010) and melanoma (Singh *et al.*, 2009) cancers. Hence, it is of great interest for the pharmaceutical community to characterize GPCRs. Being popular drug targets, GPCRs account for more than 60% of marketed drugs (Janovick *et al.*, 2009). The general mechanism of the GPCR signal-transduction pathway (Figure 5.1) starts with ligand binding to cause a conformational change in the receptor, which then leads to the coupling and activation of G proteins. Upon activation, G protein subunits can then interact with the target proteins to regulate their activities.

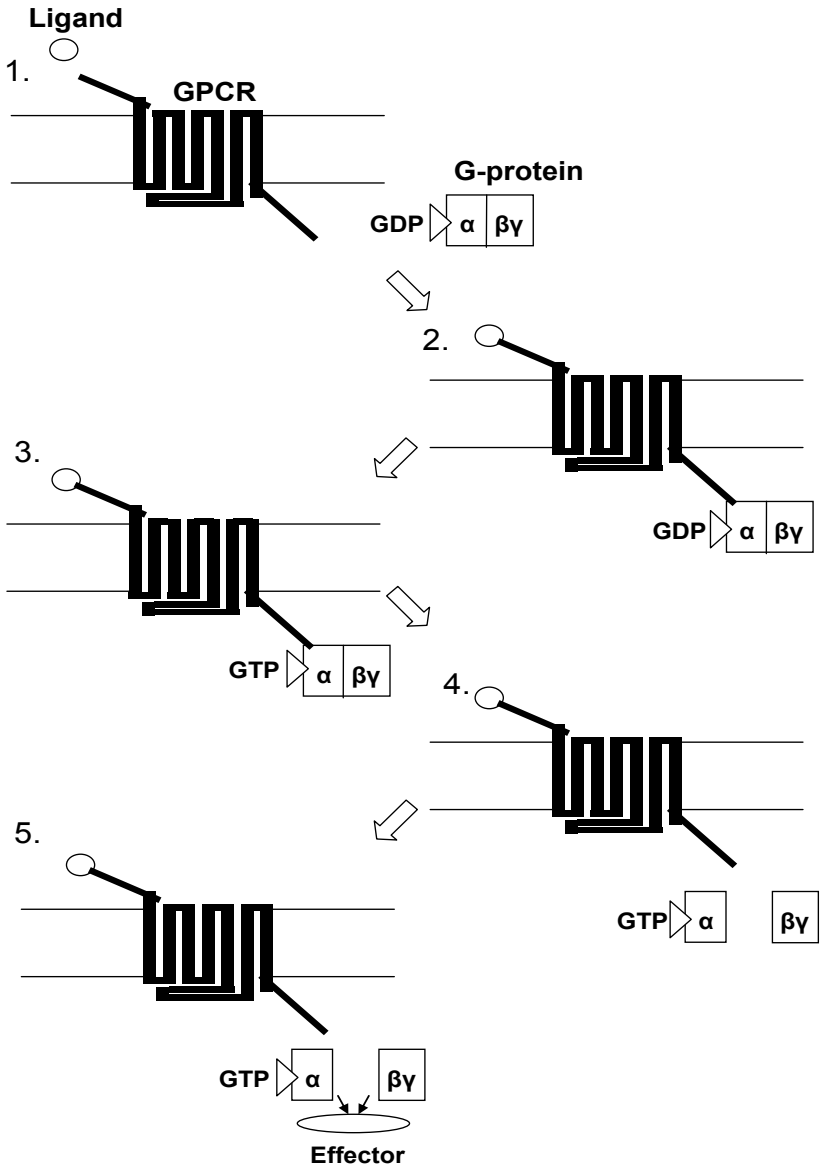


Figure 5. 1. The GPCR signaling mechanism. 1) At the resting state, the inactive GPCR and G protein are unbound to each other. 2) After binding to a ligand, the GPCR binds to a G-protein and 3) induces a conformation change in the G protein, which results in the transformation of guanosine diphosphate (GDP) to guanosine triphosphate (GTP). 4) Once the G protein is activated, the GTP-bound α subunit and βγ complex disassociate from the GPCR and each other. 5) Both the GTP-α and βγ complexes are then able to interact with the effector.

Owing to the tremendous difficulty of separating membrane proteins from the membranes to which they are attached and the restrictive conditions necessary to induce crystallization, it was not until the year 2000 that the first GPCR structure of bovine rhodopsin, 1F88 (Palczewski *et al.*, 2000), was available. Certainly, the identification of 1F88 is a remarkable aid in GPCR-based structural studies. More recently, in 2007, the first human GPCR structure, 2RH1 (Cherezov *et al.*, 2007), was solved for the beta-2 adrenergic receptor. Both 1F88 and 2RH1 contain the predicted general GPCR seven transmembrane (TM) alpha-helices, separated by three extracellular and three intracellular loops, with the N-terminus located in the extracellular region and C-terminus residing in the intracellular region (Figure 5.2), although many conflicting features have also become apparent from comparisons of the two structures. In the docking study of CCR5 carried out by Li and colleagues (Li *et al.*, 2009), the homology model, built based on 1F88, was found to be more comparable with experimentally determined results than the 2RH1 model. Attempts to map regions to 2RH1 that are highly specific to subfamilies other than the beta-2 adrenergic receptors not only could prove challenging, but may also introduce incorrect results. Nevertheless, 2RH1 should still provide important insights into how GPCRs function in humans. The 2RH1 (Figure 5.2) and other GPCR structures were solved as a chimera with other proteins or macromolecules and, therefore, are not suitable for a study of this nature. As a result, the 1F88 structure of bovine rhodopsin was chosen to depict the general 7 TM regions in this study.

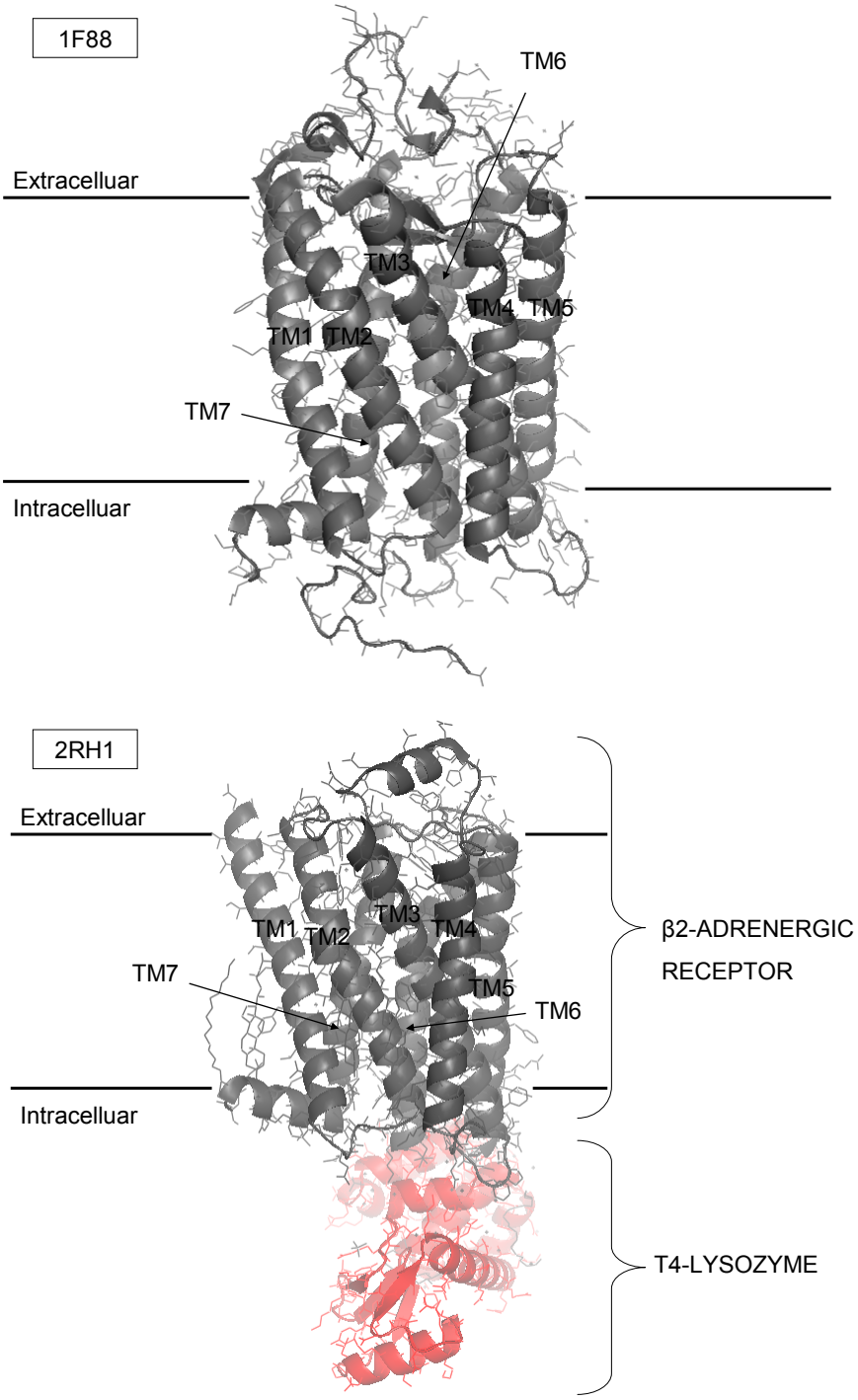


Figure 5.2. PDB structure 1F88 for bovine rhodopsin and 2RH1 for beta 2-adrenergic receptor/t4-lysozyme chimera.

It is widely understood that highly conserved regions in a protein are caused by functional or structural constraints. To identify such regions that also uniquely represent each protein family, a strategy utilizing fingerprints, (i.e. groups of short highly conserved motifs) was developed (Attwood and Findlay, 1993). More than 2000 fingerprints have been manually created in order to provide unique diagnostic signatures for a range of protein families; these can be obtained from the PRINTS database (Attwood, 2002). Essentially, the creation of PRINTS fingerprints starts by identifying appropriate orthologous sequences for a protein family from a wide range of species. Selecting divergent species is particularly important as conserved regions in closely related species are not sufficiently diverse to represent a protein family. Upon the generation of a well-aligned and family-representative multiple sequence alignment (MSA), motifs can be selected and must follow 5 criteria:

1. A motif cannot contain any gaps.
2. A motif must be between 10 to 30 residues long.
3. A motif can be conserved only within the family of interest in the appropriate hierarchical level (i.e. superfamily, family or subfamily).
4. Maximum of one residue overlap is allowed between motifs in the same fingerprint.
5. Minimum of two motifs but ideally more than three should be selected for reasonable diagnostic power.

Unlike many other protein family databases that only provide family-level information, PRINTS contains fingerprint signatures at superfamily-, family- and subfamily-levels. This enables more detailed and refined protein analyses. For each hierarchical level, motifs are selected to represent the protein family at that level and to ensure minimum overlap with higher level motifs. As shown in Figure 5.3, family-level motifs are selected in such a way that little overlap with superfamily-level motifs occurs. Similarly, subfamily-level motifs should have little overlap with family- and superfamily-level motifs.

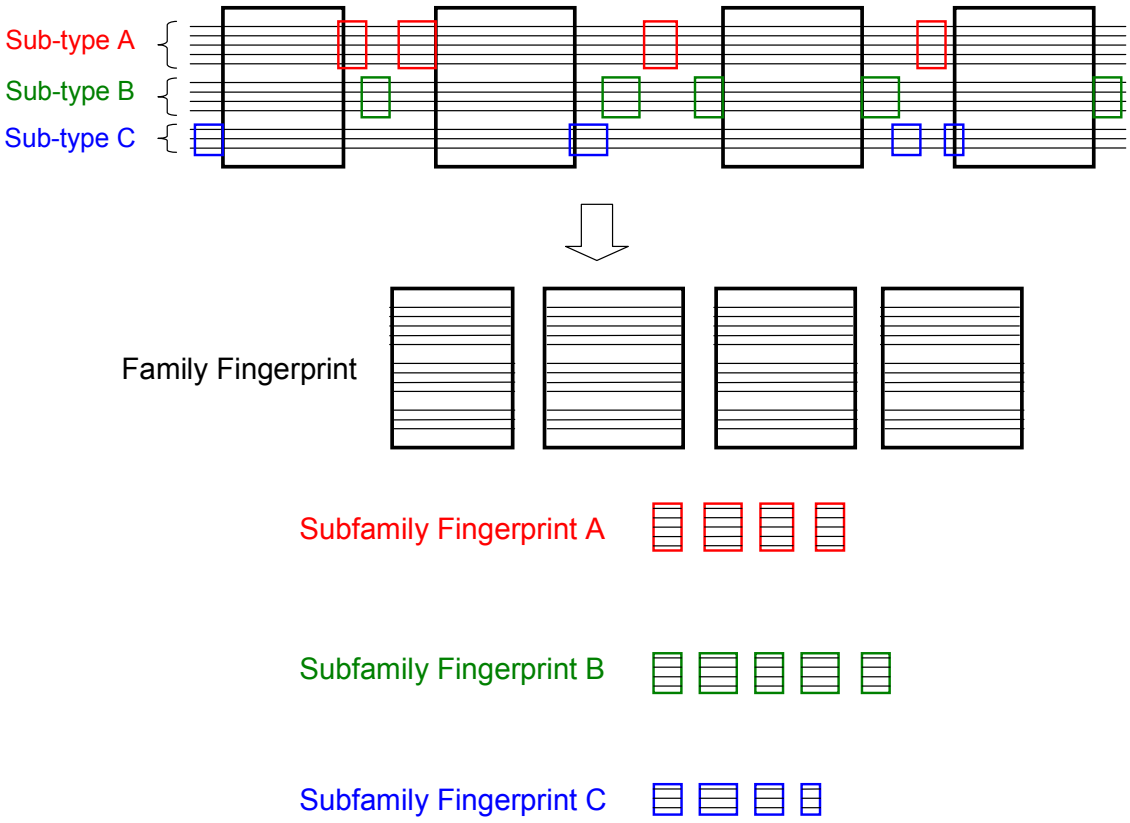


Figure 5.3. Determination of family fingerprint and subfamily fingerprints. Regions that are highly conserved for all sequences in the family are selected to represent the family-level fingerprint (black blocks). Subfamily-level fingerprints are determined from regions that are only conserved for each sub-type but not for other sub-type or family-level sequences. Three subfamily-level fingerprints A, B, and C are shown in red, green and blue blocks, respectively.

In a previous study (Gaulton, 2004), it was suggested there is a correlation between family-level motifs and experimentally verified ligand-binding sites, and, to a lesser degree, between subfamily-level motifs and G protein-coupling sites. It was pointed out by Gaulton that, as most of the GPCR families represented by the fingerprints in PRINTS bind to either the same or related ligands, it is therefore possible that the highly conserved motifs in these fingerprints are associated with ligand-binding sites. In addition, many subfamily-level motifs are located in regions (i.e. intracellular loops) of the receptors where ligands are unlikely to bind, so subfamily-level motifs were evaluated to identify a possible association with G protein-coupling sites. However, many motifs still could not be explained by these two types of binding mechanism. Since many GPCRs have been found to form homo- or hetero-oligomers (Breitwieser, 2004; Maggio *et al.*, 2005) and sometimes could also bind to proteins

other than the endogenous ligands or G-proteins (Smith *et al.*, 1999; Cheng *et al.*, 2000), it is possible that these two types of bindings could also be associated with the unexplained motifs.

Hence, to better understand important functional regions in class A rhodopsin-like GPCRs, the current study compares sites of interaction — such as those relating to ligand-binding, G protein-coupling, oligomerization and general PPIs — with PRINTS fingerprints for seven GPCRs (i.e. adrenergic, chemokine, interleukin-8, dopamine, histamine, muscarinic and serotonin receptors) at the family- and subfamily-levels. A phylogenetic tree depicting all GPCR receptors utilized for this study is shown in Figure 5.4. It should be noted that, although many GPCRs are clustered together within their own family (where a family is indicated by highlighting in a common colour), some subfamily-level GPCRs appear to be more closely related to other subfamily-level GPCRs from a different GPCR family. Nevertheless, as the aim of this study is to identify potential associations between different hierarchical level GPCR motifs and interaction sites, this should not influence the analysis results.

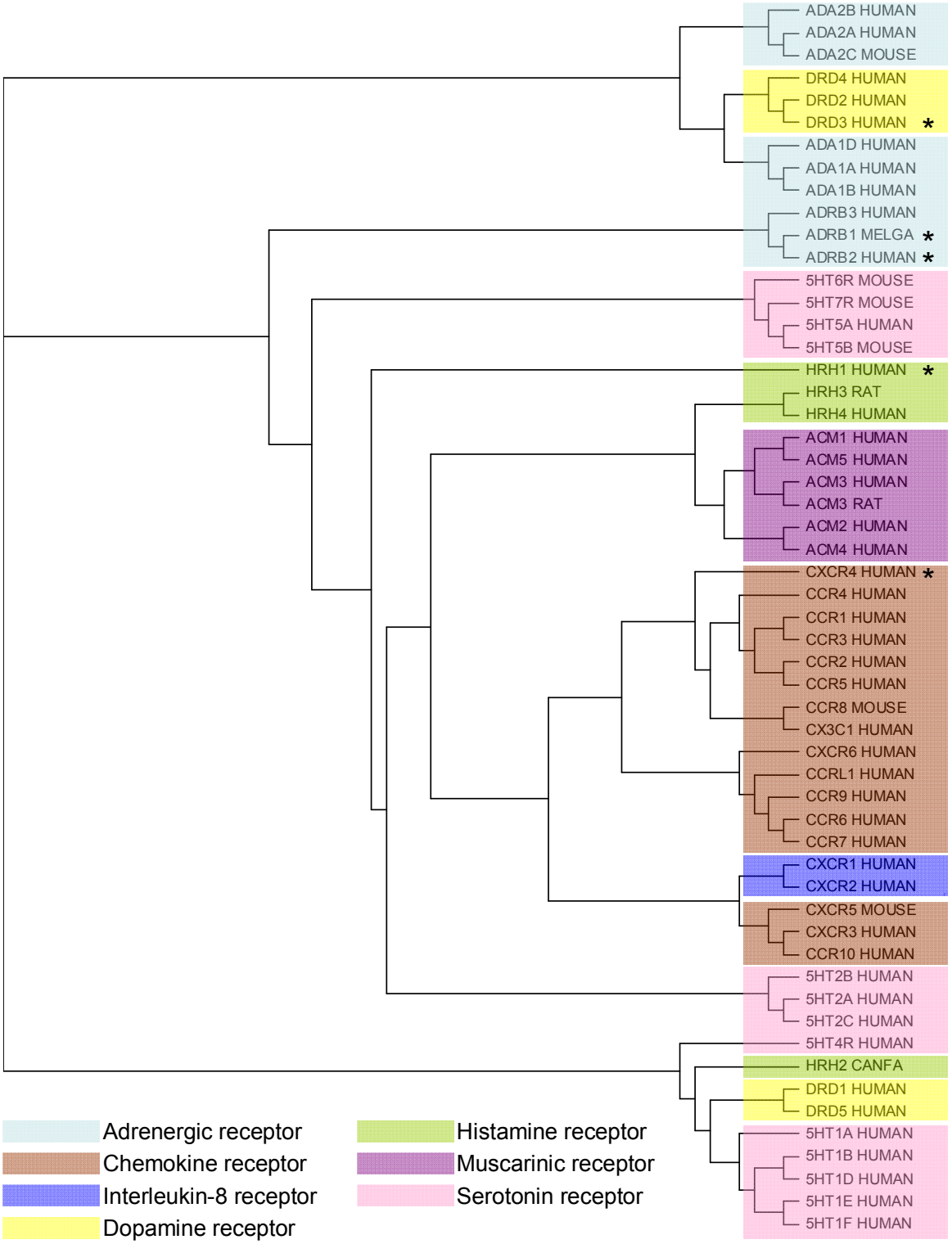


Figure 5.4. A phylogenetic tree of the GPCR families utilized for the PRINTS motif analysis. * Proteins with solved structures.

5.3. Methods

5.3.1. Data

In order to obtain a list of high-quality ligand-binding, G protein-coupling, oligomerization and PPI sites for the selected GPCR families, several resources were utilized, with the primary source being GPCRDB (http://www.gpcr.org/7tm/). Mutation data listed in this database were obtained through either MuteXt or tinyGRAP; furthermore, publications containing details of these mutations were also extracted. While the extraction of experimentally determined mutations from the literature was achieved in an automatic manner for the MuteXt data, the mutation data in tinyGRAP were extracted manually. In addition to using GPCRDB, PubMed (http://www.ncbi.nlm.nih.gov/pubmed) was searched with carefully selected combinations of keywords to obtain publications that were overlooked by GPCRDB. For the data used in this study, a total of 469 binding sites for 64 GPCR proteins were extracted manually from 290 publications obtained using the above databases. Publications for ligand binding, G protein-coupling, oligomerization and PPI sites are listed in Appendices 2, 3, 4 and 5, respectively.

All sequences for this study were obtained from UniProt (http://www.uniprot.org/) and aligned using MUSCLE (Edgar, 2004). Following that, the multiple sequence alignments were displayed in CINEMA (Pettifer *et al.*, 2004) and manual adjustments were made to ensure proper alignment. All family- and subfamily-level motifs for adrenergic, chemokine, interleukin-8, dopamine, histamine, muscarinic and serotonin receptors were obtained from the PRINTS database (http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/QuizPRINTSTX.php).

5.3.2. Residue Numbering Schemes

As different GPCRs were compared, it was important to use a universal numbering scheme for labelling residues from different receptors, to ensure compatibility of the data. Many residue numbering schemes (Ballesteros and Weinstein, 1995; Schwartz *et al.*, 1995; Baldwin *et al.*, 1997) have been proposed; however, the methods developed by Schwartz *et al.* and Baldwin *et al.* require helices with a fixed

number of residues, which is inappropriate for studying different GPCR families. Moreover, none of the proposed numbering schemes allow the comparison of loop and terminal regions. Hence, to account for large loop regions in some GPCR families, a modified Ballesteros and Weinstein scheme (Gaulton, 2004) was implemented.

The modified residue numbering scheme assigns residues based on the most conserved residue in each TM helix, in which the boundary is determined according to the bovine rhodopsin structure, 1F88. Respectively, these residues are N55, D83, R135, W161, P215, P267 and P303 for each TM helix. An index number X.50 was assigned to each of the above residues, where X represents the helix number. The rest of the residues within the boundaries of each TM domain were then labelled in relation to the X.50 indexed residues, while the non-TM residues utilize -n or +n to indicate the relative distance left or right to the closest helix. The assignment for loop residues can be rather arbitrary, but generally each loop region was divided in the middle in order to be able to allocate an appropriate helix number.

For instance, as shown in Figure 5.5, N55 was labelled as 1.50 and the residues preceding and subsequent to it were assigned 1.49 and 1.51. The last residue in the N-terminus was denoted as 1.28(-1) because it was the first residue left of the TM1 boundary (1.28). Since there were four residues in the loop region between TM1 and TM2, the two left residues were labelled according to the TM1 boundary, while the two right residues were numbered based on the TM2 boundary.

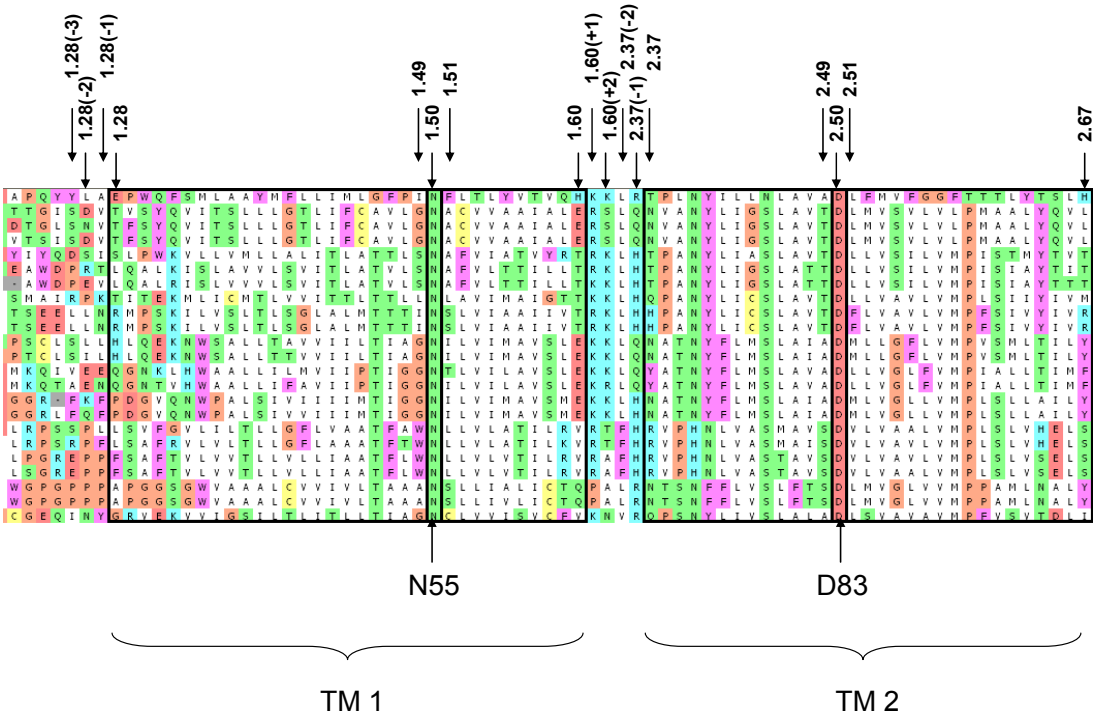


Figure 5.5. Modified Ballestros and Weinstein residue numbering scheme. A multiple sequence alignment for serotonin receptors is shown above, where the first and second transmembrane helices are boxed. Within each helix, the most conserved column is outlined and labelled 1.50 for helix 1 and 2.50 for helix 2. The non-helical regions are labelled based on the closest helix boundary.

5.3.3. Random Motif Experiment

It could be argued that, because motifs in the PRINTS database are selected manually, the process could be quite random and may therefore have no statistical basis. To examine this matter in further detail, an experiment was designed to mimic random selection of fingerprint motifs, and the resulting artificial motifs were then compared to the actual motifs obtained from PRINTS.

For each tested GPCR family and subfamily, multiple sequence alignments (MSAs) were constructed using sequences that represent each family or subfamily. Afterward, MSA regions containing short sequences were randomly selected. However, the selection process must meet certain criteria to ensure fair comparison between the artificial and original data. The artificial fingerprints must have the same number and length of motifs as the original fingerprints. However, the order of the motifs does not

necessarily need to be identical, since the order of the motifs is not a defining factor for the selection of PRINTS motifs. Furthermore, no gaps were allowed in any of the selected motifs since no gaps are allowed in real fingerprint motifs. One thousand artificial fingerprints were generated for each selected family and subfamily.

5.3.4. Surface Patch Analysis

Most of the motifs in PRINTS are not adjacent to each other in linear sequences. However, when a protein sequence is folded, the initially non-neighbouring residues can actually be seen to be in contact in a 3-dimensional structure. Hence, a surface patch analysis was carried out in order to examine whether PRINTS motifs within particular fingerprints were likely to form contiguous patches. If such contiguous patches were present on the surface, they might have some functional importance, for instance, as a potential binding interface for other molecules.

All surface residues were defined using the NACCESS program (Hubbard and Thornton, 1993) based on an algorithm developed by Lee and Richards (1971). Residues with a solvent accessibility score of ASA >= 15Å were classified as surface residues and displayed on structure 1F88. Afterward, the neighbouring surface patches were observed visually using PYMOL.

5.3.5. Statistical Analysis

As the number and length of motifs for each PRINTS fingerprint are maximized to ensure a unique identification of a specific GPCR family, it is possible that these motifs may occupy a substantial fraction of a sequence of the same family, making the probability of a functional binding site falling within or near a motif by chance quite high. It is therefore important to determine the probably of such an occurrence to ensure the results of the analysis carried out are indeed significant.

Following the approach reported by Gaulton (2004), the significance of finding a given number of binding residues that fall within or near a PRINTS motif can be calculated as follows:

$$p = \frac{m}{s}$$

where p is the probability of a residue occurring as part of a PRINTS motif. The total length of all motifs for a PRINT fingerprint is denoted as m, and the length of a protein sequence of the same family is denoted as s.

$$p(b) = \frac{n!}{b!(n-b)!} p^b (1-p)^{(n-b)}$$

where $p(b)$ is the probability that a given number, b, of binding residues fall within a PRINTS motif by chance. The total number of binding residues obtained for a GPCR family or subfamily is denoted as n.

Given the assumption that all binding residues are independent, the cumulative probability of observing b or more binding residues within a motif can be calculate as follows:

$$p(B \geq b) = \sum_{i=b}^{n} p(i)$$

5.4. Results and Discussion

5.4.1. Random Motif Experiment

PRINTS motifs form fingerprints to uniquely represent each GPCR superfamily, family and subfamily. These fingerprints are particularly powerful in characterizing GPCRs, or other types of protein derived from distantly related species, as the multiple-motif approach allows more flexible mappings than single-motif approaches. Since highly conserved regions are often associated with some functional roles, it would be intriguing to know whether these GPCR motifs actually were enriched in highly conserved regions, or whether they were merely the results of random motif selections.

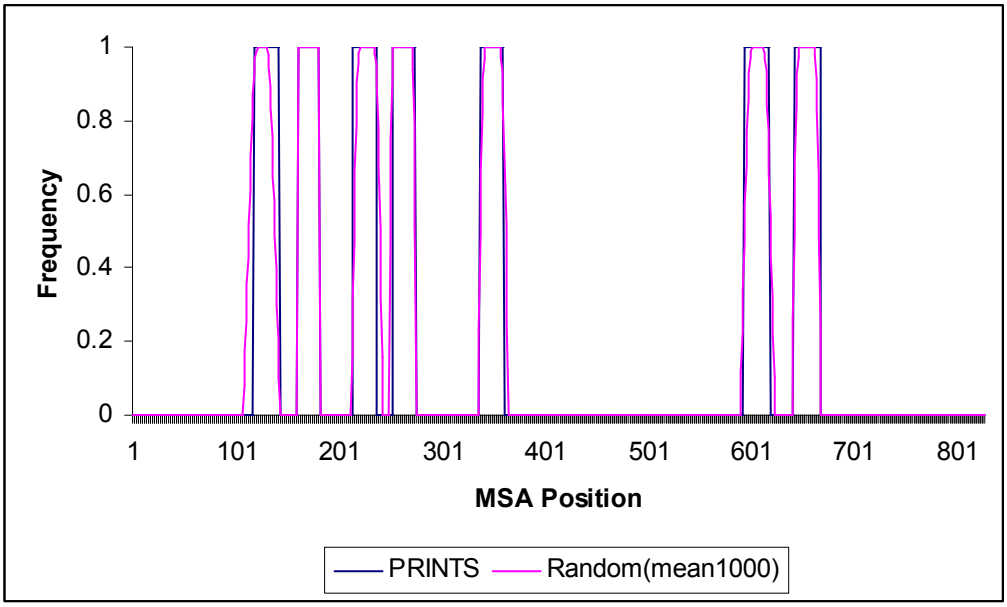
After randomly selecting motifs with similar criteria as the original motifs 1000 times, the probability of each position in the sequence as a candidate motif residue was evaluated. As shown in Figure 5.6A, the superfamily-level PRINTS motif profile nearly

superimposes the random motif profile. This is to be expected, as it is not possible to choose 7 motifs randomly from the superfamily alignment, because there are only 7 conserved regions for all GPCR sequences.

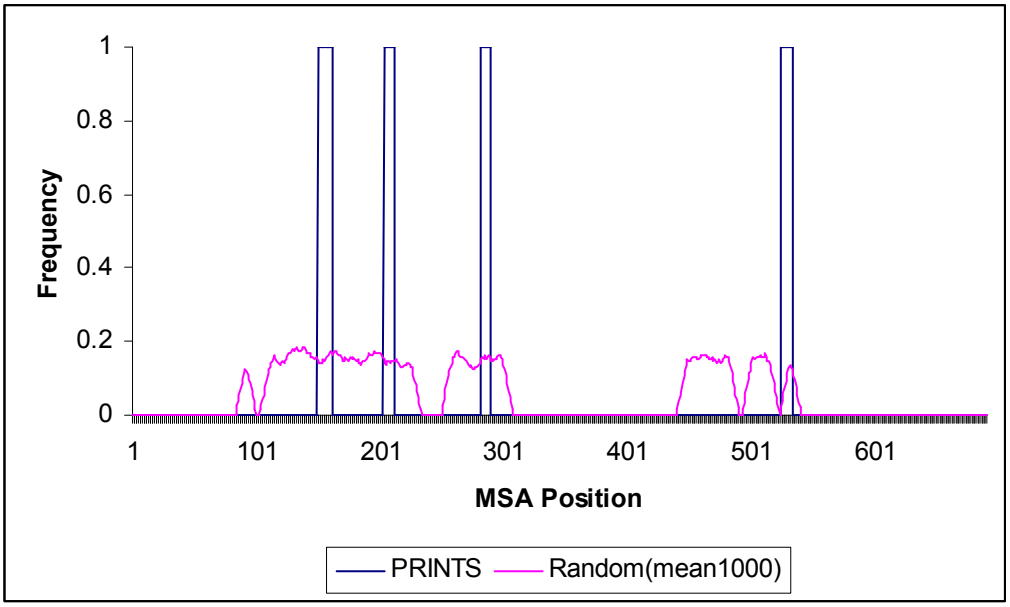
As family-level sequences are more conserved than those at the superfamily level, the regions that could be utilized to select multiple, short, non-gapped motifs increase drastically. In Figure 5.6B, six regions in the family-level adrenergic receptor MSA were shown to be suitable for motif selection, but only four short motifs were selected in PRINTS. Also, the average frequency for an adrenergic residue to be selected by chance as part of a family motif was 0.14, which is approximately seven-fold lower than the PRINTS frequency.

For subfamily-level motif selection, the whole MSA for the α1-A adrenergic receptor appeared to be a long stretch of candidate motifs, rather than the more focused six short sequences selected by PRINTS. Since sequences belonging to the same subfamilies are often highly conserved (even more so than at the family level), well-aligned MSAs without many gaps can often be produced. Hence, the selection of short conserved sequences can occur at almost any location, which reflects the single large motif-selecting region that resulted from the simulation. For adrenergic subfamily α1-A sequences, the average frequency for randomly selected residues to reside in the PRINTS-defined motif locations was determined to be 0.24. This random motif frequency is about four-fold lower than the PRINTS frequency.

A) Superfamily level: GPCR



B) Family level: adrenergic receptor



C) Subfamily level: adrenergic receptor subtype α1-A

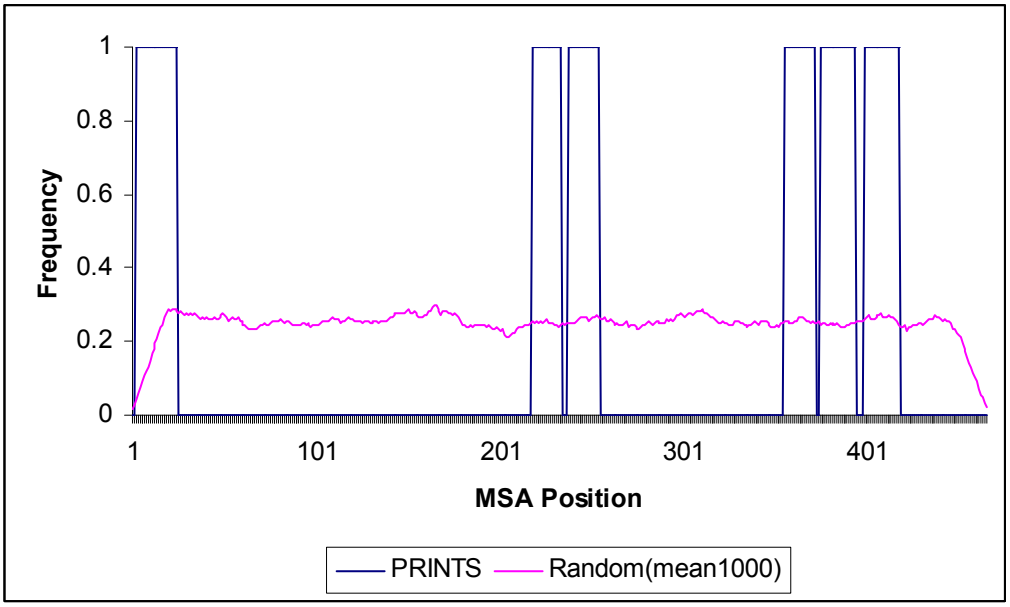


Figure 5.6. Frequency profiles for candidate motif residues. Blue lines represent the frequency of residues being selected as part of the original motifs in PRINTS. Red lines show the mean frequency of candidate motif residues selected randomly 1000 times. The profiles are shown for A) superfamily, B) family and C) subfamily.

Although different fingerprints were selected by using different MSAs (which consisted of orthologues from different species), trends similar to the adrenergic example shown here were observed for other receptors as well. The overall average background frequencies for all seven receptors evaluated in this study were calculated respectively to be 0.24 and 0.26 for the family- and subfamily-level sequences. It is evident that PRINTS motifs did not result from random generation, as the background frequencies are quite low for residues at these positions to be chosen by chance. Since these motifs were chosen by selecting conserved regions that are unique only to each receptor family or subfamily, it is likely that these signature motifs are of some functional importance.

5.4.2. Proximity between Motifs and Functional Binding Sites

As described in Section 5.2, GPCRs act as messengers that transmit signals from extracellular to intracellular spaces by first binding to ligands and, subsequently, by coupling to G proteins to activate the signal-transduction pathway. It is therefore crucial to know the locations where the receptors come in contact with the endogenous ligands

and G proteins. Recently, many GPCRs have been shown to form homo- or hetero-oligomers. Furthermore, protein-protein interactions other than ligand-binding and G protein-coupling also exist. Hence, the locations of these four functional binding sites and GPCR motifs obtained from PRINTS were compared to determine whether there is a correlation between them.

After acquiring all four types of binding residues from various literature resources, in order to generate comparable analyses, the residues were mapped to the rhodopsin structure, 1F88, using a modified Ballesteros and Weinstein numbering scheme to obtain universal positions. Subsequently, the proximity of all binding sites to PRINTS motifs was determined, and the results can be found in Appendices 6, 7, 8 and 9 for ligand-binding, G protein-coupling, oligomerization and PPIs, respectively.

Table 5.1 summarizes the proximity results for ligand-binding residues by showing the proportions of the binding sites that reside within or near family- or subfamily-level motifs. The proportions of the ligand-binding sites that lie within or near motifs vary among different receptors. While approximately half of the ligand-binding sites for interleukin-8 were found in both family- and subfamily-level motifs, only about 10 percent of the dopamine receptor ligand-binding sites were identified to be within family-level motifs, and no binding sites fell within the subfamily-level motifs.

As the PRINTS motif selection process would only select non-gapped regions that are conserved only within the family of interest at the correct level and with a minimum of 10 residues in length, it is important to also consider the neighbouring residues of all motifs to capture functionally important residues that miss a motif because of the selection requirements. For example, when a gap is inserted into a highly conserved region in a MSA and subsequently leads to the splitting of such region into two regions, with one being less than 10 residues long, the selection of motifs can then only be made at the larger region instead of the whole region. As a result, important binding sites that reside in the smaller region would not be included in this study, if the neighbouring residues are not compared. Predictably, the proportion of residues in close proximity to motifs increased for all receptors when including nearby residues. A larger increase in the number of residues was observed near motifs when the neighbouring five

residues were considered instead of only three. It is not surprising to have seen such results, as longer motifs would have higher chance of including more functional residues and other types of residues as well. However, most of the additional neighbouring binding sites were found to be within three residues away from a motif, particularly for chemokine subfamily-level motifs.

A study carried out by Gaulton (2004) has shown that there is a significant positive correlation between ligand-binding sites and family-level motifs; in addition, a negative association between the same sites and subfamily-level motifs was also indicated. However, such a trend was not evident here. Of all the ligand-binding regions acquired, 38% and 30% of the binding sites were found in or near (+/- 3 residues) family- and subfamily-level motifs, respectively. It should be noted that Gaulton's conclusion was drawn based on the analysis results of nineteen rhodopsin-like GPCR receptors instead of the seven families that were examined here. The author also indicated that ligand-binding sites for many receptor families, such as the adrenergic and bradykinin receptors, did not show a positive relationship with family-level motifs. Nevertheless, in this study, 57% of all ligand-binding sites do reside inside or within 3 residues away from either a family- or subfamily-level motif. Using the binomial probability model described in Section 5.3.5, the probability of obtaining such a result by chance was calculated to be 2.65×10^{-5}, which would suggest that the observed number of ligand-binding sites that fall within or near a PRINTS motif is significantly higher than would be predicted by random expectation. Given that more than half of the binding sites were found to correlate with PRINTS motifs, it is likely that PRINTS motifs have functional roles.

Receptor	In Motif		In or Near Motif (+/-3)		In or Near Motif (+/-5)	
	Family	Subfamily	Family	Subfamily	Family	Subfamily
Adrenergic	0.10	0.10	0.13	0.20	0.17	0.20
Chemokine	0.37	0.66	0.46	0.71	0.47	0.73
Interleukin-8	0.53	0.53	0.65	0.59	0.71	0.59
Dopamine	0.10	0.00	0.17	0.07	0.24	0.10
Muscarinic	0.42	0.00	0.58	0.00	0.71	0.00
Serotonin	0.21	0.21	0.30	0.24	0.36	0.33
Average	0.29	0.25	0.38	0.30	0.44	0.33

Table 5.1. Proportions of ligand-binding regions in proximity to family- and subfamily-level motifs.

Proximity of experimentally verified G protein-coupling sites to family- and subfamily-level motifs in the corresponding receptors was also determined, and the proportions of proximal sites are shown in Table 5.2. Of all the receptor families analyzed, four have considerably higher proportions of G protein-coupling sites that are in proximity to the subfamily-level motifs. While no G protein-coupling sites were found to be near dopamine family-level motifs, 45% of the sites were identified to be inside dopamine subfamily-level motifs. When taking into account three or five residues adjacent to subfamily-level motifs, the fraction of proximal G protein-coupling sites increased to 0.55 for the dopamine receptor. G protein-coupling regions for interleukin-8 and muscarinic receptors were detected to be closer to family-level motifs than subfamily-level motifs, whereas the other receptors showed the opposite trend. Furthermore, a large percentage of G protein-coupling sites was found outside of both the adrenergic family- and subfamily-level motifs, albeit still relatively close to motifs.

Receptor	In Motif		In or Near Motif (+/-3)		In or Near Motif (+/-5)	
	Family Level	Subfamily Level	Family Level	Subfamily Level	Family Level	Subfamily Level
Adrenergic	0.18	0.29	0.32	0.47	0.32	0.58
Chemokine	0.30	0.40	0.30	0.50	0.40	0.50
Interleukin-8	0.13	0.00	0.13	0.00	0.13	0.00
Dopamine	0.00	0.45	0.00	0.55	0.00	0.55
Muscarinic	0.25	0.22	0.33	0.22	0.33	0.22
Serotonin	0.00	0.07	0.00	0.07	0.00	0.07
Average	0.14	0.24	0.18	0.30	0.20	0.32

Table 5.2. Proportions of G protein-coupling regions in proximity to family- and subfamily-level motifs.

Overall, 45% of the G protein-coupling regions were found to be in proximal distance with either family- or subfamily-level motifs. This result is statistically significant, as the probability of such case to occur by chance was calculated to be 4.46×10^{-40}. Moreover, the analysis shows that 30% of the G protein-coupling sites are close to subfamily-level motifs and, to a much lower degree, 18% are near family-level motifs, based on the results for "in or near motif (+/-3)" in Table 5.2. The probability of obtaining the observed number of G protein-coupling sites residing in close proximity to subfamily-level motifs by chance was calculated as 1.13×10^{-55}. Indeed, this suggests that there is a positive association between subfamily-level motifs and G protein-coupling sites.

Although signal transduction is the high-level function of all GPCRs, oligomerization has also been shown recently to occur in many GPCRs. This could change the way GPCR research is being conducted, as alternative locations for ligand-binding and G protein-coupling could take place in multiple receptors. To investigate whether oligomerization has any association with PRINTS motifs, the analysis was extended to evaluate oligomerization sites. According to Table 5.3, oligomerization sites do not seem to have a strong correlation with either family- or subfamily-level motifs, as the proportion of oligomerization sites near either type of motifs are quite

similar. Essentially, 30% and 27% of oligomerization sites were found to be in, or within three residues away from, a family- or subfamily-level motif, respectively. The proportion of oligomerization sites varies greatly for different receptor families. In particular, all oligomerization sites acquired for the chemokine receptor were found near either family- or subfamily-level motifs: to be specific, 78% were identified in close proximity to family-level motifs and 44% to subfamily-level motifs. However, only one oligomerization site was found to be about five residues away from a family-level motif for the muscarinic receptor. Nevertheless, similar to the previous two comparisons, approximately half (47%) of the binding sites were identified to be associated with PRINTS motifs. The probability of finding this result by chance was calculated as 2.98×10^{-15}.

Receptor	In Motif		In or Near Motif (+/-3)		In or Near Motif (+/-5)	
	Family	Subfamily	Family	Subfamily	Family	Subfamily
Adrenergic	0.08	0.23	0.23	0.31	0.31	0.31
Chemokine	0.44	0.22	0.78	0.44	0.78	0.44
Dopamine	0.17	0.33	0.17	0.33	0.17	0.50
Muscarinic	0.00	0.00	0.00	0.00	0.50	0.00
Average	0.17	0.20	0.30	0.27	0.44	0.31

Table 5.3. Proportions of oligomerization regions in proximity to family- and subfamily-level motifs.

Consequently, in an attempt to elucidate some of the unmapped motifs from the previous comparison, the relationship between protein-protein interaction (PPI) sites and PRINTS motifs was examined. Evidently, PPI sites seem to have the strongest association with PRINTS motifs among all four types of binding sites, as 77% of PPI sites were determined to be near either a family- or subfamily-level motif, with the probability of finding such occurrence by chance as 2.15×10^{-39}. Particularly, all interleukin-8 PPI sites were found to be within family-level motifs, and all serotonin receptor PPI sites were identified to be within three residues from subfamily-level motifs (Table 5.4). Moreover, 14% more PPI sites were found to be in or within 3 residues of subfamily-level motifs when compared to family-level motifs. The probability of observing the obtained PPI residues near subfamily-level motifs by

chance was 5.85 x 10^{-50}. This would suggest a potential preference for PPI sites to be near subfamily-level motifs rather than family-level motifs.

Receptor	In Motif		In or Near Motif (+/-3)		In or Near Motif (+/-5)	
	Family	Subfamily	Family	Subfamily	Family	Subfamily
Adrenergic	0.05	0.56	0.15	0.67	0.18	0.72
Chemokine	0.24	0.47	0.29	0.53	0.35	0.53
Interleukin-8	1.00	0.00	1.00	0.00	1.00	0.20
Dopamine	0.46	0.13	0.54	0.13	0.54	0.17
Muscarinic	0.00	0.50	0.00	0.50	0.00	0.63
Serotonin	0.00	0.67	0.00	1.00	0.00	1.00
Average	0.29	0.39	0.33	0.47	0.35	0.54

Table 5.4. Proportions of protein-protein interaction regions in proximity to family- and subfamily-level motifs.

As shown above, many *p*-values are rather low, particularly for the G protein-coupling and PPI binding analyses. It should be noted that many functional binding sites that were obtained for this analysis are in blocks rather than single residues as many methods (i.e. chimera construction and mutation deletion) utilized to identify these functional sites cannot detect the precise binding locations and the sole residues that were involved in the binding. As a result, the true significance of the probability estimates through the use of binomial distribution could be over-estimated. This is especially evident in the G protein-coupling and PPI analyses where large binding regions were obtained and used for the probability estimates. In addition, residues in the same binding sites are not independent which could also lead to over-estimation of *p*-values. To demonstrate the aforementioned constraints for using binomial distribution for the analysis, simulations were carried out to randomly map a set of PRINTS motifs and functional binding sites to a protein sequence. After reiterating the simulation for 100000 times, the average number of overlapping residues between PRINTS motifs and functional binding sites were calculated. The empirical estimates of the probability suggest that the true *p*-values are less significant than the equivalent binomial estimates. For example, a *p*-value of 2.65 x 10^{-5} was derived from the application of binomial distribution to estimate the proportion of overlapping ligand binding sites and PRINTS

motifs. Thirty-five out of 49 residues were found to be in close proximity to either a family- or subfamily-level PRINTS motif in a sequence of 416 residues. After conducting the simulations, the probability of obtaining the same number of overlapping binding residues by chance was calculated to be 2.84×10^{-3} which is less significant than the binomial probability estimate, albeit still significant. This suggests that although the probability estimates derived using binomial distribution are over-estimates of the true significance, it could still be utilized as the initial statistic to filter out non-significant analysis.

5.4.3. Distribution of Motifs and Functional Binding Sites

Further analysis to investigate the association between distributions of motifs and all four types of functional binding regions was carried out. In order to increase the confidence of the analysis, all binding-site data were filtered based on the number of references and proximity to the nearest motifs. In most cases, binding sites obtained from at least two papers were retained for the subsequent analysis. However, this restriction was not applied to the interleukin-8 receptor analysis or to the PPI part of the muscarinic receptor analysis, because of the lack of binding-site data with more than one reference. Additionally, binding regions that are more than three residues away from a motif were not utilized, since an additional three residues adjacent to motifs were found to be sufficient for obtaining most of the important binding sites.

After mapping all family- and subfamily-level motifs of the seven evaluated receptor families onto a single schematic structure (see Figures 5.7 to 5.10), it is apparent that certain regions are 'motif-enriched'. Intracellular loop 1 and the N-terminal section of the C-terminus were found to be the predominant regions for family-level motifs (Figure 5.7A), whereas the highest numbers of subfamily-level motifs were identified in the C-terminal portion of N-terminus and the third intracellular loop (Figure 5.7B). The distributions of family- and subfamily-level motifs are indeed quite different. However, there seems to be a compensating selection that occurs, as the 'cold spots' (regions where only a few fingerprint motifs were found) for family-level motifs are 'hot spots' (regions where a large number of fingerprint motifs were found) for subfamily-level motifs, and vice versa. For instance, the third intracellular loop was

identified as the predominant region for subfamily-level motifs, but it was the region in which the least number of family-level motifs was found. Although one might argue that this could be due to how GPCR motifs were generated, it is not entirely true. Since the PRINTS fingerprint of each receptor family is independent, family motifs are expected to reside in different locations. While many family-level motif hot spots were found to be subfamily-level cold spots, some regions were found to have similar numbers of family- and subfamily-level motifs.

As shown in Figure 5.7, residues involved in ligand-binding were found in all TM, N-terminal and second extracellular loop regions. While some ligands bind to GPCRs to activate the signal-transduction pathway, some bind to inhibit the signaling cascade. To determine the specific regions of the binding interface for agonists (ligands for activation) and antagonists (ligands for inhibition), all ligand-binding sites were further annotated. Most of the regions for agonist binding were found in the extracellular regions, particularly the N-terminus and the second extracellular loop. In contrast, all antagonist binding sites were found only in TM regions. As for the regions that involved both agonist and antagonist binding, most were also found in TM regions, though not all near antagonist binding sites.

As for G protein-coupling regions, they were all found to reside in intracellular regions, especially the N-terminal portion of the C-terminus and the second and third intracellular loops (Figure 5.8). Not only is the third intracellular loop involved in G protein-coupling, it is also highly enriched in subfamily-level motifs. A region in the C-terminus that is also high in subfamily motifs (but to a lesser extent) was found to overlap G protein-coupling sites too. It appears that G protein-coupling sites are often found in specific intracellular regions, where subfamily-level motifs are frequently located. Furthermore, the C-terminal portion of G protein-coupling sites superimposes a family-level motif hotspot.

As shown in Figure 5.9, oligomerization sites reside in TM domains 1, 4 and 6, which are motif-poor regions. Since most of the TM regions were selected to represent the GPCR superfamily, it is less likely these regions were utilized to generate family- and subfamily-level motifs. One family- and one subfamily-level motif hotspot were found flanking TM1, which were also shown to be involved in oligomerization. While

the family-level motif hotspot is located inside the membrane, the subfamily-level hotspot resides on the outside of the membrane.

Since oligomerization in GPCRs is a relatively new concept, minimal data are currently available in comparison to the more extensively studied ligand-binding and G protein-coupling. A common technique to determine oligomerization sites is the use of chimera, which often does not identify specific binding residues. Instead, residue segments from putative functional regions are replaced with residues from other proteins. These regions are deemed to be the real functional sites if the associated function is abolished after the segment swapping. Owing to both the lack of available oligomerization data and unspecific locations of the binding interface, the association between motifs and oligomerization sites should be made with caution.

The results of the analysis, shown in Figure 5.10, reveal PPI and G protein-coupling regions overlapping in the third intracellular and the C-terminal regions. However, unlike G protein-coupling sites, PPIs do not occur in the second intracellular loop. Reflecting the proximity results from Table 5.4, most PPI sites were found near subfamily-level motifs rather than family-level motifs, and overlap all intracellular subfamily-level motif hotspots. However, the PPI region found near TM 7 in the C-terminus also overlaps a family-level motif hotspot.

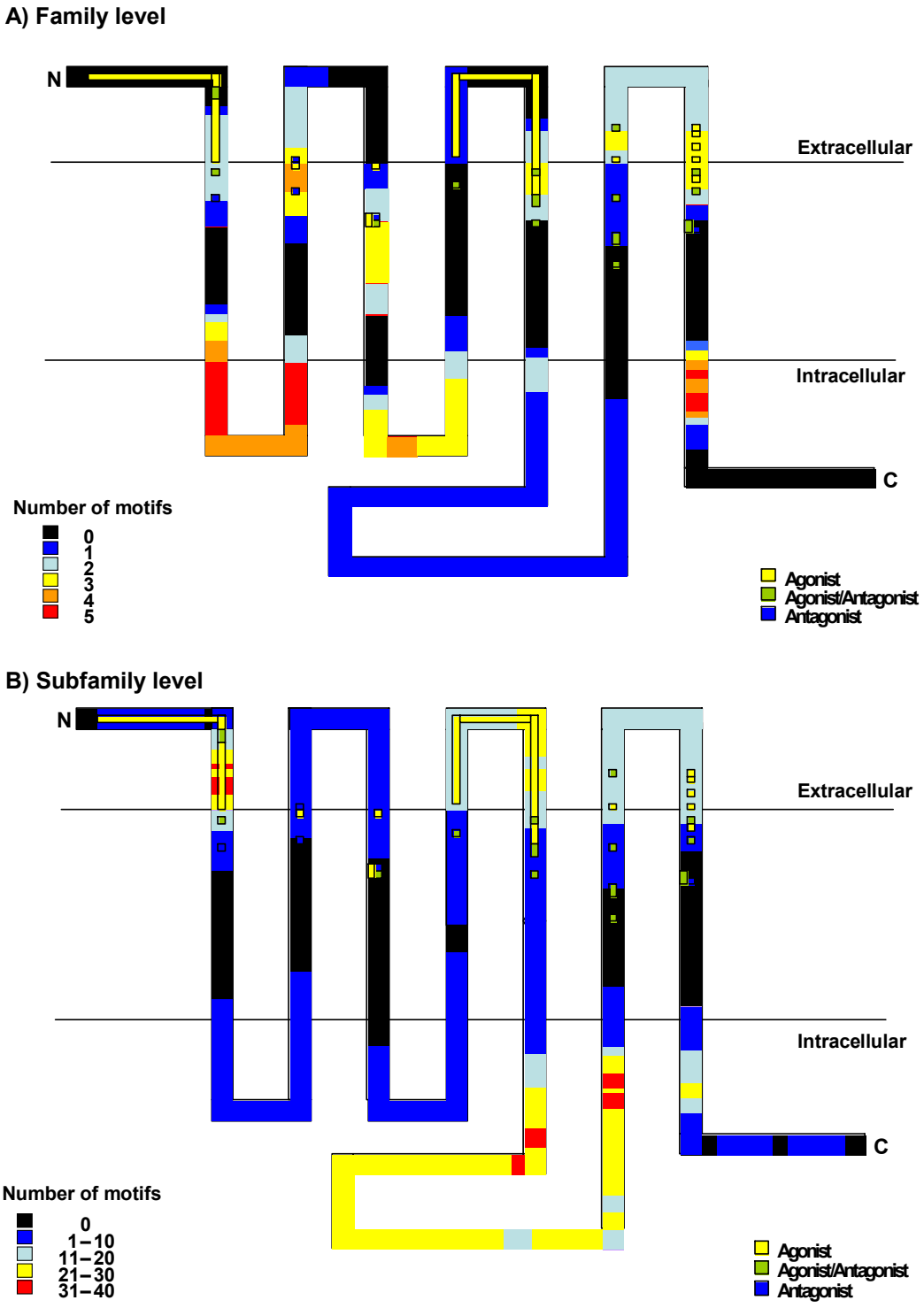


Figure 5.7. Locations of residues or regions known to be involved in ligand-binding in relation to A) family and B) subfamily-level fingerprints. Regions of fingerprints are coloured based on the number of fingerprint motifs in each region. Ligand-binding regions are represented by bars and coloured according to whether they are for agonist, antagonist, or both.

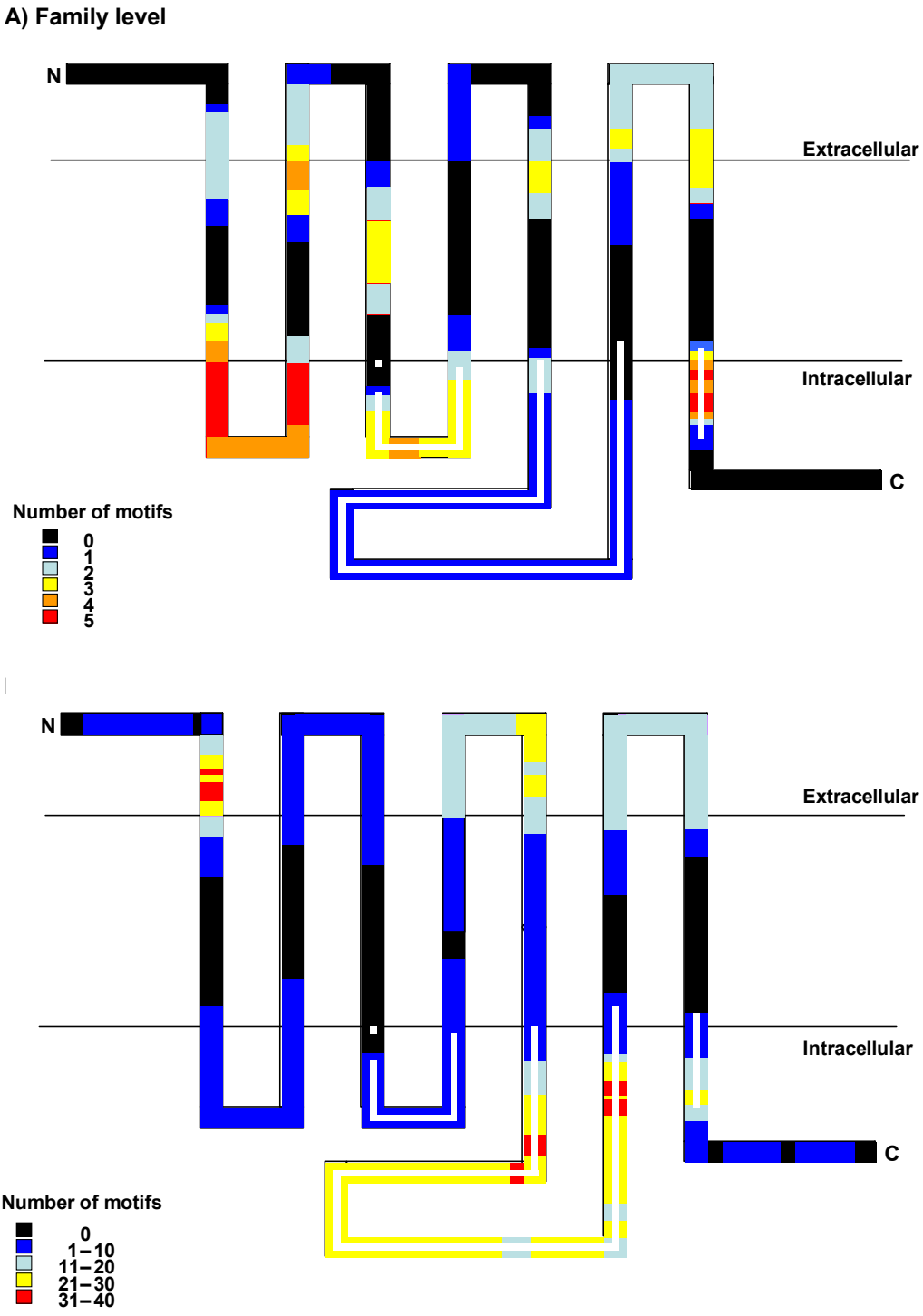


Figure 5.8. Locations of residues or regions known to be involved in G protein-coupling in relation to A) family and B) subfamily-level fingerprints. Regions of fingerprints are coloured based on the number of fingerprint motifs in each region. G protein-coupling regions are represented by white bars.

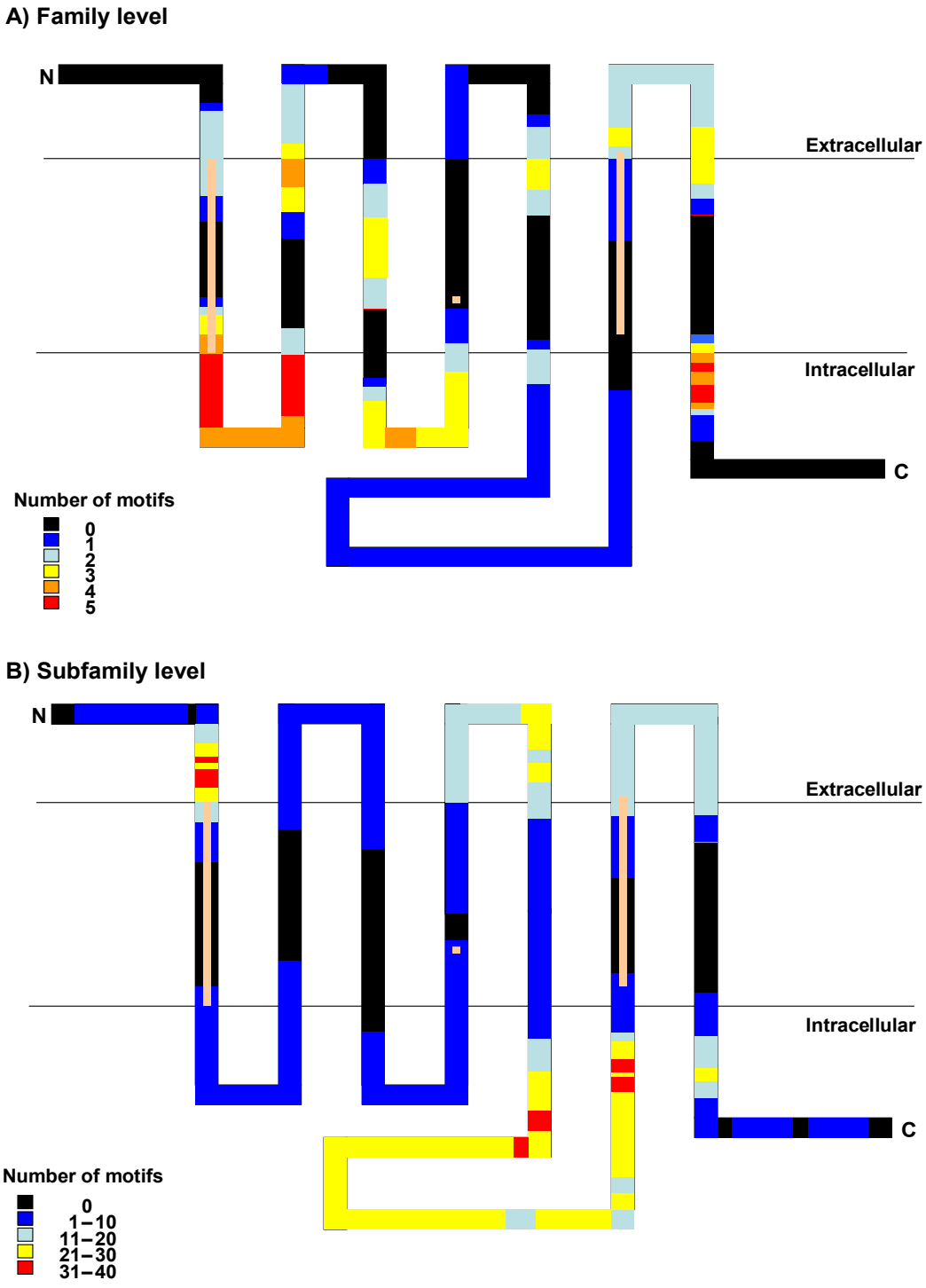


Figure 5.9. Locations of residues or regions known to be involved in oligomerization in relation to A) family and B) subfamily-level fingerprints. Regions of fingerprints are coloured based on the number of fingerprint motifs in each region. Oligomerization regions are represented by pink bars.

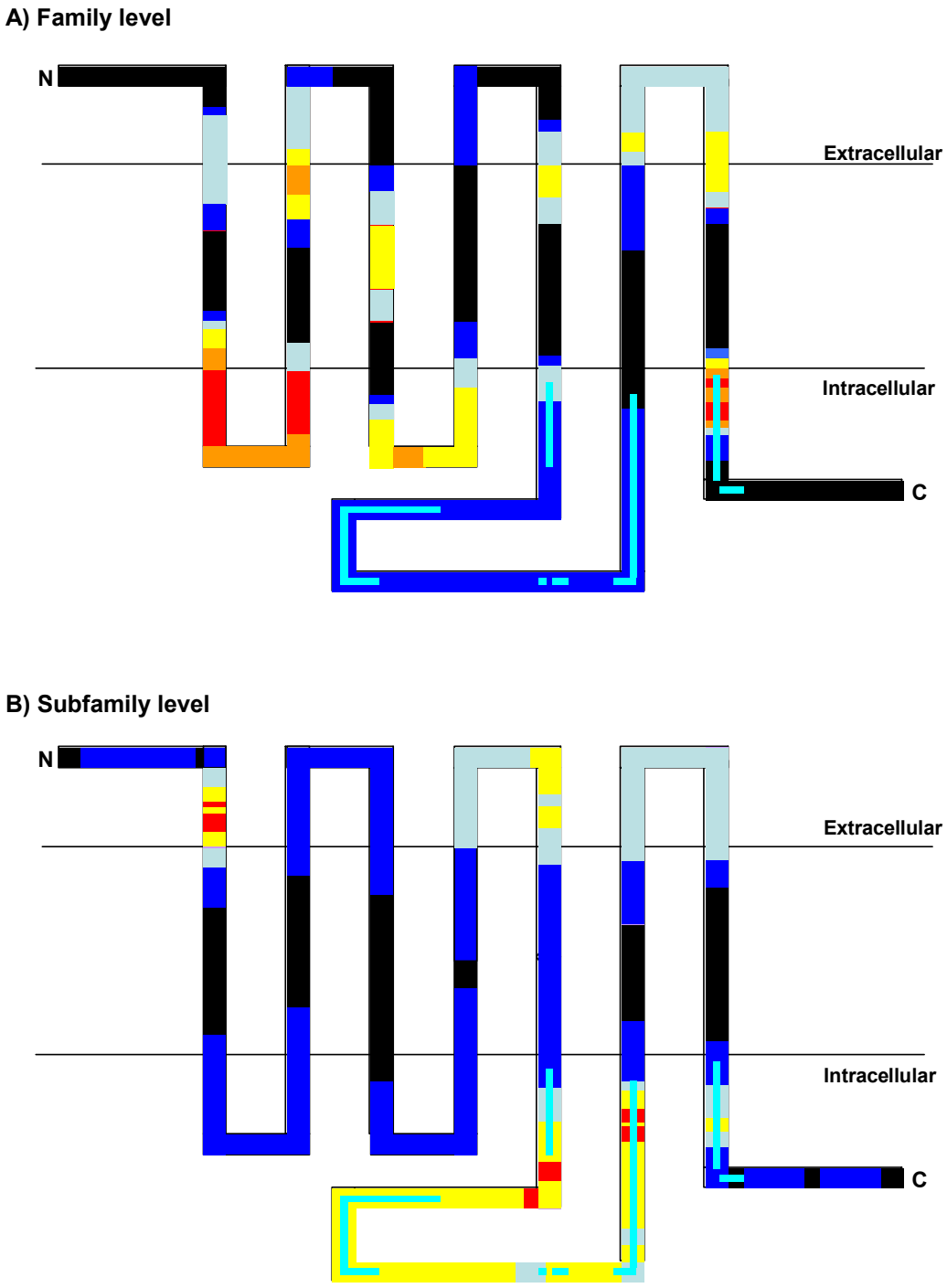


Figure 5.10. Locations of residues or regions known to be involved in protein-protein interaction in relation to A) family and B) subfamily-level fingerprints. Regions of fingerprints are coloured based on the number of fingerprint motifs in each region. Protein-protein interaction regions are represented by cyan bars.

5.4.4. Surface Patch Analysis

Many motifs were shown in the previous section to be associated with functional binding sites. However, motifs that reside in certain regions, such as intracellular loop 1, were not found to be involved in any functions analyzed. Motifs in these regions could still be functional but just not for the four types of interactions that were evaluated in this study. Alternatively, some of these motifs may reside in regions that are buried deep inside receptors and, therefore, would not be capable of interacting with other proteins or molecules. As surface residues are more likely to be involved in binding interfaces to other molecules, a surface patch analysis was carried out to examine the association between functional binding sites and motifs that reside on the surface.

Solvent accessibility for all family- and subfamily-level motifs was determined, and the results show that all family-level motifs are accessible on the surface (Figure 5.11). This excludes the histamine receptor, as no family-level motifs are available for this receptor. However, the locations of subfamily-level motifs vary for different receptors. All subfamily-level motifs for interleukin-8 are on the surface, compared with only 20% of muscarinic subfamily-level motifs are on the surface. Overall, about 65% of all subfamily-level motifs were found on the surface.

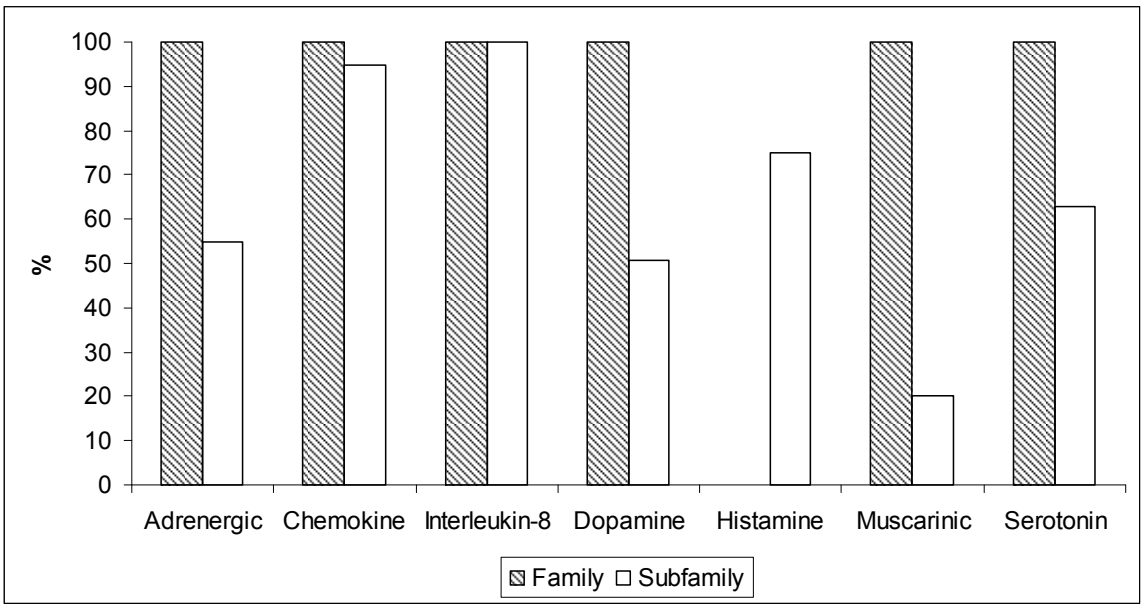


Figure 5.11. Comparison of the percentage of family- and subfamily-level motifs that reside on the surface of each receptor. Subfamily-level results are average values calculated based on all subfamily-level motifs in each receptor family. Family-level motifs for histamine receptor are not available in the PRINTS database.

Most of the PRINTS motifs appear non-contiguous on a linear sequence, and no associations among them can easily be depicted. However, when they are arranged into three-dimensional structures, it is quite evident that they form contiguous patches. Illustrated as an example in Figure 5.12, there are seven motifs in the serotonin 1A fingerprint but only two patches were observed on the surface. Motifs 1, 2 and 3 clustered into the first patch, and the second patch consisted of motifs 4 and 7. Two motifs, 5 and 6, were buried inside the structure, hence not shown on the surface. Although motifs 1 and 2 are adjacent to each other on the sequence, motif 3 is actually 153 residues away from motif 1 and 136 residues away from motif 2. Moreover, motifs 4 and 7 are 74 residues apart. To suggest a correlation in these motifs would have been rather difficult if they were not mapped to a three-dimensional structure. Additionally, both serotonin 1A surface patches were found to reside outside of the membrane. As they are not obstructed by the membrane, the probability of them interacting with other molecules is likely to be higher.

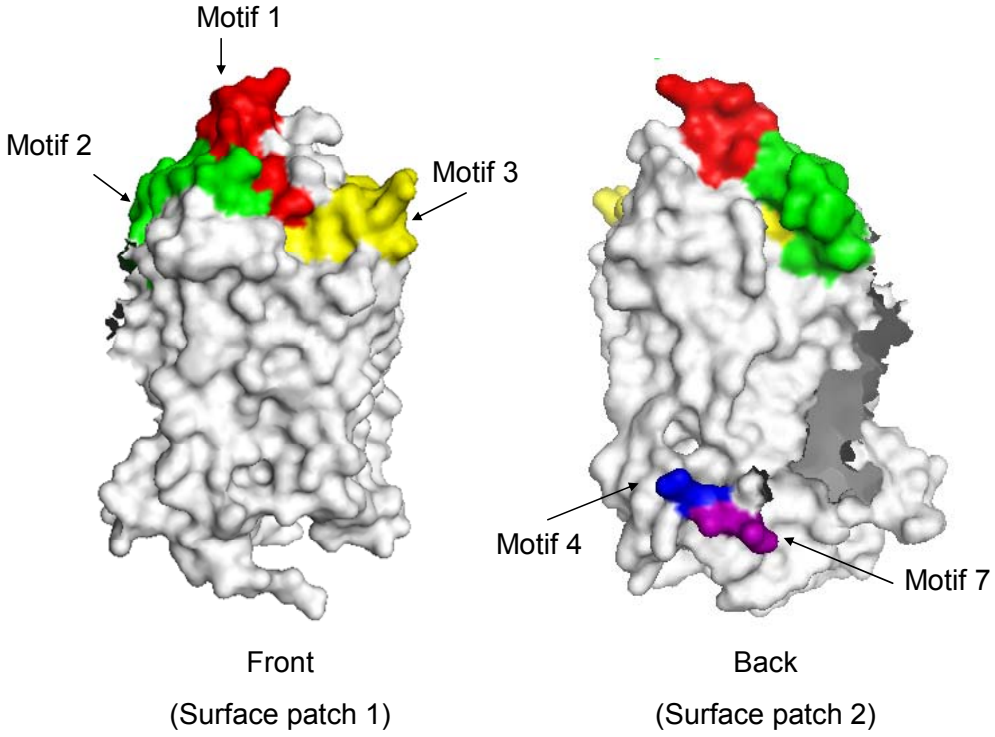


Figure 5.12. Visualization of the surface patches for serotonin 1A motifs. The front and back views of the structure are shown to illustrate two surface patches formed by motifs 1, 2, 3, 4 and 7 for the serotonin subfamily 1A receptor. The surface portion of each motif is shown in a different colour.

Surface clustering results similar to the example shown in Figure 5.12 were observed for all seven receptors examined. Most of the subfamily-level motifs clustered into two patches (Table 5.5). In addition to the clustering of subfamily-level motifs, family-level motifs form approximately four surface patches. The higher number of average clusters for family-level fingerprints would suggest that family-level motifs are probably more spread out in the sequences, whereas subfamily-level motifs reside closer together, albeit not necessarily adjacent to each other.

	Family	Subfamily (Average)
Adrenergic	4	2.00
Chemokine	2	2.27
Interleukin-8	4	2.00
Dopamine	4	2.20
Histamine	NA	2.25
Muscarinic	4	1.40
Serotonin	3	2.15

Table 5.5. Number of surface motif clusters for seven GPCR receptor families.

To further examine a possible connection to functional roles in motifs residing on the surface, the proximity of surface motif patches and all four types of binding sites were determined. In comparison to the earlier analysis that used all data (surface and non-surface), a significant increase in association between motifs and functional binding sites was observed. Eighty percent of the binding sites on the surface were found to overlap or lie within a family- or subfamily-level surface motif patch. After following the statistical analysis described in Section 5.3.5, and using only surface residues, the probability of finding 80% of the binding sites to be near a surface motif patch by chance was calculated as 3.71×10^{-5}. Such result suggests that the association between the binding sites and surface motifs is unlikely to happen by chance. In particular, 98% of all surface binding sites for the chemokine receptor were found to be associated with motifs of the same receptor on the surface. Although the earlier analysis found that G protein-coupling and protein-protein interaction sites are generally in close proximity to subfamily-level motifs, rather than family-level motifs, such association is no longer visible when restricting the analysis to surface residues only. Rather, all binding sites seem to overlap family-level motifs. This could be caused by the fact that a much higher percentage of subfamily-level motifs are buried.

Generally, ligand-binding and oligomerization regions were found to be related to family-level motifs, with the correlation between ligand-binding sites and family-level motifs more prominent and oligomerization association less evident. Such

relationships were also observed in the surface patch analysis. Considering only surface residues and binding sites falling within or adjacent to a family- or subfamily-level motif, 61% of the ligand-binding sites were determined to be correlated with family-level motifs. Once again, the result is significant, with a random expectation probability of 9.63×10^{-5}. In comparison to the previous analysis, this result shows a stronger association between ligand-binding sites and family-level motifs. Approximately twice as many ligand-binding sites were found to be near family-level motifs than subfamily-level motifs on the surface. However, according to Table 5.1, only about 8% more of ligand-binding sites (surface and non-surface) were found to be in close proximity to family-level motifs when compared to subfamily-level motifs. For the oligomerization regions, 30% were found to be associated with family-level motifs, while 27% were identified to be in or adjacent to subfamily-level motifs. Restricting the analysis to only surface residues for the oligomerization, binding regions showed no difference when compared with the results determined using all residues. It should be noted that by using binominal distribution to estimate the random expectation probability, the significance estimations for the surface patch analysis suffer the same constraints as the proximity analysis in 5.4.2. Nevertheless, the true probabilities should still be significant, albeit lower.

Figure 5.13 shows an example of all four binding types analyzed in relation to muscarinic M3 motifs. The positive association between ligand-binding sites and family-level motifs is quite evident, as all four cyan patches (ligand-binding sites) overlap red patches (family-level motifs). Furthermore, the marginally larger percentage of family-level motif association with oligomerization binding regions is also reflected, as the small yellow patch (oligomerization sites) overlaps a large family-motif patch in the second extracellular loop region. As stated in earlier analyses, G protein-coupling and PPI sites are likely to associate with subfamily-level motifs. However, such association is not observed in Figure 5.13, as neither the G protein-coupling patch (brown) nor the PPI patch (purple) overlap the subfamily-level motif patch (green). Interestingly, the PPI patch does overlap a family-level motif patch. Nevertheless, regardless of the type of motifs (family- or subfamily-level), most of the patches for functional binding sites overlap a motif patch, supporting the early finding that PRINTS motifs are likely to be associated with functional binding sites.

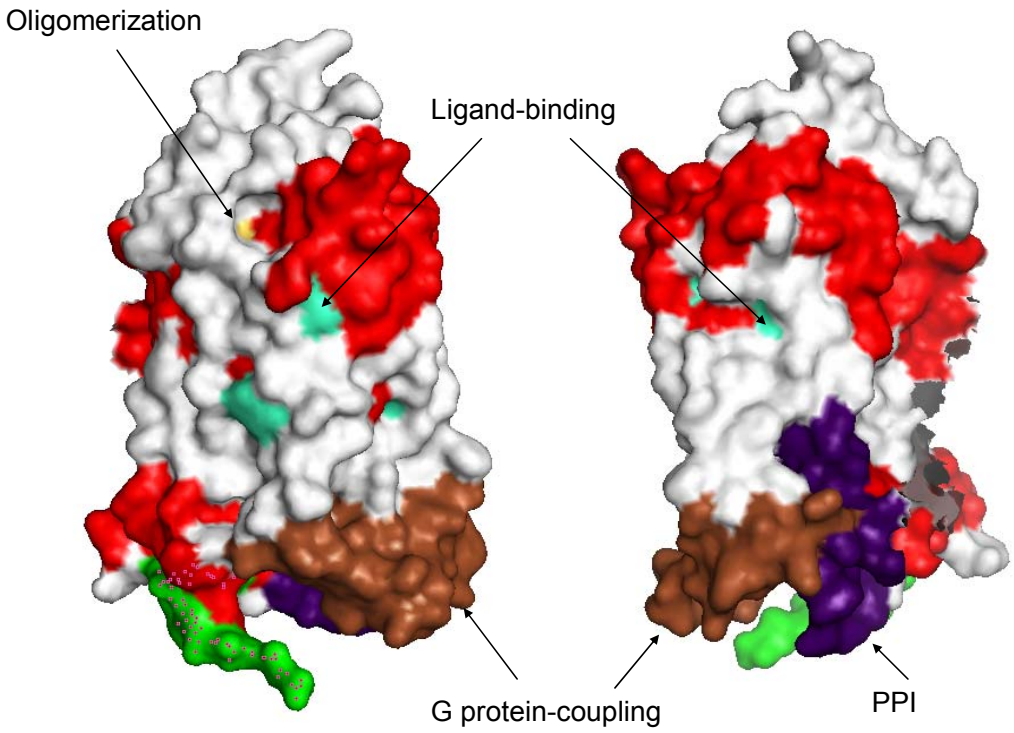


Figure 5.13. Structural representation for the muscarinic receptor motifs in relation to functional binding sites. Surface patches associated with ligand-binding, G protein-coupling, oligomerization and protein-protein interaction are shown, respectively, in cyan, brown, yellow and purple. Family- and M3 subfamily-level motifs are represented in red and green patches, respectively.

5.5. Summary

Comparisons between known functional binding sites and PRINTS fingerprints were carried out for seven G protein-coupled receptors to evaluate the possibility of utilizing fingerprints for determining important functional regions in GPCRs.

Prior to comparing functional sites and GPCR fingerprint motifs, it was important to ensure that these motifs were not the results of random motif selections caused by the manual nature of the motif selection process. A background distribution profile for each of the fingerprints analyzed was determined by randomly selecting the same number of motifs in a multiple sequence alignment representative of a specific GPCR family or subfamily; these motifs were of the same length as the original motifs. This random selection of motifs for a fingerprint was repeated 1000 times for all

receptors analyzed at the family- and subfamily-levels. Figure 5.6 shows that when comparing PRINTS fingerprints to the artificial fingerprints, a seven-fold higher probability of a residue being selected as part of a PRINTS motif was observed for the family-level adrenergic receptor motifs. Similarly, but to a lesser extent, a four-fold probability increase was observed for a subfamily-level adrenergic receptor motif residue to be selected as part of a PRINTS motif.

Subsequently, the associations between the four types of interaction and PRINTS fingerprints were analyzed. The proximity of all literature-acquired binding sites to PRINTS motifs was determined. Approximately 50% of the ligand-binding, G protein-coupling and oligomerization sites were found to be in close proximity to a motif, and 77% of the PPI sites were observed to be near a motif. Such a high proportion of binding sites found near motifs essentially suggests likely functional roles for these motifs. Furthermore, more detailed comparisons were carried out to determine whether there was an association between these binding sites and family- or subfamily-level motifs. The results show that higher proportions of ligand-binding and oligomerization sites are near family-level motifs, albeit that the preference is not significant. However, owing to the marginally higher proportion (3%) of oligomerization sites found near family-level motifs than near subfamily-level motifs, it is not possible to infer a preference between these two types of motifs for the oligomerization sites. In contrast, stronger associations between functional binding sites and PRINTS motifs were found for the G protein-coupling and PPI sites.

To deduce a more universal relationship between PRINTS motifs and functional binding sites, distributions of PRINTS motifs were determined and the locations of functional binding sites were related to these motifs. Hotspots in which family and subfamily-level motifs are frequently located were determined. Both G protein-coupling and PPI sites were found to overlap all intracellular subfamily-level motif hotspots. In addition, PPI sites also seem to overlap some of G protein-coupling regions. However, no clear association for ligand-binding and oligomerization sites could be identified from this analysis. Nonetheless, an interesting observation was obtained for ligand-binding sites that are involved in activation (agonist) or inhibition (antagonist) of the signal transduction pathway. Essentially, most ligand-binding sites involving agonists were found in the N-terminus and extracelluar loop 2, while the antagonist-binding sites

were identified to be in TM regions. In addition, ligand-binding sites that allow for both the agonist and antagonist binding were also identified, and these sites were found to be located in TM regions.

Some residues appear to be non-contiguous in a linear sequence yet actually cluster together when folded into a 3-dimensional structure. Motifs that form patches, especially the ones on the surface, could signify possible functional importance. Also, binding sites that were shown not to be in close proximity to a motif on the linear sequence could actually be in contact with a motif when shown in a structure. Hence, it was essential to carry out a surface patch analysis for all motifs and binding sites. A considerably high fraction of binding sites were found to be overlapping motif patches on the surface, confirming the suggestion of surface motifs being associated with important functional sites. In particular, considerably more ligand-binding sites were found to be near family-level motifs rather than subfamily-level motifs. Compared to the results from the analysis utilizing all residues (surface and non-surface), this association is more prominent, as a large difference in the proportion of ligand-binding sites found overlapping family- and subfamily-level motifs was observed. However, because a large fraction of subfamily-level motifs were not on the surface, no relationship between motifs and functional sites (G protein-coupling and PPI) thought to be associated with subfamily-level motifs could be derived. As for oligomerization regions, it is apparent that neither family- nor subfamily-level motifs have a strong association with this function. Common locations for oligomerization sites were found to be in TM regions, which are family- and subfamily-level motif-poor regions. Thus, it is unlikely that a meaningful inference could be made utilizing data found in these regions. Additionally, since considerably less oligomerization data is available when compared to other types of binding sites, it is possible that there is simply not enough data for the analysis to reveal a clear relationship between the acquired binding sites and motifs.

Taken as a whole, using motifs to study functional binding sites in GPCRs could definitely provide useful insights in ligand-binding, G protein-coupling and PPI. Also, once more oligomerization data become available, a more meaningful association between this function role and PRINTS motifs can be made. However, certain analysis results for different receptors could vary quite significantly. It is therefore important to

utilize PRINTS motifs in conjunction with other methods, such as site-directed mutagenesis, to ensure the accuracy of the results and to facilitate the identification of putative functional binding sites.

6. Conclusions

Proteins are vital parts of organisms because they account for more than half of the dry weight of most cells, and participate in nearly all processes within cells. However, in order to perform biological functions, proteins must interact with other proteins or macromolecules. To fully understand basic cellular organization and function, it is therefore essential to detect and predict protein-protein interactions (PPIs). Due to the explosion of sequencing projects over the past ten years, a vast number of protein sequences are now available. Although many experimental and computational methods have been developed to study these proteins, the full characterization of all protein data still remains incomplete. Hence, to help provide a better understanding of cellular functions, various bioinformatic approaches were examined and utilized for PPI predictions. This was achieved in four separate studies: 1) benchmarking of *mirrortree* based computational protein-protein interaction methods (Chapter 2), 2) domain-domain interactions of the fibrillin-1 family (Chapter 3), 3) intramolecular interaction predictions via mutual information and partial correlation (Chapter 4), and 4) characterizing functional binding sites for GPCRs in relation to PRINTS motifs (Chapter 5).

A comprehensive benchmarking study was carried out to examine the popular *mirrortree* approach for PPI predictions. As it has been reported that much of the PPI data generated using high throughput experimental techniques do not overlap, it was important to obtain multiple sets of positive and negative data to ensure an unbiased study. Furthermore, separate prokaryotic and eukaryotic datasets were generated, and various approaches were used to explore any potential bias caused by these two different cell types. The results have shown that datasets generated based on different approaches can lead to very different prediction performance. Moreover, Tan+, the positive dataset that was generated based on multiple experimental evidence performed better than Hakes+, the structure-based positive dataset. Such a finding is in agreement with an earlier study (Yeang and Haussler, 2007), which found that coevolving residue pairs are generally closer in 3-dimensional space, but physically contacting residues are not necessarily coevolving. Similar to positive datasets, different negative datasets

performed quite differently. It has been observed that higher sequence diversity produces better predictions. This trend was observed in the prediction results based on the data generated by different orthologue selection methods. For instance, the two top-hit BLAST based orthologue selection methods, BLAST-SwissProt and BLAST-Proteomes, generated MSAs with higher sequence diversity, and therefore resulted in better prediction performance. To further verify the association between sequence diversity and prediction performance, a sequence diversity experiment was systematically carried out in order to estimate the prediction performance of selected datasets with increasing and decreasing sequence diversity. Predictably, the increasing diversity test set was found to be able to distinguish between the positive and negative data better than the decreasing diversity test set. Furthermore, none of the five different distance methods evaluated appeared to be superior to the others. However, the additional speciation signal removal step substantially increased the predictive power of the TREE method. In particular, the two RNA-based speciation signal removal methods (RNA_TREE1 and RNA_TREE2) performed better than the non-RNA-based method (UAVE_TREE), with RNA_TREE1 being slightly more effective. Since RNA_TREE1 requires the identification of proteins that reflect the same evolutionary history as the equivalent small subunit rRNA tree, it is important to use complete proteomes to ensure that the proteins obtained are the most similar to the RNA. While it is necessary to use complete proteomes for the RNA_TREE1 method, RNA_TREE2 only requires the corresponding rRNA sequences, and hence can be used as an alternative method if complete proteomes are not available. In an attempt to further improve predictive power, an entropy reduction step was implemented to remove highly variable MSA columns prior to the computation for PPI predictions. However, no consistent trend was found to signify that restricting MSAs to more conserved regions can improve the prediction of PPIs.

Utilizing a multi-domain protein (fibrillin-1) family as the test set, *mirrortree* was extended from a full sequence-based PPI prediction method to identify domain-domain interactions. Each of the 56 domains was extracted from the fibrillin-1 MSA and treated as an independent MSA for the analysis. The TREE and UAVE_TREE methods were then applied to predict domain-domain interactions. As a result, ‘close to random’ (AUC = 0.5) and non-significant AUC scores were obtained for both methods. According to the benchmarking study (Chapter 2), there is a positive association

between sequence diversity and prediction performance. This perhaps explains the poor prediction results obtained for the fibrillin-1 study, as the MSAs used for the study have very low sequence diversity. Due to the high conservation of fibrilliln-1 orthologues, coevolution based approaches are not suitable to be used on this protein family for protein interaction predictions.

Once a pair of proteins is predicted to be interacting (e.g. using the *mirrortree* approach), contacting positions within the protein can then be identified using a residue-based approach based on Shannon's information theory. The mutual information (MI) statistic was utilized to measure the likelihood of two proteins contacting. As highly conserved regions in a MSA cannot be used to compute MI scores, an entropy cutoff of 0.3 was used to remove such regions. The predictions obtained after applying the entropy cutoff were more accurate than the predictions acquired without the use of any entropy cutoff, suggesting that higher sequence diversity leads to better predictions. However, MSAs consist of highly divergent sequences often contain a large number of gaps so that functionally conserved residues can be properly aligned. As such, different gap handling approaches were examined, with the best method appearing to be 21_AMINO_ACID, which treats any gap as an artificial 21st amino acid. Different from the other two methods, the 21_AMINO_ACID method retains all residues for the calculation of MI scores; hence the overall entropy is not reduced, and no important functional residues are accidentally removed. Furthermore, different reduced alphabet groupings were examined and compared to the original 20 amino acid alphabet, which was found to be the best performing alphabet grouping. As all coevolution based studies use MSAs, it is inevitable that they all suffer from the phylogenetic bias. In an attempt to remove such background bias, a speciation signal correction method, APC, was implemented. Higher AUC scores were obtained after the application of APC, indicating the effectiveness of this method. To further improve the prediction of intramolecular protein interactions with additional information from a third position, the partial correlation coefficient score was computed for each position pair. This approach appears to be highly effective, as contacting positions were detected with high accuracy. Furthermore, clusters consisting of three residues were predicted by extracting the position combinations with the highest partial correlation coefficients. Clusters identified using this approach often consist of two close positions, with the third position slightly farther away.

Lastly, the characteristics of four types of interactions in G protein-coupled receptors (GPCRs) were determined in relation to PRINTS fingerprints. Prior to comparing the GPCR binding sites and the PRINTS fingerprints, it was important to ensure the significance of PRINTS fingerprints. To do this, a randomization experiment was carried out to determine the background motif distribution for each fingerprint analyzed in the study. This was achieved by randomly selecting motifs using similar criteria as the PRINTS motifs. The background motif distribution was determined after 1000 iterations. The probability of a residue being selected as part of a PRINTS fingerprint motif was estimated to be 7 times higher than an artificial motif residue at the same position for family-level motifs. To a lesser extent, the same trend was observed for subfamily-level motifs, as subfamily PRINTS motif residues were 4 times more likely to reside in their locations when compared to the artificial motif residues. This suggests that it is unlikely that PRINTS motifs were generated by random chance. As all PRINTS motifs in a fingerprint were selected from highly conserved regions that uniquely represent the corresponding family or subfamily, it is likely that these conserved regions have some functional roles. Furthermore, to determine whether there is an association between PRINTS motifs and functionally important sites, binding sites for ligand-binding, G protein-coupling, oligomerization and general PPI were examined and compared to the corresponding PRINTS fingerprint motifs at family- and subfamily-level. More than half of all four types of binding sites were found to be in close proximity to PRINTS motifs, and higher proportions of ligand-binding sites were found to be near family-level motifs. Although the same trend was also observed for oligomerization sites, the difference between family-level motif and subfamily-level motif association was fairly small. As for the G protein-coupling and PPI sites, they appear to have stronger associations with subfamily-level motifs than they do to family-level motifs. In order to identify a more universal relationship between the PRINTS motifs and the four functional sites, PRINTS motif distribution was determined and compared to functional binding sites. Several PRINTS motif hotspots were identified. While G protein-coupling and PPI sites were found to overlap all intracellular hotspots, no apparent association between ligand-binding and oligomerization sites was identified. Additionally, all PPI sites overlap some of the G protein-coupling sites. Finally, a surface patch analysis was carried out to determine whether motifs cluster together and form patches on the surface of a structure, and if so, what the relationship between these surface patches and their functional binding sites is. Indeed, many of the

non-adjacent motifs in linear sequence cluster together when folded in a 3-dimensional structure, and approximately 80% of functional binding sites were found to overlap a motif patch on the surface. Consistent with an earlier observation, a positive association between ligand-binding sites and family-level motifs was observed. Such an association is even more prominent than in linear sequences.

A recent study (Wu *et al.*, 2010) has reported the completion of several crystal structures consisting of the CXCR4 chemokine GPCR and two of its antagonists. As this is the first time a reported GPCR structure was activated by its indigenous antagonist, valuable knowledge can be gained by studying these structures. Although the CXCR4 structures do have the common seven transmembrane helices as all the other GPCRs, certain regions in CXCR4 differ quite substantially from other known GPCR structures. This makes homology modeling quite challenging. As a possible future project, using the methods studied in this thesis and the structures of the newly solved CXCR4 GPCR, various types of GPCR features can be examined to gain insights into cancer metastasis and HIV infection. As an initial step, *mirrortree* can be utilized to identify additional binding PPI partners for CXCR4. It should be noted that, as suggested by various studies (Pazos *et al.*, 1997; Yeang and Haussler, 2007; Burger and van Nimwegen, 2008), coevolving partners are generally closer in 3-dimensional space, but physically contacting partners are not necessarily coevolving. Therefore, it is inevitable that some true interactions will not be identified by the *mirrortree* approach. Nevertheless, the *mirrortree* approach still can be utilized for intermolecular interactions where solved crystal structures are not available. Since sequence diversity appears to be a crucial factor for coevolution based methods, it is important to obtain a large number of CXCR4 proteins from a wide range of proteomes. In the most recent UniProt release (UniProtKB Release 2011_07), CXCR4 sequences were available for 152 species. Such a large number of available CXCR4 sequences exceeds the suggestion from Martin *et al.* (2005) that a minimum of 125 sequences be used in order to attain meaningful results for residue-residue contacting predictions. Once putative interacting partners are identified, mutual information based methods can be applied to further determine the exact interaction interface. Although it was not examined in this thesis, it may also be possible to use mutual information for identifying interacting partners. Since contacting residues generally have higher MI statistics, it is reasonable to assume that interacting protein pairs will also have higher average MI scores than

non-interacting protein pairs. If this assumption remains valid after being tested on a number of known interactions, this could potentially be used as a supplementary approach to *mirrortree* for identifying physically contacting interacting partners. As shown in the GPCR study in Chapter 5, PRINTS motifs appear to have a strong association with different GPCR functions. It would be interesting to see whether the trends observed based on the rhodopsin structure still remain valid on the CXCR4 structures. In addition, the MI analysis can be combined with subfamily PRINTS motifs to identify potential PPI binding sites for CXCR4. Although due to the low sequence diversity of the test MSAs, the fibrillin-1 analysis in Chapter 3 did not result in a definitive conclusion that *mirrortree* is useful for predicting domain-domain interactions, domain-domain interactions can be examined again using the large number of available CXCR4 sequences. As family-level PRINTS motifs have been shown to be associated with ligand-binding in GPCR, family-level PRINTS motifs can be combined with *mirrortree* to predict GPCR protein ligand specifications. Essentially, each family-level motif for CXCR4 can be used as a separate MSA, and compared with the MSA of a potential ligand. Certainly, as a possible future project, the CXCR4 chemokine GPCR can be used to further examine and integrate the methods presented in this thesis and, potentially, the analysis can also be extended to other GPCR proteins for interaction predictions.

References

Ai LS, Liao F (2002) Mutating the four extracellular cysteines in the chemokine receptor CCR6 reveals their differing roles in receptor trafficking, ligand binding, and signaling. Biochemistry 41(26): 8332-8341

Albert PR, Morris SJ, Ghahremani MH, Storring JM, Lembo PM (1998) A putative alpha-helical G beta gamma-coupling domain in the second intracellular loop of the 5-HT1A receptor. Ann N Y Acad Sci 861: 146-161

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, P. W (2002) Molecular Biology of the Cell - Fourth Edition, 4 edn. New York: Garland Science.

Alberts GL, Pregenzer JF, Im WB (1998) Contributions of cysteine 114 of the human D3 dopamine receptor to ligand binding and sensitivity to external oxidizing agents. Br J Pharmacol 125(4): 705-710

Allman K, Page KM, Curtis CA, Hulme EC (2000) Scanning mutagenesis identifies amino acid side chains in transmembrane domain 5 of the M(1) muscarinic receptor that participate in binding the acetyl methyl group of acetylcholine. Mol Pharmacol 58(1): 175-184

Almaula N, Ebersole BJ, Ballesteros JA, Weinstein H, Sealfon SC (1996a) Contribution of a helix 5 locus to selectivity of hallucinogenic and nonhallucinogenic ligands for the human 5-hydroxytryptamine2A and 5-hydroxytryptamine2C receptors: direct and indirect effects on ligand affinity mediated by the same locus. Mol Pharmacol 50(1): 34-42

Almaula N, Ebersole BJ, Zhang D, Weinstein H, Sealfon SC (1996b) Mapping the binding site pocket of the serotonin 5-Hydroxytryptamine2A receptor. Ser3.36(159) provides a second interaction site for the protonated amine of serotonin but not of lysergic acid diethylamide or bufotenin. J Biol Chem 271(25): 14672-14675

Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L, Russell RB (2004) Structure-based assembly of protein complexes in yeast. Science 303(5666): 2026-2029

Aloy P, Russell RB (2004) Ten thousand interactions for the molecular biologist. Nat Biotechnol 22(10): 1317-1321

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3): 403-410

Arai H, Monteclaro FS, Tsou CL, Franci C, Charo IF (1997) Dissociation of chemotaxis from agonist-induced receptor internalization in a lymphocyte cell line transfected with CCR2B. Evidence that directed migration does not require rapid modulation of signaling at the receptor level. J Biol Chem 272(40): 25037-25042

Ashworth JL, Kelly V, Wilson R, Shuttleworth CA, Kielty CM (1999) Fibrillin assembly: dimer formation mediated by amino-terminal sequences. J Cell Sci 112 (Pt 20): 3549-3558

Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol Biol Evol 17(1): 164-178

Attwood TK (2002) The PRINTS database: a resource for identification of protein families. Brief Bioinform 3(3): 252-263

Attwood TK, Beck ME, Bleasby AJ, Parry-Smith DJ (1994) PRINTS--a database of protein motif fingerprints. Nucleic Acids Res 22(17): 3590-3596

Attwood TK, Findlay JB (1993) Design of a discriminating fingerprint for G-protein-coupled receptors. Protein engineering 6(2): 167-176

Auger GA, Pease JE, Shen X, Xanthou G, Barker MD (2002) Alanine scanning mutagenesis of CCR3 reveals that the three intracellular loops are essential for functional receptor expression. Eur J Immunol 32(4): 1052-1058

Bacardit J, Stout M, Hirst JD, Valencia A, Smith RE, Krasnogor N (2009) Automated alphabet reduction for protein datasets. BMC Bioinformatics 10: 6

Bader GD, Betel D, Hogue CW (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 31(1): 248-250

Bagowski CP, Bruins W, Te Velthuis AJ (2010) The nature of protein domain evolution: shaping the interaction network. Curr Genomics 11(5): 368-376

Baldwin JM, Schertler GF, Unger VM (1997) An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. J Mol Biol 272(1): 144-164

Ballesteros J, Weinstein H (1995) Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in g protein coupled receptors. Methods in Neuroscience 25: 366-428

Becamel C, Figge A, Poliak S, Dumuis A, Peles E, Bockaert J, Lubbert H, Ullmer C (2001) Interaction of serotonin 5-hydroxytryptamine type 2C receptors with PDZ10 of the multi-PDZ domain protein MUPP1. J Biol Chem 276(16): 12974-12982

Ben-Baruch A, Bengali KM, Biragyn A, Johnston JJ, Wang JM, Kim J, Chuntharapai A, Michiel DF, Oppenheim JJ, Kelvin DJ (1995) Interleukin-8 receptor beta. The role of the carboxyl terminus in signal transduction. J Biol Chem 270(16): 9121-9128

Ben-Hur A, Noble WS (2006) Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics 7 Suppl 1: S2

Berkhout TA, Blaney FE, Bridges AM, Cooper DG, Forbes IT, Gribble AD, Groot PH, Hardy A, Ife RJ, Kaur R, Moores KE, Shillito H, Willetts J, Witherington J (2003) CCR2: characterization of the antagonist binding site from a combined receptor modeling/mutagenesis approach. J Med Chem 46(19): 4070-4086

Bermak JC, Li M, Bullock C, Zhou QY (2001) Regulation of transport of the dopamine D1 receptor by a new membrane-associated ER protein. Nat Cell Biol 3(5): 492-498

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28(1): 235-242

Bieniasz PD, Fridell RA, Aramori I, Ferguson SS, Caron MG, Cullen BR (1997) HIV-1-induced cell fusion is mediated by multiple regions within both the viral envelope and the CCR-5 co-receptor. Embo J 16(10): 2599-2609

Blanpain C, Doranz BJ, Bondue A, Govaerts C, De Leener A, Vassart G, Doms RW, Proudfoot A, Parmentier M (2003) The core domain of chemokines binds CCR5 extracellular domains while their amino terminus interacts with the transmembrane helix bundle. J Biol Chem 278(7): 5179-5187

Blanpain C, Doranz BJ, Vakili J, Rucker J, Govaerts C, Baik SS, Lorthioir O, Migeotte I, Libert F, Baleux F, Vassart G, Doms RW, Parmentier M (1999) Multiple charged and aromatic residues in CCR5 amino-terminal domain are involved in high affinity binding of both chemokines and HIV-1 Env protein. J Biol Chem 274(49): 34719-34727

Blanpain C, Lee B, Tackoen M, Puffer B, Boom A, Libert F, Sharron M, Wittamer V, Vassart G, Doms RW, Parmentier M (2000) Multiple nonfunctional alleles of CCR5 are frequent in various human populations. Blood 96(5): 1638-1645

Blin N, Yun J, Wess J (1995) Mapping of single amino acid residues required for selective activation of Gq/11 by the m3 muscarinic acetylcholine receptor. J Biol Chem 270(30): 17741-17748

Bluml K, Mutschler E, Wess J (1994a) Functional role in ligand binding and receptor activation of an asparagine residue present in the sixth transmembrane domain of all muscarinic acetylcholine receptors. J Biol Chem 269(29): 18870-18876

Bluml K, Mutschler E, Wess J (1994b) Functional role of a cytoplasmic aromatic amino acid in muscarinic receptor-mediated activation of phospholipase C. J Biol Chem 269(15): 11537-11541

Bluml K, Mutschler E, Wess J (1994c) Insertion mutagenesis as a tool to predict the secondary structure of a muscarinic receptor domain determining specificity of G-protein coupling. Proc Natl Acad Sci U S A 91(17): 7980-7984

Bock JR, Gough DA (2001) Predicting protein--protein interactions from primary structure. Bioinformatics 17(5): 455-460

Boeckler F, Lanig H, Gmeiner P (2005) Modeling the similarity and divergence of dopamine D2-like receptors and identification of validated ligand-receptor complexes. J Med Chem 48(3): 694-709

Boess FG, Monsma FJ, Jr., Meyer V, Zwingelstein C, Sleight AJ (1997) Interaction of tryptamine and ergoline compounds with threonine 196 in the ligand binding site of the 5-hydroxytryptamine6 receptor. Mol Pharmacol 52(3): 515-523

Boess FG, Monsma FJ, Jr., Sleight AJ (1998) Identification of residues in transmembrane regions III and VI that contribute to the ligand binding site of the serotonin 5-HT6 receptor. J Neurochem 71(5): 2169-2177

Bofill-Cardona E, Kudlacek O, Yang Q, Ahorn H, Freissmuth M, Nanoff C (2000) Binding of calmodulin to the D2-dopamine receptor reduces receptor signaling by arresting the G protein activation switch. J Biol Chem 275(42): 32672-32680

Bornberg-Bauer E, Huylmans AK, Sikosek T (2010) How do new proteins arise? Curr Opin Struct Biol 20(3): 390-396

Bradford JR, Westhead DR (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics 21(8): 1487-1494

Breitwieser GE (2004) G protein-coupled receptor oligomerization: implications for G protein activation and cell signaling. Circ Res 94(1): 17-27

Brelot A, Heveker N, Adema K, Hosie MJ, Willett B, Alizon M (1999) Effect of mutations in the second extracellular loop of CXCR4 on its utilization by human and feline immunodeficiency viruses. J Virol 73(4): 2576-2586

Brelot A, Heveker N, Montes M, Alizon M (2000) Identification of residues of CXCR4 critical for human immunodeficiency virus coreceptor and chemokine receptor activities. J Biol Chem 275(31): 23736-23744

Brelot A, Heveker N, Pleskoff O, Sol N, Alizon M (1997) Role of the first and third extracellular domains of CXCR-4 in human immunodeficiency virus coreceptor activity. J Virol 71(6): 4744-4751

Bremm S, Schreck T, Boba P, Held S, Hamacher K (2010) Computing and visually analyzing mutual information in molecular co-evolution. BMC Bioinformatics 11: 330

Brown CA, Brown KS (2010) Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my! PLoS One 5(6): e10779

Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D (2005) The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res 33(Database issue): D212-215

Burger L, van Nimwegen E (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol 4: 165

Burstein ES, Spalding TA, Brann MR (1996) Amino acid side chains that define muscarinic receptor/G-protein coupling. Studies of the third intracellular loop. J Biol Chem 271(6): 2882-2885

Burstein ES, Spalding TA, Brann MR (1998) Structure/function relationships of a G-protein coupling pocket formed by the third intracellular loop of the m5 muscarinic receptor. Biochemistry 37(12): 4052-4058

Burstein ES, Spalding TA, Hill-Eubanks D, Brann MR (1995) Structure-function of muscarinic receptor coupling to G proteins. Random saturation mutagenesis identifies a critical determinant of receptor affinity for G proteins. J Biol Chem 270(7): 3141-3146

Buslje CM, Santos J, Delfino JM, Nielsen M (2009) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics 25(9): 1125-1131

Canals M, Marcellino D, Fanelli F, Ciruela F, de Benedetti P, Goldberg SR, Neve K, Fuxe K, Agnati LF, Woods AS, Ferre S, Lluis C, Bouvier M, Franco R (2003) Adenosine A2A-dopamine D2 receptor-receptor heteromerization: qualitative and quantitative assessment by fluorescence and bioluminescence energy transfer. J Biol Chem 278(47): 46741-46749

Cao TT, Deacon HW, Reczek D, Bretscher A, von Zastrow M (1999) A kinase-regulated PDZ-domain interaction controls endocytic sorting of the beta2-adrenergic receptor. Nature 401(6750): 286-290

Carrillo JJ, Lopez-Gimenez JF, Milligan G (2004) Multiple interactions between transmembrane helices generate the oligomeric alpha1b-adrenoceptor. Mol Pharmacol 66(5): 1123-1137

Carrillo JJ, Pediani J, Milligan G (2003) Dimers of class A G protein-coupled receptors function via agonist-mediated trans-activation of associated G proteins. J Biol Chem 278(43): 42578-42587

Cavalli A, Fanelli F, Taddei C, De Benedetti PG, Cotecchia S (1996) Amino acids of the alpha1B-adrenergic receptor involved in agonist binding: differences in docking catecholamines to receptor subtypes. FEBS Lett 399(1-2): 9-13

Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G (2009) MINT, the molecular interaction database: 2009 update. Nucleic Acids Res 38(Database issue): D532-539

Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G (2010) MINT, the molecular interaction database: 2009 update. Nucleic Acids Res 38(Database issue): D532-539

Chaar ZY, Jackson A, Tiberi M (2001) The cytoplasmic tail of the D1A receptor subtype: identification of specific domains controlling dopamine cellular responsiveness. J Neurochem 79(5): 1047-1058

Chabot DJ, Zhang PF, Quinnan GV, Broder CC (1999) Mutagenesis of CXCR4 identifies important domains for human immunodeficiency virus type 1 X4 isolate envelope-mediated membrane fusion and virus entry and reveals cryptic coreceptor activity for R5 isolates. J Virol 73(8): 6598-6609

Chanda PK, Minchin MC, Davis AR, Greenberg L, Reilly Y, McGregor WH, Bhat R, Lubeck MD, Mizutani S, Hung PP (1993) Identification of residues important for ligand binding to the human 5-hydroxytryptamine1A serotonin receptor. Mol Pharmacol 43(4): 516-520

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res 35(Database issue): D572-574

Chen R, Snyder M (2010) Yeast proteomics and protein microarrays. J Proteomics 73(11): 2147-2157

Chen S, Xu M, Lin F, Lee D, Riek P, Graham RM (1999) Phe310 in transmembrane VI of the alpha1B-adrenergic receptor is a key switch residue involved in activation and catecholamine ring aromatic bonding. J Biol Chem 274(23): 16320-16330

Chen XW, Liu M (2005) Prediction of protein-protein interactions using random decision forest framework. Bioinformatics 21(24): 4394-4400

Chen Y, Green SR, Almazan F, Quehenberger O (2006) The amino terminus and the third extracellular loop of CX3CR1 contain determinants critical for distinct receptor functions. Mol Pharmacol 69(3): 857-865

Cheng ZJ, Zhao J, Sun Y, Hu W, Wu YL, Cen B, Wu GX, Pei G (2000) beta-arrestin differentially regulates the chemokine receptor CXCR4-mediated signaling and receptor internalization, and this implicates multiple interaction sites between beta-arrestin and CXCR4. J Biol Chem 275(4): 2479-2485

Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, Choi HJ, Kuhn P, Weis WI, Kobilka BK, Stevens RC (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 318(5854): 1258-1265

Cho W, Taylor LP, Akil H (1996) Mutagenesis of residues adjacent to transmembrane prolines alters D1 dopamine receptor binding and signal transduction. Mol Pharmacol 50(5): 1338-1345

Cho W, Taylor LP, Mansour A, Akil H (1995) Hydrophobic residues of the D2 dopamine receptor are important for binding and signal transduction. J Neurochem 65(5): 2105-2115

Choi SS, Li W, Lahn BT (2005) Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis. Nat Genet 37(12): 1367-1371

Chothia C, Janin J (1975) Principles of protein-protein recognition. Nature 256(5520): 705-708

Chothia C, Lesk AM (1982) Evolution of proteins formed by beta-sheets. I. Plastocyanin and azurin. J Mol Biol 160(2): 309-323

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4): 823-826

Choudhary MS, Craigo S, Roth BL (1993) A single point mutation (Phe340-->Leu340) of a conserved phenylalanine abolishes 4-[125I]iodo-(2,5-dimethoxy)phenylisopropylamine and [3H]mesulergine but not [3H]ketanserin binding to 5-hydroxytryptamine2 receptors. Mol Pharmacol 43(5): 755-761

Choudhary MS, Sachs N, Uluer A, Glennon RA, Westkaemper RB, Roth BL (1995) Differential ergoline and ergopeptine binding to 5-hydroxytryptamine2A receptors: ergolines require an aromatic residue at position 340 for high affinity binding. Mol Pharmacol 47(3): 450-457

Chung FZ, Wang CD, Potter PC, Venter JC, Fraser CM (1988) Site-directed mutagenesis and continuous expression of human beta-adrenergic receptors. Identification of a conserved aspartate residue involved in agonist binding and receptor activation. J Biol Chem 263(9): 4052-4055

Ciruela F, Burgueno J, Casado V, Canals M, Marcellino D, Goldberg SR, Bader M, Fuxe K, Agnati LF, Lluis C, Franco R, Ferre S, Woods AS (2004) Combining mass spectrometry and pull-down techniques for the study of receptor heteromerization. Direct epitope-epitope electrostatic interactions between adenosine A2A and dopamine D2 receptors. Anal Chem 76(18): 5354-5363

Cohen-Gihon I, Nussinov R, Sharan R (2007) Comprehensive analysis of co-occurring domain sets in yeast proteins. BMC Genomics 8: 161

Coley C, Woodward R, Johansson AM, Strange PG, Naylor LH (2000) Effect of multiple serine/alanine mutations in the transmembrane spanning region V of the D2 dopamine receptor on ligand binding. J Neurochem 74(1): 358-366

Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics 6(3): 439-450

Colvin RA, Campanella GS, Manice LA, Luster AD (2006) CXCR3 requires tyrosine sulfation for ligand binding and a second extracellular loop arginine residue for ligand-induced chemotaxis. Mol Cell Biol 26(15): 5838-5849

Cormier EG, Persuh M, Thompson DA, Lin SW, Sakmar TP, Olson WC, Dragic T (2000) Specific interaction of CCR5 amino-terminal domain peptides containing sulfotyrosines with HIV-1 envelope glycoprotein gp120. Proc Natl Acad Sci U S A 97(11): 5762-5767

Cornell M, Paton NW, Oliver SG (2004) A critical and integrated view of the yeast interactome. Comparative and Functional Genomics 5(5): 382-402

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, Prinz J, St Onge RP, VanderSluis B, Makhnevych T, Vizeacoumar FJ, Alizadeh S, Bahr S, Brost RL, Chen Y, Cokol M, Deshpande R, Li Z, Lin ZY, Liang W, Marback M, Paw J, San Luis BJ, Shuteriqi E, Tong AH, van Dyk N, Wallace IM, Whitney JA, Weirauch MT, Zhong G, Zhu H, Houry WA, Brudno M, Ragibizadeh S, Papp B, Pal C, Roth FP, Giaever G, Nislow C, Troyanskaya OG, Bussey H, Bader GD, Gingras AC, Morris QD, Kim PM, Kaiser CA, Myers CL, Andrews BJ, Boone C (2010) The genetic landscape of a cell. Science 327(5964): 425-431

Cotecchia S, Exum S, Caron MG, Lefkowitz RJ (1990) Regions of the alpha 1-adrenergic receptor involved in coupling to phosphatidylinositol hydrolysis and enhanced sensitivity of biological function. Proc Natl Acad Sci U S A 87(8): 2896-2900

Cox BA, Henningsen RA, Spanoyannis A, Neve RL, Neve KA (1992) Contributions of conserved serine residues to the interactions of ligands with dopamine D2 receptors. J Neurochem 59(2): 627-635

Craig RA, Liao L (2007) Improving protein protein interaction prediction based on phylogenetic information using a least-squares support vector machine. Ann N Y Acad Sci 1115: 154-167

Cravchik A, Gejman PV (1999) Functional analysis of the human D5 dopamine receptor missense and nonsense variants: differences in dopamine binding affinities. Pharmacogenetics 9(2): 199-206

Damaj BB, McColl SR, Neote K, Songqing N, Ogborn KT, Hebert CA, Naccache PH (1996) Identification of G-protein binding sites of the human interleukin-8 receptors by functional mapping of the intracellular loops. Faseb J 10(12): 1426-1434

Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends in biochemical sciences 23(9): 324-328

Daniell SJ, Strange PG, Naylor LH (1994) Site-directed mutagenesis of Tyr417 in the rat D2 dopamine receptor. Biochem Soc Trans 22(2): 144S

Davidson JN, Chen KC, Jamison RS, Musmanno LA, Kern CB (1993) The evolutionary history of the first three enzymes in pyrimidine biosynthesis. Bioessays 15(3): 157-164

de Folter S, Immink RG (2011) Yeast protein-protein interaction assays and screens. Methods Mol Biol 754: 145-165

de Juan D, Mellado M, Rodriguez-Frade JM, Hernanz-Falcon P, Serrano A, del Sol A, Valencia A, Martinez AC, Rojas AM (2005) A framework for computational and experimental methods: identifying dimerization residues in CCR chemokine receptors. Bioinformatics 21 Suppl 2: ii13-18

de Mendonca FL, da Fonseca PC, Phillips RM, Saldanha JW, Williams TJ, Pease JE (2005) Site-directed mutagenesis of CC chemokine receptor 1 reveals the mechanism of action of UCB 35625, a small molecule chemokine receptor antagonist. J Biol Chem 280(6): 4808-4816

Deane CM, Salwinski L, Xenarios I, Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics 1(5): 349-356

DeGraff JL, Gurevich VV, Benovic JL (2002) The third intracellular loop of alpha 2-adrenergic receptors determines subtype specificity of arrestin interaction. J Biol Chem 277(45): 43247-43252

Del Tredici AL, Schiffer HH, Burstein ES, Lameh J, Mohell N, Hacksell U, Brann MR, Weiner DM (2004) Pharmacology of polymorphic variants of the human 5-HT1A receptor. Biochem Pharmacol 67(3): 479-490

Deng M, Mehta S, Sun F, Chen T (2002) Inferring domain-domain interactions from protein-protein interactions. Genome research 12(10): 1540-1548

D'Haeseleer P, Church GM (2004) Estimating and improving protein interaction error rates. Proc IEEE Comput Syst Bioinform Conf: 216-223

Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, Trave G, Gibson TJ (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci 13: 6580-6603

Diviani D, Lattion AL, Abuin L, Staub O, Cotecchia S (2003) The adaptor complex 2 directly interacts with the alpha 1b-adrenergic receptor and plays a role in receptor endocytosis. J Biol Chem 278(21): 19331-19340

Dohlman HG, Caron MG, Strader CD, Amlaiky N, Lefkowitz RJ (1988) Identification and sequence of a binding site peptide of the beta 2-adrenergic receptor. Biochemistry 27(6): 1813-1817

Doranz BJ, Lu ZH, Rucker J, Zhang TY, Sharron M, Cen YH, Wang ZX, Guo HH, Du JG, Accavitti MA, Doms RW, Peiper SC (1997) Two distinct CCR5 domains can mediate coreceptor usage by human immunodeficiency virus type 1. J Virol 71(9): 6305-6314

Doranz BJ, Orsini MJ, Turner JD, Hoffman TL, Berson JF, Hoxie JA, Peiper SC, Brass LF, Doms RW (1999) Identification of CXCR4 domains that support coreceptor and chemokine receptor functions. J Virol 73(4): 2752-2761

Dorsam RT, Gutkind JS (2007) G-protein-coupled receptors and cancer. Nat Rev Cancer 7(2): 79-94

Dragic T, Trkola A, Lin SW, Nagashima KA, Kajumo F, Zhao L, Olson WC, Wu L, Mackay CR, Allaway GP, Sakmar TP, Moore JP, Maddon PJ (1998) Amino-terminal substitutions in the CCR5 coreceptor impair gp120 binding and human immunodeficiency virus type 1 entry. J Virol 72(1): 279-285

Dragic T, Trkola A, Thompson DA, Cormier EG, Kajumo FA, Maxwell E, Lin SW, Ying W, Smith SO, Sakmar TP, Moore JP (2000) A binding pocket for a small molecule inhibitor of HIV-1 entry within the transmembrane helices of CCR5. Proc Natl Acad Sci U S A 97(10): 5639-5644

Duerson K, Carroll R, Clapham D (1993) Alpha-helical distorting substitution disrupt coupling between m3 muscarinic receptor and G proteins. FEBS Lett 324(1): 103-108

Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24(3): 333-340

Eason MG, Liggett SB (1995) Identification of a Gs coupling domain in the amino terminus of the third intracellular loop of the alpha 2A-adrenergic receptor. Evidence for distinct structural determinants that confer Gs versus Gi coupling. J Biol Chem 270(42): 24753-24760

Eason MG, Liggett SB (1996) Chimeric mutagenesis of putative G-protein coupling domains of the alpha2A-adrenergic receptor. Localization of two redundant and fully competent gi coupling domains. J Biol Chem 271(22): 12826-12832

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5): 1792-1797

Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 18(10): 529-536

Ehrlich PR, Raven PH (1964) Butterflies and plants: a study in coevolution. Evolution 18: 586-608

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95(25): 14863-14868

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402(6757): 86-90

Fan GH, Yang W, Sai J, Richmond A (2001a) Phosphorylation-independent association of CXCR2 with the protein phosphatase 2A core enzyme. J Biol Chem 276(20): 16960-16968

Fan GH, Yang W, Sai J, Richmond A (2002) Hsc/Hsp70 interacting protein (hip) associates with CXCR2 and regulates the receptor signaling and trafficking. J Biol Chem 277(8): 6590-6597

Fan GH, Yang W, Wang XJ, Qian Q, Richmond A (2001b) Identification of a motif in the carboxyl terminus of CXCR2 that is involved in adaptin 2 binding and receptor internalization. Biochemistry 40(3): 791-800

Farzan M, Babcock GJ, Vasilieva N, Wright PL, Kiprilov E, Mirzabekov T, Choe H (2002) The role of post-translational modifications of the CXCR4 amino terminus in stromal-derived factor 1 alpha association and HIV-1 entry. J Biol Chem 277(33): 29484-29489

Farzan M, Choe H, Vaca L, Martin K, Sun Y, Desjardins E, Ruffing N, Wu L, Wyatt R, Gerard N, Gerard C, Sodroski J (1998) A tyrosine-rich region in the N terminus of CCR5 is important for human immunodeficiency virus type 1 entry and mediates an association between gp120 and CCR5. J Virol 72(2): 1160-1164

Felsenstein J (1985) Phylogenies and the comparative method. American Naturalist 125: 1-15

Ferrer-Costa C, Orozco M, de la Cruz X (2007) Characterization of compensated mutations in terms of structural and physico-chemical properties. J Mol Biol 365(1): 249-256

Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. Nature 340(6230): 245-246

Filteau F, Veilleux F, Levesque D (1999) Effects of reciprocal chimeras between the C-terminal portion of third intracellular loops of the human dopamine D2 and D3 receptors. FEBS Lett 447(2-3): 251-256

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. Nucleic Acids Res 38(Database issue): D211-222

Fitch WM (1971) Rate of change of concomitantly variable codons. Journal of molecular evolution 1(1): 84-96

Fitch WM, Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4(5): 579-593

Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S (2008) Ensembl 2008. Nucleic Acids Res 36(Database issue): D707-714

Fodor AA, Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 56(2): 211-221

Fong AM, Alam SM, Imai T, Haribabu B, Patel DD (2002) CX3CR1 tyrosine sulfation enhances fractalkine-induced cell adhesion. J Biol Chem 277(22): 19418-19423

Fraser CM (1989) Site-directed mutagenesis of beta-adrenergic receptors. Identification of conserved cysteine residues that independently affect ligand binding and receptor activation. J Biol Chem 264(16): 9266-9270

Fraser CM, Chung FZ, Wang CD, Venter JC (1988) Site-directed mutagenesis of human beta-adrenergic receptors: substitution of aspartic acid-130 by asparagine produces a receptor with high-affinity agonist binding that is uncoupled from adenylate cyclase. Proc Natl Acad Sci U S A 85(15): 5478-5482

Fraser CM, Wang CD, Robinson DA, Gocayne JD, Venter JC (1989) Site-directed mutagenesis of m1 muscarinic acetylcholine receptors: conserved aspartic acids play important roles in receptor function. Mol Pharmacol 36(6): 840-847

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296(5568): 750-752

Fraser HB, Hirsh AE, Wall DP, Eisen MB (2004) Coevolution of gene expression among interacting proteins. Proc Natl Acad Sci U S A 101(24): 9033-9038

Fryxell KJ (1996) The coevolution of gene family trees. Trends Genet 12(9): 364-369

Fu D, Ballesteros JA, Weinstein H, Chen J, Javitch JA (1996) Residues in the seventh membrane-spanning segment of the dopamine D2 receptor accessible in the binding-site crevice. Biochemistry 35(35): 11278-11285

Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38(3): 285-293

Gantz I, DelValle J, Wang LD, Tashiro T, Munzert G, Guo YJ, Konda Y, Yamada T (1992) Molecular basis for the interaction of histamine with the histamine H2 receptor. J Biol Chem 267(29): 20840-20843

Gaulton A (2004) G Protein-Coupled Receptors: Relating Sequence Motifs to Structure and Function. PhD Thesis, University of Manchester,

Gaulton A, Attwood TK (2003) Bioinformatics approaches for the classification of G-protein-coupled receptors. Curr Opin Pharmacol 3(2): 114-120

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440(7084): 631-636

Gelber EI, Kroeze WK, Willins DL, Gray JA, Sinar CA, Hyde EG, Gurevich V, Benovic J, Roth BL (1999) Structure and function of the third intracellular loop of the 5-hydroxytryptamine2A receptor: the third intracellular loop is alpha-helical and binds purified arrestins. J Neurochem 72(5): 2206-2214

Gerlach LO, Skerlj RT, Bridger GJ, Schwartz TW (2001) Molecular interactions of cyclam and bicyclam non-peptide antagonists with the CXCR4 chemokine receptor. J Biol Chem 276(17): 14153-14160

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS (2003) Global analysis of protein expression in yeast. Nature 425(6959): 737-741

Glennon RA, Hong SS, Bondarev M, Law H, Dukat M, Rakhi S, Power P, Fan E, Kinneau D, Kamboj R, Teitler M, Herrick-Davis K, Smith C (1996) Binding of O-alkyl derivatives of serotonin at human 5-HT1D beta receptors. J Med Chem 39(1): 314-322

Gloor GB, Martin LC, Wahl LM, Dunn SD (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 44(19): 7156-7165

Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18(4): 309-317

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274(5287): 546, 563-547

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE (2000) Co-evolution of proteins with their interaction partners. J Mol Biol 299(2): 283-293

Gosling J, Monteclaro FS, Atchison RE, Arai H, Tsou CL, Goldsmith MA, Charo IF (1997) Molecular uncoupling of C-C chemokine receptor 5-induced chemotaxis and signal transduction from HIV-1 coreceptor activity. Proc Natl Acad Sci U S A 94(10): 5061-5066

Gouldson PR, Dean MK, Snell CR, Bywater RP, Gkoutos G, Reynolds CA (2001) Lipid-facing correlated mutations and dimerization in G-protein coupled receptors. Protein engineering 14(10): 759-767

Gouldson PR, Snell CR, Reynolds CA (1997) A new approach to docking in the beta 2-adrenergic receptor that exploits the domain structure of G-protein-coupled receptors. J Med Chem 40(24): 3871-3886

Gouveia-Oliveira R, Roque FS, Wernersson R, Sicheritz-Ponten T, Sackett PW, Molgaard A, Pedersen AG (2009) InterMap3D: predicting and visualizing co-evolving protein residues. Bioinformatics 25(15): 1963-1965

Govaerts C, Blanpain C, Deupi X, Ballet S, Ballesteros JA, Wodak SJ, Vassart G, Pardo L, Parmentier M (2001) The TXP motif in the second transmembrane helix of CCR5. A structural determinant of chemokine-induced activation. J Biol Chem 276(16): 13217-13225

Granas C, Larhammar D (1999) Identification of an amino acid residue important for binding of methiothepin and sumatriptan to the human 5-HT(1B) receptor. Eur J Pharmacol 380(2-3): 171-181

Granas C, Nordquist J, Mohell N, Larhammar D (2001) Site-directed mutagenesis of the 5-HT1B receptor increases the affinity of 5-HT for the agonist low-affinity conformation and reduces the intrinsic activity of 5-HT. Eur J Pharmacol 421(2): 69-76

Granas C, Nordvall G, Larhammar D (1998) Site-directed mutagenesis of the human 5-HT1B receptor. Eur J Pharmacol 349(2-3): 367-375

Graur D, Li WH (2000) Fundamentals of Molecular Evolution, 2 edn.: Sinauer Associates, Inc.

Greasley PJ, Fanelli F, Scheer A, Abuin L, Nenniger-Tosato M, DeBenedetti PG, Cotecchia S (2001) Mutational and computational analysis of the alpha(1b)-adrenergic receptor. Involvement of basic and hydrophobic residues in receptor activation and G protein coupling. J Biol Chem 276(49): 46485-46494

Green SA, Cole G, Jacinto M, Innis M, Liggett SB (1993) A polymorphism of the human beta 2-adrenergic receptor within the fourth transmembrane domain alters ligand binding and functional properties of the receptor. J Biol Chem 268(31): 23116-23121

Guan XM, Peroutka SJ, Kobilka BK (1992) Identification of a single amino acid residue responsible for the binding of a class of beta-adrenergic receptor antagonists to 5-hydroxytryptamine1A receptors. Mol Pharmacol 41(4): 695-698

Guo W, Shi L, Javitch JA (2003) The fourth transmembrane segment forms the interface of the dopamine D2 receptor homodimer. J Biol Chem 278(7): 4385-4388

Gutierrez J, Kremer L, Zaballos A, Goya I, Martinez AC, Marquez G (2004) Analysis of post-translational CCR8 modifications and their influence on receptor activity. J Biol Chem 279(15): 14726-14733

Hakes L, Lovell SC, Oliver SG, Robertson DL (2007) Specificity in protein interactions and its relationship with sequence diversity and coevolution. Proc Natl Acad Sci U S A

Hall RA, Ostedgaard LS, Premont RT, Blitzer JT, Rahman N, Welsh MJ, Lefkowitz RJ (1998) A C-terminal motif found in the beta2-adrenergic receptor, P2Y1 receptor and cystic fibrosis transmembrane conductance regulator determines binding to the Na+/H+ exchanger regulatory factor family of PDZ proteins. Proc Natl Acad Sci U S A 95(15): 8496-8501

Hamacher K (2008) Relating sequence evolution of HIV1-protease to its underlying molecular mechanics. Gene 422(1-2): 30-36

Hamaguchi N, True TA, Saussy DL, Jr., Jeffs PW (1996) Phenylalanine in the second membrane-spanning domain of alpha 1A-adrenergic receptor determines subtype selectivity of dihydropyridine antagonists. Biochemistry 35(45): 14312-14317

Hartwell LH (2004) Yeast and cancer. Biosci Rep 24(4-5): 523-544

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology, Oxford: Oxford University Press.

Hatse S, Princen K, Gerlach LO, Bridger G, Henson G, De Clercq E, Schwartz TW, Schols D (2001) Mutation of Asp(171) and Asp(262) of the chemokine receptor CXCR4 impairs its coreceptor function for human immunodeficiency virus-1 entry and abrogates the antagonistic activity of AMD3100. Mol Pharmacol 60(1): 164-173

Hawes BE, Luttrell LM, Exum ST, Lefkowitz RJ (1994) Inhibition of G protein-coupled receptor signaling by expression of cytoplasmic domains of the receptor. J Biol Chem 269(22): 15776-15785

He J, Bellini M, Inuzuka H, Xu J, Xiong Y, Yang X, Castleberry AM, Hall RA (2006) Proteomic analysis of beta1-adrenergic receptor interactions with PDZ scaffold proteins. J Biol Chem 281(5): 2820-2827

He J, Bellini M, Xu J, Castleberry AM, Hall RA (2004) Interaction with cystic fibrosis transmembrane conductance regulator-associated ligand (CAL) inhibits beta1-adrenergic receptor surface expression. J Biol Chem 279(48): 50190-50196

Hebert CA, Chuntharapai A, Smith M, Colby T, Kim J, Horuk R (1993) Partial functional mapping of the human interleukin-8 type A receptor. Identification of a major ligand binding domain. J Biol Chem 268(25): 18549-18553

Hebert TE, Moffett S, Morello JP, Loisel TP, Bichet DG, Barret C, Bouvier M (1996) A peptide derived from a beta2-adrenergic receptor transmembrane domain inhibits both receptor dimerization and activation. J Biol Chem 271(27): 16384-16392

Heitz F, Holzwarth JA, Gies JP, Pruss RM, Trumpp-Kallmeyer S, Hibert MF, Guenet C (1999) Site-directed mutagenesis of the putative human muscarinic M2 receptor binding site. Eur J Pharmacol 380(2-3): 183-195

Henikoff JG, Henikoff S (1996) Blocks database and its applications. Methods Enzymol 266: 88-105

Henrick K, Thornton JM (1998) PQS: a protein quaternary structure file server. Trends in biochemical sciences 23(9): 358-361

Hernanz-Falcon P, Rodriguez-Frade JM, Serrano A, Juan D, del Sol A, Soriano SF, Roncal F, Gomez L, Valencia A, Martinez AC, Mellado M (2004) Identification of amino acid residues crucial for chemokine receptor dimerization. Nat Immunol 5(2): 216-223

Herrick-Davis K, Grinde E, Harrigan TJ, Mazurkiewicz JE (2005) Inhibition of serotonin 5-hydroxytryptamine2c receptor function through heterodimerization: receptor dimers bind two molecules of ligand and one G-protein. J Biol Chem 280(48): 40144-40151

Hieble JP, Hehr A, Li YO, Ruffolo RR, Jr. (1998) Molecular basis for the stereoselective interactions of catecholamines with alpha-adrenoceptors. Proc West Pharmacol Soc 41: 225-228

Hill CM, Kwon D, Jones M, Davis CB, Marmon S, Daugherty BL, DeMartino JA, Springer MS, Unutmaz D, Littman DR (1998) The amino terminus of human CCR5 is required for its function as a receptor for diverse human and simian immunodeficiency virus envelope glycoproteins. Virology 248(2): 357-371

Hill-Eubanks D, Burstein ES, Spalding TA, Brauner-Osborne H, Brann MR (1996) Structure of a G-protein-coupling domain of a muscarinic receptor predicted by random saturation mutagenesis. J Biol Chem 271(6): 3058-3065

Ho BY, Karschin A, Raymond JR, Branchek T, Lester HA, Davidson N (1992) Expression in animal cells of the 5-HT1A receptor by a vaccinia virus vector system. FEBS Lett 301(3): 303-306

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868): 180-183

Hoffmann R, Krallinger M, Andres E, Tamames J, Blaschke C, Valencia A (2005) Text mining for metabolic pathways, signaling cascades, and protein networks. Sci STKE 2005(283): pe21

Howard OM, Shirakawa AK, Turpin JA, Maynard A, Tobin GJ, Carrington M, Oppenheim JJ, Dean M (1999) Naturally occurring CCR5 extracellular and transmembrane domain variants affect HIV-1 Co-receptor and ligand binding function. J Biol Chem 274(23): 16228-16234

Hu LA, Chen W, Martin NP, Whalen EJ, Premont RT, Lefkowitz RJ (2003) GIPC interacts with the beta1-adrenergic receptor and regulates beta1-adrenergic receptor-mediated ERK activation. J Biol Chem 278(28): 26295-26301

Hu LA, Tang Y, Miller WE, Cong M, Lau AG, Lefkowitz RJ, Hall RA (2000) beta 1-adrenergic receptor association with PSD-95. Inhibition of receptor internalization and facilitation of beta 1-adrenergic receptor interaction with N-methyl-D-aspartate receptors. J Biol Chem 275(49): 38659-38666

Huang XP, Nagy PI, Williams FE, Peseckis SM, Messer WS, Jr. (1999a) Roles of threonine 192 and asparagine 382 in agonist and antagonist interactions with M1 muscarinic receptors. Br J Pharmacol 126(3): 735-745

Huang XP, Williams FE, Peseckis SM, Messer WS, Jr. (1999b) Differential modulation of agonist potency and receptor coupling by mutations of Ser388Tyr and Thr389Pro at the junction of transmembrane domain VI and the third extracellular loop of human M(1) muscarinic acetylcholine receptors. Mol Pharmacol 56(4): 775-783

Hubbard SJ, Thornton JM. (1993) 'NACCESS'. Department of Biochemistry and Molecular Biology, University College of London.

Hubmacher D, El-Hallous EI, Nelea V, Kaartinen MT, Lee ER, Reinhardt DP (2008) Biogenesis of extracellular microfibrils: Multimerization of the fibrillin-1 C terminus into bead-like structures enables self-assembly. Proc Natl Acad Sci U S A 105(18): 6548-6553

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37(Database issue): D211-215

Huttenrauch F, Nitzki A, Lin FT, Honing S, Oppermann M (2002) Beta-arrestin binding to CC chemokine receptor 5 requires multiple C-terminal receptor phosphorylation sites and involves a conserved Asp-Arg-Tyr sequence motif. J Biol Chem 277(34): 30769-30777

Huynen M, Snel B, Lathe W, 3rd, Bork P (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome research 10(8): 1204-1210

Hwa J, Graham RM, Perez DM (1996) Chimeras of alpha1-adrenergic receptor subtypes identify critical residues that modulate active state isomerization. J Biol Chem 271(14): 7956-7964

Ilani T, Fishburn CS, Levavi-Sivan B, Carmon S, Raveh L, Fuchs S (2002) Coupling of dopamine receptors to G proteins: studies with chimeric D2/D3 dopamine receptors. Cell Mol Neurobiol 22(1): 47-56

Isogaya M, Yamagiwa Y, Fujita S, Sugimoto Y, Nagao T, Kurose H (1998) Identification of a key amino acid of the beta2-adrenergic receptor for high affinity binding of salmeterol. Mol Pharmacol 54(4): 616-622

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98(8): 4569-4574

Janovick JA, Patny A, Mosley R, Goulet MT, Altman MD, Rush TS, 3rd, Cornea A, Conn PM (2009) Molecular mechanism of action of pharmacoperone rescue of misrouted GPCR mutants: the GnRH receptor. Mol Endocrinol 23(2): 157-168

Javitch JA, Ballesteros JA, Chen J, Chiappa V, Simpson MM (1999) Electrostatic and aromatic microdomains within the binding-site crevice of the D2 receptor: contributions of the second membrane-spanning segment. Biochemistry 38(25): 7961-7968

Javitch JA, Ballesteros JA, Weinstein H, Chen J (1998) A cluster of aromatic residues in the sixth membrane-spanning segment of the dopamine D2 receptor is accessible in the binding-site crevice. Biochemistry 37(4): 998-1006

Javitch JA, Fu D, Chen J (1995) Residues in the fifth membrane-spanning segment of the dopamine D2 receptor exposed in the binding-site crevice. Biochemistry 34(50): 16433-16439

Javitch JA, Fu D, Chen J (1996) Differentiating dopamine D2 ligands by their sensitivities to modification of the cysteine exposed in the binding-site crevice. Mol Pharmacol 49(4): 692-698

Javitch JA, Fu D, Liapakis G, Chen J (1997) Constitutive activation of the beta2 adrenergic receptor alters the orientation of its sixth membrane-spanning segment. J Biol Chem 272(30): 18546-18549

Javitch JA, Shi L, Simpson MM, Chen J, Chiappa V, Visiers I, Weinstein H, Ballesteros JA (2000) The fourth transmembrane segment of the dopamine D2 receptor: accessibility in the binding-site crevice and position in the transmembrane bundle. Biochemistry 39(40): 12190-12199

Jeanneteau F, Diaz J, Sokoloff P, Griffon N (2004) Interactions of GIPC with dopamine D2, D3 but not D4 receptors define a novel mode of regulation of G protein-coupled receptors. Mol Biol Cell 15(2): 696-705

Jensen AA, Pedersen UB, Kiemer A, Din N, Andersen PH (1995) Functional importance of the carboxyl tail cysteine residues in the human D1 dopamine receptor. J Neurochem 65(3): 1325-1331

Johnson MP, Loncharich RJ, Baez M, Nelson DL (1994) Species variations in transmembrane region V of the 5-hydroxytryptamine type 2A receptor alter the structure-activity relationship of certain ergolines and tryptamines. Mol Pharmacol 45(2): 277-286

Johnson MP, Wainscott DB, Lucaites VL, Baez M, Nelson DL (1997) Mutations of transmembrane IV and V serines indicate that all tryptamines do not bind to the rat 5-HT2A receptor in the same manner. Brain Res Mol Brain Res 49(1-2): 1-6

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8(3): 275-282

Jones RB, Gordus A, Krall JA, MacBeath G (2006) A quantitative protein interaction network for the ErbB receptors using protein microarrays. Nature 439(7073): 168-174

Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc Natl Acad Sci U S A 93(1): 13-20

Jothi R, Kann MG, Przytycka TM (2005) Predicting protein-protein interaction by searching evolutionary tree automorphism space. Bioinformatics 21 Suppl 1: i241-250

Juan D, Pazos F, Valencia A (2008a) Co-evolution and co-adaptation in protein networks. FEBS Lett 582(8): 1225-1230

Juan D, Pazos F, Valencia A (2008b) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc Natl Acad Sci U S A 105(3): 934-939

Kalani MY, Vaidehi N, Hall SE, Trabanino RJ, Freddolino PL, Kalani MA, Floriano WB, Kam VW, Goddard WA, 3rd (2004) The predicted 3D structure of the human D2 dopamine receptor and the binding site and binding affinities for agonists and antagonists. Proc Natl Acad Sci U S A 101(11): 3815-3820

Kann MG, Jothi R, Cherukuri PF, Przytycka TM (2007) Predicting protein domain interactions from coevolution of conserved regions. Proteins 67(4): 811-820

Kann MG, Shoemaker BA, Panchenko AR, Przytycka TM (2009) Correlated evolution of interacting proteins: looking behind the mirrortree. J Mol Biol 385(1): 91-98

Kao HT, Adham N, Olsen MA, Weinshank RL, Branchek TA, Hartig PR (1992) Site-directed mutagenesis of a single residue changes the binding properties of the serotonin 5-HT2 receptor from a human to a rat pharmacology. FEBS Lett 307(3): 324-328

Karlin S, Altschul SF (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. Proc Natl Acad Sci U S A 90(12): 5873-5877

Katancik JA, Sharma A, de Nardin E (2000) Interleukin 8, neutrophil-activating peptide-2 and GRO-alpha bind to and elicit cell activation via specific and different amino acid residues of CXCR2. Cytokine 12(10): 1480-1488

Keane MP, Belperio JA, Xue YY, Burdick MD, Strieter RM (2004) Depletion of CXCR2 inhibits tumor growth and angiogenesis in a murine model of lung cancer. J Immunol 172(5): 2853-2860

Keene DR, Maddox BK, Kuo HJ, Sakai LY, Glanville RW (1991) Extraction of extendable beaded structures and their identification as fibrillin-containing extracellular matrix microfibrils. J Histochem Cytochem 39(4): 441-449

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H (2007) IntAct--open source resource for molecular interaction data. Nucleic Acids Res 35(Database issue): D561-565

Kikkawa H, Isogaya M, Nagao T, Kurose H (1998) The role of the seventh transmembrane region in high affinity binding of a beta 2-selective agonist TA-2005. Mol Pharmacol 53(1): 128-134

Kim WK, Bolser DM, Park JH (2004) Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). Bioinformatics 20(7): 1138-1150

Klein U, Ramirez MT, Kobilka BK, von Zastrow M (1997) A novel interaction between adrenergic receptors and the alpha-subunit of eukaryotic initiation factor 2B. J Biol Chem 272(31): 19099-19102

Ko J, Jang SW, Kim YS, Kim IS, Sung HJ, Kim HH, Park JY, Lee YH, Kim J, Na DS (2004) Human LZIP binds to CCR1 and differentially affects the chemotactic activities of CCR1-dependent chemokines. Faseb J 18(7): 890-892

Kohen R, Fashingbauer LA, Heidmann DE, Guthrie CR, Hamblin MW (2001) Cloning of the mouse 5-HT6 serotonin receptor and mutagenesis studies of the third cytoplasmic loop. Brain Res Mol Brain Res 90(2): 110-117

Kostenis E, Conklin BR, Wess J (1997a) Molecular basis of receptor/G protein coupling selectivity studied by coexpression of wild type and mutant m2 muscarinic receptors with mutant G alpha(q) subunits. Biochemistry 36(6): 1487-1495

Kostenis E, Gomeza J, Lerche C, Wess J (1997b) Genetic analysis of receptor-Galphaq coupling selectivity. J Biol Chem 272(38): 23675-23681

Kraft K, Olbrich H, Majoul I, Mack M, Proudfoot A, Oppermann M (2001) Characterization of sequence determinants within the carboxyl-terminal domain of chemokine receptor CCR5 that regulate signaling and receptor internalization. J Biol Chem 276(37): 34408-34418

Kristiansen K, Kroeze WK, Willins DL, Gelber EI, Savage JE, Glennon RA, Roth BL (2000) A highly conserved aspartic acid (Asp-155) anchors the terminal amine moiety of tryptamines and is involved in membrane targeting of the 5-HT(2A) serotonin receptor but does not participate in activation via a "salt-bridge disruption" mechanism. J Pharmacol Exp Ther 293(3): 735-746

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440(7084): 637-643

Kuhmann SE, Platt EJ, Kozak SL, Kabat D (1997) Polymorphisms in the CCR5 genes of African green monkeys and mice implicate specific amino acids in infections by simian and human immunodeficiency viruses. J Virol 71(11): 8642-8656

Kuipers W, Link R, Standaar PJ, Stoit AR, Van Wijngaarden I, Leurs R, Ijzerman AP (1997) Study of the interaction between aryloxypropanolamines and Asn386 in helix VII of the human 5-hydroxytryptamine1A receptor. Mol Pharmacol 51(5): 889-896

Kulathinal RJ, Bettencourt BR, Hartl DL (2004) Compensated deleterious mutations in insect genomes. Science 306(5701): 1553-1554

Kung LA, Snyder M (2006) Proteome chips for whole-organism assays. Nat Rev Mol Cell Biol 7(8): 617-622

Kung LA, Tao SC, Qian J, Smith MG, Snyder M, Zhu H (2009) Global analysis of the glycoproteome in Saccharomyces cerevisiae reveals new roles for protein glycosylation in eukaryotes. Mol Syst Biol 5: 308

Kurtenbach E, Curtis CA, Pedder EK, Aitken A, Harris AC, Hulme EC (1990) Muscarinic acetylcholine receptors. Peptide sequencing identifies residues involved in antagonist binding and disulfide bond formation. J Biol Chem 265(23): 13702-13708

Kushwaha N, Albert PR (2005) Coupling of 5-HT1A autoreceptors to inhibition of mitogen-activated protein kinase activation via G beta gamma subunit signaling. Eur J Neurosci 21(3): 721-732

Kushwaha N, Harwood SC, Wilson AM, Berger M, Tecott LH, Roth BL, Albert PR (2006) Molecular determinants in the second intracellular loop of the 5-hydroxytryptamine-1A receptor for G-protein coupling. Mol Pharmacol 69(5): 1518-1526

Lee B, Godfrey M, Vitale E, Hori H, Mattei MG, Sarfarazi M, Tsipouras P, Ramirez F, Hollister DW (1991) Linkage of Marfan syndrome and a phenotypically related disorder to two different fibrillin genes. Nature 352(6333): 330-334

Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55(3): 379-400

Lee FJ, Xue S, Pei L, Vukusic B, Chery N, Wang Y, Wang YT, Niznik HB, Yu XM, Liu F (2002) Dual regulation of NMDA receptor functions by direct protein-protein interactions with the dopamine D1 receptor. Cell 111(2): 219-230

Lee KB, Ptasienski JA, Pals-Rylaarsdam R, Gurevich VV, Hosey MM (2000) Arrestin binding to the M(2) muscarinic acetylcholine receptor is precluded by an inhibitory element in the third intracellular loop of the receptor. J Biol Chem 275(13): 9284-9289

Lee SM, Shin H, Jang SW, Shim JJ, Song IS, Son KN, Hwang J, Shin YH, Kim HH, Lee CK, Ko J, Na DS, Kwon BS, Kim J (2004) PLP2/A4 interacts with CCR1 and stimulates migration of CCR1-expressing HOS cells. Biochem Biophys Res Commun 324(2): 768-772

Lee SP, O'Dowd BF, Ng GYK, Varghese G, Akil H, Mansour A, Nguyen T, George SR (2000) Inhibition of cell surface expression by mutant receptors demonstrates that D2 dopamine receptors exist as oligomers in the cell. Molecular Pharmacology 58: 120-128

Lei B, Morris DP, Smith MP, Svetkey LP, Newman MF, Rotter JI, Buchanan TA, Beckstrom-Sternberg SM, Green ED, Schwinn DA (2005) Novel human alpha1a-adrenoceptor single nucleotide polymorphisms alter receptor pharmacology and biological function. Naunyn Schmiedebergs Arch Pharmacol 371(3): 229-239

Lemaire R, Bayle J, Lafyatis R (2006) Fibrillin in Marfan syndrome and tight skin mice provides new insights into transforming growth factor-beta regulation and systemic sclerosis. Curr Opin Rheumatol 18(6): 582-587

Lembo PM, Ghahremani MH, Morris SJ, Albert PR (1997) A conserved threonine residue in the second intracellular loop of the 5-hydroxytryptamine 1A receptor directs signaling specificity. Mol Pharmacol 52(1): 164-171

Leong SR, Kabakoff RC, Hebert CA (1994) Complete mutagenesis of the extracellular domain of interleukin-8 (IL-8) type A receptor identifies charged residues mediating IL-8 binding and signal transduction. J Biol Chem 269(30): 19343-19348

Leppik RA, Miller RC, Eck M, Paquet JL (1994) Role of acidic amino acids in the allosteric modulation by gallamine of antagonist binding at the m2 muscarinic acetylcholine receptor. Mol Pharmacol 45(5): 983-990

Lesk AM, Chothia C (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 136(3): 225-270

Leurs R, Smit MJ, Meeder R, Ter Laak AM, Timmerman H (1995) Lysine200 located in the fifth transmembrane domain of the histamine H1 receptor interacts with histamine but not with all H1 agonists. Biochem Biophys Res Commun 214(1): 110-117

Leurs R, Smit MJ, Tensen CP, Ter Laak AM, Timmerman H (1994) Site-directed mutagenesis of the histamine H1-receptor reveals a selective interaction of asparagine207 with subclasses of H1-receptor agonists. Biochem Biophys Res Commun 201(1): 295-301

Li G, Haney KM, Kellogg GE, Zhang Y (2009) Comparative docking study of anibamine as the first natural product CCR5 antagonist in CCR5 homology models. J Chem Inf Model 49(1): 120-132

Li M, Bermak JC, Wang ZW, Zhou QY (2000) Modulation of dopamine D(2) receptor signaling by actin-binding protein (ABP-280). Mol Pharmacol 57(3): 446-452

Liapakis G, Ballesteros JA, Papachristou S, Chan WC, Chen X, Javitch JA (2000) The forgotten serine. A critical role for Ser-2035.42 in ligand binding to and activation of the beta 2-adrenergic receptor. J Biol Chem 275(48): 37779-37788

Liggett SB, Caron MG, Lefkowitz RJ, Hnatowich M (1991) Coupling of a mutated form of the human beta 2-adrenergic receptor to Gi and Gs. Requirement for multiple cytoplasmic domains in the coupling process. J Biol Chem 266(8): 4816-4821

Ligneau X, Morisset S, Tardivel-Lacombe J, Gbahou F, Ganellin CR, Stark H, Schunack W, Schwartz JC, Arrang JM (2000) Distinct pharmacology of rat and human histamine H(3) receptors: role of two amino acids in the third transmembrane domain. Br J Pharmacol 131(7): 1247-1250

Lima BL, Santos EJ, Fernandes GR, Merkel C, Mello MR, Gomes JP, Soukoyan M, Kerkis A, Massironi SM, Visintin JA, Pereira LV (2010) A new mouse model for marfan syndrome presents phenotypic variability associated with the genetic background and overall levels of Fbn1 expression. PLoS One 5(11): e14136

Lin G, Baribaud F, Romano J, Doms RW, Hoxie JA (2003) Identification of gp120 binding sites on CXCR4 by using CD4-independent human immunodeficiency virus type 2 Env proteins. J Virol 77(2): 931-942

Lin G, Tiedemann K, Vollbrandt T, Peters H, Batge B, Brinckmann J, Reinhardt DP (2002) Homo- and heterotypic fibrillin-1 and -2 interactions constitute the basis for the assembly of microfibrils. J Biol Chem 277(52): 50795-50804

Liotta LA, Espina V, Mehta AI, Calvert V, Rosenblatt K, Geho D, Munson PJ, Young L, Wulfkuhle J, Petricoin EF, 3rd (2003) Protein microarrays: meeting analytical challenges for clinical applications. Cancer Cell 3(4): 317-325

Little DY, Chen L (2009) Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. PLoS One 4(3): e4762

Liu F, Wan Q, Pristupa ZB, Yu XM, Wang YT, Niznik HB (2000) Direct protein-protein coupling enables cross-talk between dopamine D5 and gamma-aminobutyric acid A receptors. Nature 403(6767): 274-280

Liu J, Blin N, Conklin BR, Wess J (1996) Molecular mechanisms involved in muscarinic acetylcholine receptor-mediated G protein activation studied by insertion mutagenesis. J Biol Chem 271(11): 6172-6178

Liu J, Conklin BR, Blin N, Yun J, Wess J (1995) Identification of a receptor/G-protein contact site critical for signaling specificity and G-protein activation. Proc Natl Acad Sci U S A 92(25): 11642-11646

Liu Y, Buck DC, Macey TA, Lan H, Neve KA (2007) Evidence that calmodulin binding to the dopamine D2 receptor enhances receptor signaling. J Recept Signal Transduct Res 27(1): 47-65

Lopez-Gimenez JF, Canals M, Pediani JD, Milligan G (2007) The alpha1b-adrenoceptor exists as a higher-order oligomer: effective oligomerization is required for receptor maturation, surface delivery, and function. Mol Pharmacol 71(4): 1015-1029

Lu Z, Berson JF, Chen Y, Turner JD, Zhang T, Sharron M, Jenks MH, Wang Z, Kim J, Rucker J, Hoxie JA, Peiper SC, Doms RW (1997) Evolution of HIV-1 coreceptor usage through interactions with distinct CCR5 and CXCR4 domains. Proc Natl Acad Sci U S A 94(12): 6426-6431

Lu ZL, Hulme EC (1999) The functional topography of transmembrane domain 3 of the M1 muscarinic acetylcholine receptor, revealed by scanning mutagenesis. J Biol Chem 274(11): 7309-7315

Lu ZL, Saldanha JW, Hulme EC (2001) Transmembrane domains 4 and 7 of the M(1) muscarinic acetylcholine receptor are critical for ligand binding and the receptor activation switch. J Biol Chem 276(36): 34098-34104

Lucas JL, Wang D, Sadee W (2006) Calmodulin binding to peptides derived from the i3 loop of muscarinic receptors. Pharm Res 23(4): 647-653

Lundstrom K, Turpin MP, Large C, Robertson G, Thomas P, Lewell XQ (1998) Mapping of dopamine D3 receptor binding site by pharmacological characterization of mutants expressed in CHO cells with the Semliki Forest virus system. J Recept Signal Transduct Res 18(2-3): 133-150

Luo Z, Butcher DJ, Huang Z (1997) Molecular modeling of interleukin-8 receptor beta and analysis of the receptor-ligand interaction. Protein engineering 10(9): 1039-1045

MacBeath G, Schreiber SL (2000) Printing proteins as microarrays for high-throughput function determination. Science 289(5485): 1760-1763

Maggio R, Novi F, Scarselli M, Corsini GU (2005) The impact of G-protein-coupled receptor hetero-oligomerization on function and pharmacology. FEBS J 272(12): 2939-2946

Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011: bar009

Malek D, Munch G, Palm D (1993) Two sites in the third inner loop of the dopamine D2 receptor are involved in functional G protein-mediated coupling to adenylate cyclase. FEBS Lett 325(3): 215-219

Malmberg A, Strange PG (2000) Site-directed mutations in the third intracellular loop of the serotonin 5-HT(1A) receptor alter G protein coupling from G(i) to G(s) in a ligand-dependent manner. J Neurochem 75(3): 1283-1293

Manivet P, Schneider B, Smith JC, Choi DS, Maroteaux L, Kellermann O, Launay JM (2002) The serotonin binding site of human and murine 5-HT2B receptors: molecular modeling and site-directed mutagenesis. J Biol Chem 277(19): 17170-17178

Mansour A, Meng F, Meador-Woodruff JH, Taylor LP, Civelli O, Akil H (1992) Site-directed mutagenesis of the human dopamine D2 receptor. Eur J Pharmacol 227(2): 205-214

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285(5428): 751-753

Marion S, Oakley RH, Kim KM, Caron MG, Barak LS (2006) A beta-arrestin binding determinant common to the second intracellular loops of rhodopsin family G protein-coupled receptors. J Biol Chem 281(5): 2932-2938

Marjamaki A, Pihlavisto M, Cockcroft V, Heinonen P, Savola JM, Scheinin M (1998) Chloroethylclonidine binds irreversibly to exposed cysteines in the fifth membrane-spanning domain of the human alpha2A-adrenergic receptor. Mol Pharmacol 53(3): 370-376

Marson A, Rock MJ, Cain SA, Freeman LJ, Morgan A, Mellody K, Shuttleworth CA, Baldock C, Kielty CM (2005) Homotypic fibrillin-1 interactions in microfibril assembly. J Biol Chem 280(6): 5013-5021

Martin LC, Gloor GB, Dunn SD, Wahl LM (2005) Using information theory to search for co-evolving residues in proteins. Bioinformatics 21(22): 4116-4124

Matsui H, Lazareno S, Birdsall NJ (1995) Probing of the location of the allosteric site on m1 muscarinic receptors by site-directed mutagenesis. Mol Pharmacol 47(1): 88-98

Matsuo Y, Raimondo M, Woodward TA, Wallace MB, Gill KR, Tong Z, Burdick MD, Yang Z, Strieter RM, Hoffman RM, Guha S (2009) CXC-chemokine/CXCR2 biological axis promotes angiogenesis in vitro and in vivo in pancreatic cancer. Int J Cancer 125(5): 1027-1037

Meszaros B, Tompa P, Simon I, Dosztanyi Z (2007) Molecular principles of the interactions of disordered proteins. J Mol Biol 372(2): 549-561

Mialet J, Dahmoune Y, Lezoualc'h F, Berque-Bestel I, Eftekhari P, Hoebeke J, Sicsic S, Langlois M, Fischmeister R (2000) Exploration of the ligand binding site of the human 5-HT(4) receptor by site-directed mutagenesis and molecular modeling. Br J Pharmacol 130(3): 527-538

Mintseris J, Weng Z (2003) Atomic contact vectors in protein-protein recognition. Proteins 53(3): 629-639

Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci U S A 102(31): 10930-10935

Mirzadegan T, Diehl F, Ebi B, Bhakta S, Polsky I, McCarley D, Mulkins M, Weatherhead GS, Lapierre JM, Dankwardt J, Morgans D, Jr., Wilhelm R, Jarnagin K (2000) Identification of the binding site for a novel class of CCR2b chemokine receptor antagonists: binding to a common chemokine receptor motif within the helical bundle. J Biol Chem 275(33): 25562-25571

Moguilevsky N, Varsalona F, Guillaume JP, Noyer M, Gillard M, Daliers J, Henichart JP, Bollen A (1995) Pharmacological and functional characterisation of the wild-type and site-directed mutants of the human H1 histamine receptor stably expressed in CHO cells. J Recept Signal Transduct Res 15(1-4): 91-102

Monteclaro FS, Charo IF (1996) The amino-terminal extracellular domain of the MCP-1 receptor, but not the RANTES/MIP-1alpha receptor, confers chemokine selectivity. Evidence for a two-step mechanism for MCP-1 receptor activation. J Biol Chem 271(32): 19084-19092

Monteclaro FS, Charo IF (1997) The amino-terminal domain of CCR2 is both necessary and sufficient for high affinity binding of monocyte chemoattractant protein 1. Receptor activation by a pseudo-tethered ligand. J Biol Chem 272(37): 23186-23190

Moro O, Lameh J, Hogger P, Sadee W (1993) Hydrophobic amino acid in the i2 loop plays a key role in receptor-G protein coupling. J Biol Chem 268(30): 22273-22276

Mosser VA, Amana IJ, Schimerlik MI (2002) Kinetic analysis of M2 muscarinic receptor activation of Gi in Sf9 insect cell membranes. J Biol Chem 277(2): 922-931

Moyle WR, Campbell RK, Myers RV, Bernard MP, Han Y, Wang X (1994) Co-evolution of ligand-receptor pairs. Nature 368(6468): 251-255

Nasman J, Jansson CC, Akerman KE (1997) The second intracellular loop of the alpha2-adrenergic receptors determines subtype-specific coupling to cAMP production. J Biol Chem 272(15): 9703-9708

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3): 443-453

Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics: Oxford University Press.

Nelesen S, Liu K, Zhao D, Linder CR, Warnow T (2008) The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses. Pacific Symposium on Biocomputing: 25-36

Neve KA, Cox BA, Henningsen RA, Spanoyannis A, Neve RL (1991) Pivotal role for aspartate-80 in the regulation of dopamine D2 receptor affinity for drugs and inhibition of adenylyl cyclase. Mol Pharmacol 39(6): 733-739

Neve KA, Cumbay MG, Thompson KR, Yang R, Buck DC, Watts VJ, DuRand CJ, Teeter MM (2001) Modeling and mutational analysis of a putative sodium-binding pocket on the dopamine D2 receptor. Mol Pharmacol 60(2): 373-381

Ng GY, O'Dowd BF, Lee SP, Chung HT, Brann MR, Seeman P, George SR (1996) Dopamine D2 receptor dimers and receptor-blocking peptides. Biochem Biophys Res Commun 227(1): 200-204

Nijman SM (2011) Synthetic lethality: general principles, utility and detection using genetic screens in human cells. FEBS Lett 585(1): 1-6

Noivirt O, Eisenstein M, Horovitz A (2005) Detection and reduction of evolutionary noise in correlated mutation analysis. Protein Eng Des Sel 18(5): 247-253

Nonaka H, Otaki S, Ohshima E, Kono M, Kase H, Ohta K, Fukui H, Ichimura M (1998) Unique binding pocket for KW-4679 in the histamine H1 receptor. Eur J Pharmacol 345(1): 111-117

Nooren IM, Thornton JM (2003) Diversity of protein-protein interactions. EMBO J 22(14): 3486-3492

Obosi LA, Hen R, Beadle DJ, Bermudez I, King LA (1997) Mutational analysis of the mouse 5-HT7 receptor: importance of the third intracellular loop for receptor-G-protein interaction. FEBS Lett 412(2): 321-324

O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33(Database issue): D476-480

O'Dowd BF, Hnatowich M, Regan JW, Leader WM, Caron MG, Lefkowitz RJ (1988) Site-directed mutagenesis of the cytoplasmic domains of the human beta 2-adrenergic receptor. Localization of regions involved in G protein-receptor coupling. J Biol Chem 263(31): 15985-15992

Ohta K, Hayashi H, Mizuguchi H, Kagamiyama H, Fujimoto K, Fukui H (1994) Site-directed mutagenesis of the histamine H1 receptor: roles of aspartic acid107, asparagine198 and threonine194. Biochem Biophys Res Commun 203(2): 1096-1101

Oksenberg D, Havlik S, Peroutka SJ, Ashkenazi A (1995) The third intracellular loop of the 5-hydroxytryptamine2A receptor determines effector coupling specificity. J Neurochem 64(4): 1440-1447

Oksenberg D, Marsters SA, O'Dowd BF, Jin H, Havlik S, Peroutka SJ, Ashkenazi A (1992) A single amino-acid difference confers major pharmacological variation between human and rodent 5-HT1B receptors. Nature 360(6400): 161-163

Olmea O, Valencia A (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 2(3): S25-32

Page KM, Curtis CA, Jones PG, Hulme EC (1995) The functional role of the binding site aspartate in muscarinic acetylcholine receptors, probed by site-directed mutagenesis. Eur J Pharmacol 289(3): 429-437

Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW, Ruepp A, Frishman D (2005) The MIPS mammalian protein-protein interaction database. Bioinformatics 21(6): 832-834

Pages S, Belaich A, Belaich JP, Morag E, Lamed R, Shoham Y, Bayer EA (1997) Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: prediction of specificity determinants of the dockerin domain. Proteins 29(4): 517-527

Pak Y, Pham N, Rotin D (2002) Direct binding of the beta1 adrenergic receptor to the cyclic AMP-dependent guanine nucleotide exchange factor CNrasGEF leads to Ras activation. Mol Cell Biol 22(22): 7942-7952

Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. Science 289(5480): 739-745

Papin J, Subramaniam S (2004) Bioinformatics and cellular signaling. Curr Opin Biotechnol 15(1): 78-81

Papoucheva E, Dumuis A, Sebben M, Richter DW, Ponimaskin EG (2004) The 5-hydroxytryptamine(1A) receptor is stably palmitoylated, and acylation is critical for communication of receptor with Gi protein. J Biol Chem 279(5): 3280-3291

Park J, Lappe M, Teichmann SA (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J Mol Biol 307(3): 929-938

Parker EM, Grisel DA, Iben LG, Shapiro RA (1993) A single amino acid difference accounts for the pharmacological distinctions between the rat and human 5-hydroxytryptamine1B receptors. J Neurochem 60(1): 380-383

Parker LL, Backstrom JR, Sanders-Bush E, Shieh BH (2003) Agonist-induced phosphorylation of the serotonin 5-HT2C receptor regulates its interaction with multiple PDZ protein 1. J Biol Chem 278(24): 21576-21583

Pauling L (1940) A theory of the structure and process of formation of antibodies. J Am Chem Soc 62: 2643-2657

Pauwels PJ, Colpaert FC (2000) Disparate ligand-mediated Ca(2+) responses by wild-type, mutant Ser(200)Ala and Ser(204)Ala alpha(2A)-adrenoceptor: G(alpha15) fusion proteins: evidence for multiple ligand-activation binding sites. Br J Pharmacol 130(7): 1505-1512

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J Mol Biol 271(4): 511-523

Pazos F, Ranea JA, Juan D, Sternberg MJ (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol 352(4): 1002-1015

Pazos F, Valencia A (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein engineering 14(9): 609-614

Pazos F, Valencia A (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins 47(2): 219-227

Pazos F, Valencia A (2008) Protein co-evolution, co-adaptation and interactions. Embo J 27(20): 2648-2655

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85(8): 2444-2448

Pei J, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17(8): 700-712

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96(8): 4285-4288

Peltonen JM, Nyronen T, Wurster S, Pihlavisto M, Hoffren AM, Marjamaki A, Xhaard H, Kanerva L, Savola JM, Johnson MS, Scheinin M (2003) Molecular mechanisms of ligand-receptor interactions in transmembrane domain V of the alpha2A-adrenoceptor. Br J Pharmacol 140(2): 347-358

Pereira L, D'Alessio M, Ramirez F, Lynch JR, Sykes B, Pangilinan T, Bonadio J (1993) Genomic organization of the sequence coding for fibrillin, the defective gene product in Marfan syndrome. Hum Mol Genet 2(7): 961-968

Perez DM, Hwa J, Zhao MM, Porter J (1998) Molecular mechanisms of ligand binding and activation in alpha 1-adrenergic receptors. Adv Pharmacol 42: 398-403

Pettifer SR, Sinnott JR, Attwood TK (2004) UTOPIA-User-Friendly Tools for Operating Informatics Applications. Comp Funct Genomics 5(1): 56-60

Pollock DD, Taylor WR, Goldman N (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 287(1): 187-198

Pollock NJ, Manelli AM, Hutchins CW, Steffey ME, MacKenzie RG, Frail DE (1992) Serine mutations in transmembrane V of the dopamine D1 receptor affect ligand interactions and receptor activation. J Biol Chem 267(25): 17780-17786

Porter JE, Edelmann SE, Waugh DJ, Piascik MT, Perez DM (1998) The agonism and synergistic potentiation of weak partial agonists by triethylamine in alpha 1-adrenergic receptor activation: evidence for a salt bridge as the initiating process. Mol Pharmacol 53(4): 766-771

Porter JE, Hwa J, Perez DM (1996) Activation of the alpha1b-adrenergic receptor is initiated by disruption of an interhelical salt bridge constraint. J Biol Chem 271(45): 28318-28323

Porter JE, Perez DM (1999) Characteristics for a salt-bridge switch mutation of the alpha(1b) adrenergic receptor. Altered pharmacology and rescue of constitutive activity. J Biol Chem 274(49): 34535-34538

Prakash MK (2011) Insights on the Role of (Dis)order from Protein-Protein Interaction Linear Free-Energy Relationships. J Am Chem Soc 133(26): 9976-9979

Preobrazhensky AA, Dragan S, Kawano T, Gavrilin MA, Gulina IV, Chakravarty L, Kolattukudy PE (2000) Monocyte chemotactic protein-1 receptor CCR2B is a glycoprotein that has tyrosine sulfation in a conserved extracellular N-terminal region. J Immunol 165(9): 5295-5303

Price RD, Weiner DM, Chang MS, Sanders-Bush E (2001) RNA editing of the human serotonin 5-HT2C receptor alters receptor-mediated activation of G13 protein. J Biol Chem 276(48): 44663-44668

Pritchard L, Dufton MJ (2000) Do proteins learn to evolve? The Hopfield network as a basis for the understanding of protein evolution. J Theor Biol 202(1): 77-86

Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Seraphin B (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods (San Diego, Calif 24(3): 218-229

Qi Y, Klein-Seetharaman J, Bar-Joseph Z (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. Pacific Symposium on Biocomputing: 531-542

Rabut GE, Konner JA, Kajumo F, Moore JP, Dragic T (1998) Alanine substitutions of polar and nonpolar residues in the amino-terminal domain of CCR5 differently impair entry of macrophage- and dualtropic isolates of human immunodeficiency virus type 1. J Virol 72(4): 3464-3468

Rasmussen SG, Choi HJ, Rosenbaum DM, Kobilka TS, Thian FS, Edwards PC, Burghammer M, Ratnala VR, Sanishvili R, Fischetti RF, Schertler GF, Weis WI, Kobilka BK (2007) Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. Nature 450(7168): 383-387

Read RJ, Brayer GD, Jurasek L, James MN (1984) Critical evaluation of comparative model building of Streptomyces griseus trypsin. Biochemistry 23(26): 6570-6575

Reiland J, Furcht LT, McCarthy JB (1999) CXC-chemokines stimulate invasion and chemotaxis in prostate carcinoma cells through the CXCR2 receptor. Prostate 41(2): 78-88

Reinhardt DP, Keene DR, Corson GM, Poschl E, Bachinger HP, Gambee JE, Sakai LY (1996) Fibrillin-1: organization in microfibrils and structural properties. J Mol Biol 258(1): 104-116

Rey M, Vicente-Manzanares M, Viedma F, Yanez-Mo M, Urzainqui A, Barreiro O, Vazquez J, Sanchez-Madrid F (2002) Cutting edge: association of the motor protein nonmuscle myosin heavy chain-IIA with the C terminus of the chemokine receptor CXCR4 in T lymphocytes. J Immunol 169(10): 5410-5414

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol 17(10): 1030-1032

Rogov SI, Nekrasov AN (2001) A numerical measure of amino acid residues similarity based on the analysis of their surroundings in natural protein sequences. Protein engineering 14(7): 459-463

Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE (2011) The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 39(Database issue): D392-401

Ross TM, Bieniasz PD, Cullen BR (1998) Multiple residues contribute to the inability of murine CCR-5 to function as a coreceptor for macrophage-tropic human immunodeficiency virus type 1 isolates. J Virol 72(3): 1918-1924

Roth BL, Choudhary MS, Craigo S (1993) Mutagenesis of 5-HT2 serotonin receptors: what does an analysis of many mutant receptors tell us? Med Chem Res 3: 297-305

Roth BL, Shoham M, Choudhary MS, Khan N (1997) Identification of conserved aromatic residues essential for agonist binding and second messenger production at 5-hydroxytryptamine2A receptors. Mol Pharmacol 52(2): 259-266

Rudling JE, Kennedy K, Evans PD (1999) The effect of site-directed mutagenesis of two transmembrane serine residues on agonist-specific coupling of a cloned human alpha2A-adrenoceptor to adenylyl cyclase. Br J Pharmacol 127(4): 877-886

Sakai LY, Keene DR, Glanville RW, Bachinger HP (1991) Purification and partial characterization of fibrillin, a cysteine-rich structural component of connective tissue microfibrils. J Biol Chem 266(22): 14763-14770

Salwinski L, Eisenberg D (2003) Computational methods of analysis of protein-protein interactions. Curr Opin Struct Biol 13(3): 377-382

Samson M, LaRosa G, Libert F, Paindavoine P, Detheux M, Vassart G, Parmentier M (1997) The second extracellular loop of CCR5 is the major determinant of ligand specificity. J Biol Chem 272(40): 24934-24941

Sandhu KS (2009) Intrinsic disorder explains diverse nuclear roles of chromatin remodeling proteins. J Mol Recognit 22(1): 1-8

Sartania N, Strange PG (1999) Role of conserved serine residues in the interaction of agonists with D3 dopamine receptors. J Neurochem 72(6): 2621-2624

Sato T, Kobayashi H, Nagao T, Kurose H (1999) Ser203 as well as Ser204 and Ser207 in fifth transmembrane domain of the human beta2-adrenoceptor contributes to agonist binding and receptor activation. Br J Pharmacol 128(2): 272-274

Sato T, Yamanishi Y, Horimoto K, Kanehisa M, Toh H (2006) Partial correlation coefficient between distance matrices as a new indicator of protein-protein interactions. Bioinformatics 22(20): 2488-2492

Sato T, Yamanishi Y, Kanehisa M, Toh H (2005) The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21(17): 3482-3489

Savarese TM, Wang CD, Fraser CM (1992) Site-directed mutagenesis of the rat m1 muscarinic acetylcholine receptor. Role of conserved cysteines in receptor function. J Biol Chem 267(16): 11439-11448

Scapin G (2006) Structural biology and drug discovery. Curr Pharm Des 12(17): 2087-2097

Schetz JA, Benjamin PS, Sibley DR (2000) Nonconserved residues in the second transmembrane-spanning domain of the D(4) dopamine receptor are molecular determinants of D(4)-selective pharmacology. Mol Pharmacol 57(1): 144-152

Schwartz T, Gether U, Schambye H, Hjorth S (1995) Molecular mechanism of action of non-peptide ligands for peptide receptors. Current Pharmaceutical Design 1: 325-342

Scordis P, Flower DR, Attwood TK (1999) FingerPRINTScan: intelligent searching of the PRINTS motif database. Bioinformatics 15(10): 799-806

Shackelford G, Karplus K (2007) Contact prediction using mutual information and neural nets. Proteins 69 Suppl 8: 159-164

Shapiro DA, Kristiansen K, Kroeze WK, Roth BL (2000) Differential modes of agonist binding to 5-hydroxytryptamine(2A) serotonin receptors revealed by mutation and molecular modeling of conserved residues in transmembrane region 5. Mol Pharmacol 58(5): 877-886

Shin N, Coates E, Murgolo NJ, Morse KL, Bayne M, Strader CD, Monsma FJ, Jr. (2002) Molecular modeling and site-specific mutagenesis of the histamine-binding site of the histamine H4 receptor. Mol Pharmacol 62(1): 38-47

Siciliano SJ, Kuhmann SE, Weng Y, Madani N, Springer MS, Lineberger JE, Danzeisen R, Miller MD, Kavanaugh MP, DeMartino JA, Kabat D (1999) A critical site in the core of the CCR5 chemokine receptor required for binding and infectivity of human immunodeficiency virus type 1. J Biol Chem 274(4): 1905-1913

Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38(Database issue): D161-166

Simpson MM, Ballesteros JA, Chiappa V, Chen J, Suehiro M, Hartman DS, Godel T, Snyder LA, Sakmar TP, Javitch JA (1999) Dopamine D4/D2 receptor selectivity is determined by A divergent aromatic

microdomain contained within the second, third, and seventh membrane-spanning segments. Mol Pharmacol 56(6): 1116-1126

Singh S, Nannuru KC, Sadanandam A, Varney ML, Singh RK (2009) CXCR1 and CXCR2 enhances human melanoma tumourigenesis, growth and invasion. Br J Cancer 100(10): 1638-1646

Skelton NJ, Quan C, Reilly D, Lowman H (1999) Structure of a CXC chemokine-receptor fragment in complex with interleukin-8. Structure 7(2): 157-168

Smith FD, Oxford GS, Milgram SL (1999) Association of the D2 dopamine receptor third cytoplasmic loop with spinophilin, a protein phosphatase-1-interacting protein. J Biol Chem 274(28): 19894-19900

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1): 195-197

Spalding TA, Birdsall NJ, Curtis CA, Hulme EC (1994) Acetylcholine mustard labels the binding site aspartate in muscarinic acetylcholine receptors. J Biol Chem 269(6): 4092-4097

Spalding TA, Burstein ES, Brauner-Osborne H, Hill-Eubanks D, Brann MR (1995) Pharmacology of a constitutively active muscarinic receptor generated by random mutagenesis. J Pharmacol Exp Ther 275(3): 1274-1279

Sprinzak E, Margalit H (2001) Correlated sequence-signatures as markers of protein-protein interaction. J Mol Biol 311(4): 681-692

Stanasila L, Perez JB, Vogel H, Cotecchia S (2003) Oligomerization of the alpha 1a- and alpha 1b-adrenergic receptor subtypes. Potential implications in receptor internalization. J Biol Chem 278(41): 40239-40251

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34(Database issue): D535-539

Strader CD, Candelore MR, Hill WS, Sigal IS, Dixon RA (1989) Identification of two serine residues involved in agonist activation of the beta-adrenergic receptor. J Biol Chem 264(23): 13572-13578

Strader CD, Dixon RA (1991) Genetic analysis of the beta-adrenergic receptor. Adv Exp Med Biol 287: 209-220

Strader CD, Sigal IS, Blake AD, Cheung AH, Register RB, Rands E, Zemcik BA, Candelore MR, Dixon RA (1987) The carboxyl terminus of the hamster beta-adrenergic receptor expressed in mouse L cells is not required for receptor sequestration. Cell 49(6): 855-863

Strader CD, Sigal IS, Candelore MR, Rands E, Hill WS, Dixon RA (1988) Conserved aspartic acid residues 79 and 113 of the beta-adrenergic receptor have different roles in receptor function. J Biol Chem 263(21): 10267-10271

Sugimoto Y, Fujisawa R, Tanimura R, Lattion AL, Cotecchia S, Tsujimoto G, Nagao T, Kurose H (2002) Beta(1)-selective agonist (-)-1-(3,4-dimethoxyphenetylamino)-3-(3,4-dihydroxy)-2-propanol [(-)-RO363] differentially interacts with key amino acids responsible for beta(1)-selective binding in resting and active states. J Pharmacol Exp Ther 301(1): 51-58

Suryanarayana S, Daunt DA, Von Zastrow M, Kobilka BK (1991) A point mutation in the seventh hydrophobic domain of the alpha 2 adrenergic receptor increases its affinity for a family of beta receptor antagonists. J Biol Chem 266(23): 15488-15492

Suryanarayana S, Kobilka BK (1993) Amino acid substitutions at position 312 in the seventh hydrophobic segment of the beta 2-adrenergic receptor modify ligand-binding specificity. Mol Pharmacol 44(1): 111-114

Tamames J, Casari G, Ouzounis C, Valencia A (1997) Conserved clusters of functionally related genes in two bacterial genomes. Journal of molecular evolution 44(1): 66-73

Tan SH, Zhang Z, Ng SK (2004) ADVICE: Automated Detection and Validation of Interaction by Co-Evolution. Nucleic Acids Res 32(Web Server issue): W69-72

Teichmann SA, Babu MM (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. Trends in biotechnology 20(10): 407-410; discussion 410

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22): 4673-4680

Tomic M, Seeman P, George SR, O'Dowd BF (1993) Dopamine D1 receptor mutagenesis: role of amino acids in agonist and antagonist binding. Biochem Biophys Res Commun 191(3): 1020-1027

Tsai CJ, Xu D, Nussinov R (1998) Protein folding via binding and vice versa. Fold Des 3(4): R71-80

Tsai J, Gerstein M (2002) Calculations of protein volumes: sensitivity analysis and parameter database. Bioinformatics 18(7): 985-995

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, 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 JM (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770): 623-627

Valiquette M, Bonin H, Bouvier M (1993) Mutation of tyrosine-350 impairs the coupling of the beta 2-adrenergic receptor to the stimulatory guanine nucleotide binding protein without interfering with receptor down-regulation. Biochemistry 32(19): 4979-4985

Vogel WK, Peterson GL, Broderick DJ, Mosser VA, Schimerlik MI (1999) Double mutant cycle analysis of aspartate 69, 97, and 103 to asparagine mutants in the m2 muscarinic acetylcholine receptor. Arch Biochem Biophys 361(2): 283-294

Vogel WK, Sheehan DM, Schimerlik MI (1997) Site-directed mutagenesis on the m2 muscarinic acetylcholine receptor: the significance of Tyr403 in the binding of agonists and functional coupling. Mol Pharmacol 52(6): 1087-1094

Volkel P, Le Faou P, Angrand PO (2010) Interaction proteomics: characterization of protein complexes using tandem affinity purification-mass spectrometry. Biochem Soc Trans 38(4): 883-887

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417(6887): 399-403

Wade SM, Lim WK, Lan KL, Chung DA, Nanamori M, Neubig RR (1999) G(i) activator region of alpha(2A)-adrenergic receptors: distinct basic residues mediate G(i) versus G(s) activation. Mol Pharmacol 56(5): 1005-1013

Wang CD, Buck MA, Fraser CM (1991) Site-directed mutagenesis of alpha 2A-adrenergic receptors: identification of amino acids involved in ligand binding and receptor activation by agonists. Mol Pharmacol 40(2): 168-179

Wang CD, Gallaher TK, Shih JC (1993) Site-directed mutagenesis of the serotonin 5-hydroxytrypamine2 receptor: identification of amino acids necessary for ligand binding and receptor activation. Mol Pharmacol 43(6): 931-940

Wang Q, Limbird LE (2002) Regulated interactions of the alpha 2A adrenergic receptor with spinophilin, 14-3-3zeta, and arrestin 3. J Biol Chem 277(52): 50589-50596

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337(3): 635-645

Ward SD, Curtis CA, Hulme EC (1999) Alanine-scanning mutagenesis of transmembrane domain 6 of the M(1) muscarinic acetylcholine receptor suggests that Tyr381 plays key roles in receptor function. Mol Pharmacol 56(5): 1031-1041

Waugh DJ, Gaivin RJ, Zuscik MJ, Gonzalez-Cabrera P, Ross SA, Yun J, Perez DM (2001) Phe-308 and Phe-312 in transmembrane domain 7 are major sites of alpha 1-adrenergic receptor antagonist binding. Imidazoline agonists bind like antagonists. J Biol Chem 276(27): 25366-25371

Waugh DJ, Zhao MM, Zuscik MJ, Perez DM (2000) Novel aromatic residues in transmembrane domains IV and V involved in agonist binding at alpha(1a)-adrenergic receptors. J Biol Chem 275(16): 11698-11705

Weinrich D, Jonkheijm P, Niemeyer CM, Waldmann H (2009) Applications of protein biochips in biomedical and biotechnological research. Angew Chem Int Ed Engl 48(42): 7744-7751

Wess J, Blin N, Mutschler E, Bluml K (1995) Muscarinic acetylcholine receptors: structural basis of ligand binding and G protein coupling. Life Sci 56(11-12): 915-922

Wess J, Gdula D, Brann MR (1991) Site-directed mutagenesis of the m3 muscarinic receptor: identification of a series of threonine and tyrosine residues involved in agonist but not antagonist binding. Embo J 10(12): 3729-3734

Wess J, Maggio R, Palmer JR, Vogel Z (1992) Role of conserved threonine and tyrosine residues in acetylcholine binding and muscarinic receptor activation. A study with m3 muscarinic receptor point mutants. J Biol Chem 267(27): 19313-19319

Wess J, Nanavati S, Vogel Z, Maggio R (1993) Functional role of proline and tryptophan residues highly conserved among G protein-coupled receptors studied by mutational analysis of the m3 muscarinic receptor. Embo J 12(1): 331-338

Wieland K, Laak AM, Smit MJ, Kuhne R, Timmerman H, Leurs R (1999) Mutational analysis of the antagonist-binding site of the histamine H(1) receptor. J Biol Chem 274(42): 29994-30000

Wieland K, Zuurmond HM, Krasel C, Ijzerman AP, Lohse MJ (1996) Involvement of Asn-293 in stereospecific agonist recognition and in activation of the beta 2-adrenergic receptor. Proc Natl Acad Sci U S A 93(17): 9276-9281

Wilcox RE, Huang WH, Brusniak MY, Wilcox DM, Pearlman RS, Teeter MM, DuRand CJ, Wiens BL, Neve KA (2000) CoMFA-based prediction of agonist affinities at recombinant wild type versus serine to alanine point mutated D2 dopamine receptors. J Med Chem 43(16): 3005-3019

Wong SK, Parker EM, Ross EM (1990) Chimeric muscarinic cholinergic: beta-adrenergic receptors that activate Gs in response to muscarinic agonists. J Biol Chem 265(11): 6219-6224

Wong SK, Slaughter C, Ruoho AE, Ross EM (1988) The catecholamine binding site of the beta-adrenergic receptor is formed by juxtaposed membrane-spanning domains. J Biol Chem 263(17): 7925-7928

Woods AS, Ciruela F, Fuxe K, Agnati LF, Lluis C, Franco R, Ferre S (2005) Role of electrostatic interaction in receptor-receptor heteromerization. J Mol Neurosci 26(2-3): 125-132

Woods AS, Marcellino D, Jackson SN, Franco R, Ferre S, Agnati LF, Fuxe K (2008) How calmodulin interacts with the adenosine A(2A) and the dopamine D(2) receptors. J Proteome Res 7(8): 3428-3434

Woodward R, Coley C, Daniell S, Naylor LH, Strange PG (1996) Investigation of the role of conserved serine residues in the long form of the rat D2 dopamine receptor using site-directed mutagenesis. J Neurochem 66(1): 394-402

Woodward R, Daniell SJ, Strange PG, Naylor LH (1994) Structural studies on D2 dopamine receptors: mutation of a histidine residue specifically affects the binding of a subgroup of substituted benzamide drugs. J Neurochem 62(5): 1664-1669

Wright W, Scordis P, Attwood TK (1999) BLAST PRINTS--alternative perspectives on sequence similarity. Bioinformatics 15(6): 523-524

Wu B, Chien EY, Mol CD, Fenalti G, Liu W, Katritch V, Abagyan R, Brooun A, Wells P, Bi FC, Hamel DJ, Kuhn P, Handel TM, Cherezov V, Stevens RC (2010) Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. Science 330(6007): 1066-1071

Wu G, Bogatkevich GS, Mukhin YV, Benovic JL, Hildebrandt JD, Lanier SM (2000) Identification of Gbetagamma binding sites in the third intracellular loop of the M(3)-muscarinic receptor and their role in receptor regulation. J Biol Chem 275(12): 9026-9034

Wu L, LaRosa G, Kassam N, Gordon CJ, Heath H, Ruffing N, Chen H, Humblias J, Samson M, Parmentier M, Moore JP, Mackay CR (1997) Interaction of chemokine receptor CCR5 with its ligands: multiple domains for HIV-1 gp120 binding and a single domain for chemokine binding. J Exp Med 186(8): 1373-1381

Wurch T, Boutet-Robinet EA, Palmier C, Colpaert FC, Pauwels PJ (2003) Constitutive coupling of a chimeric dopamine D2/alpha 1B receptor to the phospholipase C pathway: inverse agonism to silent antagonism by neuroleptic drugs. J Pharmacol Exp Ther 304(1): 380-390

Xanthou G, Williams TJ, Pease JE (2003) Molecular characterization of the chemokine receptor CXCR3: evidence for the involvement of distinct extracellular domains in a multi-step model of ligand binding and receptor activation. Eur J Immunol 33(10): 2927-2936

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D (2000) DIP: the database of interacting proteins. Nucleic Acids Res 28(1): 289-291

Xie W, Jiang H, Wu Y, Wu D (1997) Two basic amino acids in the second inner loop of the interleukin-8 receptor are essential for Galpha16 coupling. J Biol Chem 272(40): 24948-24951

Xu J, He J, Castleberry AM, Balasubramanian S, Lau AG, Hall RA (2003) Heterodimerization of alpha 2A- and beta 1-adrenergic receptors. J Biol Chem 278(12): 10770-10777

Xu J, Paquet M, Lau AG, Wood JD, Ross CA, Hall RA (2001) beta 1-adrenergic receptor association with the synaptic scaffolding protein membrane-associated guanylate kinase inverted-2 (MAGI-2). Differential regulation of receptor internalization by MAGI-2 and PSD-95. J Biol Chem 276(44): 41310-41317

Yang G, Rosen DG, Liu G, Yang F, Guo X, Xiao X, Xue F, Mercado-Uribe I, Huang J, Lin SH, Mills GB, Liu J (2010) CXCR2 promotes ovarian cancer growth through dysregulated cell cycle, diminished apoptosis, and enhanced angiogenesis. Clin Cancer Res 16(15): 3875-3886

Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. Mol Syst Biol 1: 2005 0026

Yeang CH, Haussler D (2007) Detecting coevolution in and among protein domains. PLoS Comput Biol 3(11): e211

Yip KY, Patel P, Kim PM, Engelman DM, McDermott D, Gerstein M (2008) An integrated system for studying residue coevolution in proteins. Bioinformatics 24(2): 290-292

Youn BS, Yu KY, Alkhatib G, Kwon BS (2001) The seventh transmembrane domain of cc chemokine receptor 5 is critical for MIP-1beta binding and receptor activation: role of MET 287. Biochem Biophys Res Commun 281(3): 627-633

Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet AS, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabasi AL, Tavernier J, Hill DE, Vidal M (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322(5898): 104-110

Yu X, Schneiderhan-Marra N, Joos TO (2010) Protein microarrays for personalized medicine. Clin Chem 56(3): 376-387

Zeng FY, Wess J (1999) Identification and molecular characterization of m3 muscarinic receptor dimers. J Biol Chem 274(27): 19487-19497

Zhang T, Xu Q, Chen FR, Han QD, Zhang YY (2004) Yeast two-hybrid screening for proteins that interact with alpha1-adrenergic receptors. Acta Pharmacol Sin 25(11): 1471-1478

Zhou N, Luo Z, Luo J, Liu D, Hall JW, Pomerantz RJ, Huang Z (2001) Structural and functional characterization of human CXCR4 as a chemokine receptor and HIV-1 co-receptor by mutagenesis and molecular modeling studies. J Biol Chem 276(46): 42826-42833

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 RA, Gerstein M, Snyder M (2001) Global analysis of protein activities using proteome chips. Science 293(5537): 2101-2105

Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, Klemic KG, Smith D, Gerstein M, Reed MA, Snyder M (2000) Analysis of yeast protein kinases using protein chips. Nat Genet 26(3): 283-289

Zoffmann S, Chollet A, Galzi JL (2002) Identification of the extracellular loop 2 as the point of interaction between the N terminus of the chemokine MIP-1alpha and its CCR1 receptor. Mol Pharmacol 62(3): 729-736

Zuscik MJ, Porter JE, Gaivin R, Perez DM (1998) Identification of a conserved switch residue responsible for selective constitutive activation of the beta2-adrenergic receptor. J Biol Chem 273(6): 3401-3407

Zuurmond HM, Hessling J, Bluml K, Lohse M, Ijzerman AP (1999) Study of interaction between agonists and asn293 in helix VI of human beta(2)-adrenergic receptor. Mol Pharmacol 56(5): 909-916

Appendix 1. Benchmarking Study Data

List of benchmarking study proteins for various datasets.

Pazos+ (Validation method: DIP database)

ACKA_ECOLI	ATP6_ECOLI	ATPA_ECOLI	ATPB_ECOLI	ATPD_ECOLI	ATPE_ECOLI
ATPF_ECOLI	ATPG_ECOLI	ATPL_ECOLI	BGLG_ECOLI	BIOF_ECOLI	CARA_ECOLI
CARB_ECOLI	CCME_ECOLI	CCMF_ECOLI	CCMH_ECOLI	CH10_ECOLI	CH60_ECOLI
CHEA_ECOLI	CHER_ECOLI	CHEY_ECOLI	CLPA_ECOLI	CLPP_ECOLI	CLPS_ECOLI
CLPX_ECOLI	CRP_ECOLI	CYSD_ECOLI	CYSN_ECOLI	CYTR_ECOLI	DADA_ECOLI
DHSA_ECOLI	DHSB_ECOLI	DHSC_ECOLI	DLDH_ECOLI	DNAA_ECOLI	DNAB_ECOLI
DNAC_ECOLI	DNAJ_ECOLI	DNAK_ECOLI	DSBC_ECOLI	DSBD_ECOLI	ENO_ECOLI
ENVZ_ECOLI	ERA_ECOLI	EX1_ECOLI	EXBB_ECOLI	EXBD_ECOLI	FER_ECOLI
FIS_ECOLI	FLIF_ECOLI	FLIG_ECOLI	FLIM_ECOLI	FLIN_ECOLI	FTNA_ECOLI
FTSA_ECOLI	FTSQ_ECOLI	FTSZ_ECOLI	G3P1_ECOLI	GATY_ECOLI	GLF_ECOLI
GLNB_ECOLI	GLPK_ECOLI	GRPE_ECOLI	GYRA_ECOLI	HSCA_ECOLI	HSCB_ECOLI
HSLU_ECOLI	HSLV_ECOLI	IHFA_ECOLI	IHFB_ECOLI	ILVI_ECOLI	ISCS_ECOLI
LEU1_ECOLI	LEXA_ECOLI	LIPA_ECOLI	LOLD_ECOLI	MALE_ECOLI	MALG_ECOLI
MAZG_ECOLI	MCP2_ECOLI	METE_ECOLI	METF_ECOLI	METK_ECOLI	MINC_ECOLI
MIND_ECOLI	MOBA_ECOLI	MOBB_ECOLI	MOEA_ECOLI	MOG_ECOLI	MUTL_ECOLI
MUTS_ECOLI	NADA_ECOLI	NADB_ECOLI	NAGD_ECOLI	NIFU_ECOLI	NTRB_ECOLI
NUSA_ECOLI	NUSB_ECOLI	NUSG_ECOLI	OMPR_ECOLI	OXAA_ECOLI	PABA_ECOLI
PABB_ECOLI	PAL_ECOLI	PNP_ECOLI	PNTA_ECOLI	PNTB_ECOLI	PRIM_ECOLI
PSTB_ECOLI	PT1_ECOLI	PTA_ECOLI	PTGA_ECOLI	PTHP_ECOLI	PUR7_ECOLI
PYRB_ECOLI	RECA_ECOLI	RECF_ECOLI	RECO_ECOLI	RECR_ECOLI	RHO_ECOLI
RIR1_ECOLI	RIR2_ECOLI	RL34_ECOLI	RL7_ECOLI	RL9_ECOLI	RNE_ECOLI
RP32_ECOLI	RPOA_ECOLI	RPOB_ECOLI	RPOC_ECOLI	RPOD_ECOLI	RPOZ_ECOLI
RS10_ECOLI	RS2_ECOLI	RUVA_ECOLI	RUVB_ECOLI	SBCC_ECOLI	SBCD_ECOLI
SECA_ECOLI	SECB_ECOLI	SECE_ECOLI	SECG_ECOLI	SECY_ECOLI	SPEA_ECOLI
SSB_ECOLI	SSPB_ECOLI	SUCC_ECOLI	SUCD_ECOLI	SYFA_ECOLI	SYFB_ECOLI
SYGB_ECOLI	SYT_ECOLI	TATB_ECOLI	TATC_ECOLI	TDH_ECOLI	THID_ECOLI
THIG_ECOLI	THIO_ECOLI	TOLB_ECOLI	TRPA_ECOLI	TRPB_ECOLI	TRXB_ECOLI
UPP_ECOLI	UVRD_ECOLI				

Pazos- (Validation method: random selection of two proteins from the Pazos+ set)

ACKA_ECOLI	ACRA_ECOLI	ACRB_ECOLI	AES_ECOLI	ARCB_ECOLI	ATP6_ECOLI
ATPA_ECOLI	ATPD_ECOLI	ATPE_ECOLI	ATPF_ECOLI	ATPG_ECOLI	ATPL_ECOLI
BARA_ECOLI	BGLG_ECOLI	BIOF_ECOLI	BTUB_ECOLI	BTUC_ECOLI	CARA_ECOLI
CARB_ECOLI	CCME_ECOLI	CCMF_ECOLI	CCMH_ECOLI	CH10_ECOLI	CH60_ECOLI
CHEA_ECOLI	CHER_ECOLI	CLPA_ECOLI	CLPP_ECOLI	CLPS_ECOLI	CLPX_ECOLI
CRP_ECOLI	CYSD_ECOLI	DADA_ECOLI	DBHA_ECOLI	DBHB_ECOLI	DHSA_ECOLI
DHSB_ECOLI	DHSC_ECOLI	DLDH_ECOLI	DNAA_ECOLI	DNAC_ECOLI	DNAJ_ECOLI
DNAK_ECOLI	DSBD_ECOLI	EFTS_ECOLI	ENO_ECOLI	ENVZ_ECOLI	ERA_ECOLI
EX1_ECOLI	EXBB_ECOLI	EXBD_ECOLI	FER_ECOLI	FHUA_ECOLI	FIS_ECOLI
FLIF_ECOLI	FLIN_ECOLI	FTNA_ECOLI	FTSA_ECOLI	FTSZ_ECOLI	GALU_ECOLI
GATY_ECOLI	GCVA_ECOLI	GLNB_ECOLI	GLPK_ECOLI	GRPE_ECOLI	GYRA_ECOLI
HSLU_ECOLI	HSLV_ECOLI	IHFB_ECOLI	ILVI_ECOLI	ISCS_ECOLI	LEU1_ECOLI
LEXA_ECOLI	LOLD_ECOLI	MALE_ECOLI	MALF_ECOLI	MALG_ECOLI	MALY_ECOLI
MAZG_ECOLI	METE_ECOLI	METF_ECOLI	METK_ECOLI	MINC_ECOLI	MIND_ECOLI
MOAE_ECOLI	MOBA_ECOLI	MOEA_ECOLI	MOEB_ECOLI	MOG_ECOLI	MUTL_ECOLI
MUTS_ECOLI	NADB_ECOLI	NIFU_ECOLI	NTRB_ECOLI	NUSA_ECOLI	NUSB_ECOLI

NUSG_ECOLI	OMPA_ECOLI	OMPR_ECOLI	OXAA_ECOLI	PABA_ECOLI	PABB_ECOLI
PNP_ECOLI	PNTA_ECOLI	PNTB_ECOLI	PRIM_ECOLI	PSTB_ECOLI	PT1_ECOLI
PTA_ECOLI	PTGA_ECOLI	PTHP_ECOLI	PUR7_ECOLI	PYRB_ECOLI	RECA_ECOLI
RECF_ECOLI	RECO_ECOLI	RECR_ECOLI	RHO_ECOLI	RIR1_ECOLI	RIR2_ECOLI
RIR3_ECOLI	RL34_ECOLI	RL7_ECOLI	RL9_ECOLI	RNE_ECOLI	RP32_ECOLI
RPOA_ECOLI	RPOB_ECOLI	RPOC_ECOLI	RPOE_ECOLI	RPOZ_ECOLI	RS10_ECOLI
RS20_ECOLI	RS2_ECOLI	RUVB_ECOLI	SBCC_ECOLI	SBCD_ECOLI	SECB_ECOLI
SECE_ECOLI	SECG_ECOLI	SECY_ECOLI	SOPA_ECOLI	SPEA_ECOLI	SSB_ECOLI
SUCC_ECOLI	SUCD_ECOLI	SYFA_ECOLI	SYFB_ECOLI	SYGB_ECOLI	TDH_ECOLI
THID_ECOLI	THIG_ECOLI	TOLQ_ECOLI	TRPA_ECOLI	TRPB_ECOLI	UPP_ECOLI
YBGF_ECOLI					

Tan+ (Validation method: minimum three experimental results)

ADA2_YEAST	AP1B1_YEAST	AP1G1_YEAST	AP1M1_YEAST	ARO1_YEAST	ARPC1_YEAST
ARPC3_YEAST	ARX1_YEAST	ATPG_YEAST	ATPO_YEAST	BRX1_YEAST	CALM_YEAST
CARB_YEAST	CDC11_YEAST	CDC12_YEAST	CDC28_YEAST	CDC53_YEAST	CG23_YEAST
CKS1_YEAST	COAC_YEAST	CSK22_YEAST	CSK2B_YEAST	CSK2C_YEAST	DIP2_YEAST
DRS1_YEAST	EI2BA_YEAST	EI2BB_YEAST	EI2BD_YEAST	ERB1_YEAST	FAS1_YEAST
FAS2_YEAST	FBRL_YEAST	GAL83_YEAST	GCN5_YEAST	HAS1_YEAST	IF32_YEAST
IF34_YEAST	IF38_YEAST	IF4E_YEAST	IF4F1_YEAST	KAPB_YEAST	KAPR_YEAST
LSM2_YEAST	LSM7_YEAST	MYO2_YEAST	NOG1_YEAST	NOP12_YEAST	NOP4_YEAST
NOP58_YEAST	NUG1_YEAST	ODO1_YEAST	ODO2_YEAST	ODP2_YEAST	ODPA_YEAST
ODPB_YEAST	ODPX_YEAST	PRI1_YEAST	PRI2_YEAST	PRP3_YEAST	PRP4_YEAST
PRS10_YEAST	PRS6A_YEAST	PRS6B_YEAST	PSB3_YEAST	PSB7_YEAST	PSDA_YEAST
PUF6_YEAST	PYC1_YEAST	PYC2_YEAST	RFC2_YEAST	RFC3_YEAST	RFC4_YEAST
RFC5_YEAST	RLP24_YEAST	RPA1_YEAST	RPA2_YEAST	RPB1_YEAST	RPB3_YEAST
RPB5_YEAST	RPB6_YEAST	RPC19_YEAST	RPC1_YEAST	RPC2_YEAST	RPC5_YEAST
RPN10_YEAST	RPN11_YEAST	RPN12_YEAST	RPN3_YEAST	RPN7_YEAST	RPN8_YEAST
RRP1_YEAST	RRP41_YEAST	RRP45_YEAST	RT05_YEAST	RT09_YEAST	SEC13_YEAST
SIK1_YEAST	SIP2_YEAST	SKP1_YEAST	SNF1_YEAST	SNF4_YEAST	SPT5_YEAST
SUI1_YEAST	SYFA_YEAST	SYFB_YEAST	TPS1_YEAST	TPS2_YEAST	TRPE_YEAST
TRPG_YEAST	UTP13_YEAST	VA0D_YEAST	VATA_YEAST	VATB_YEAST	VATD_YEAST
VATE_YEAST	VATF_YEAST	VPH1_YEAST	WEB1_YEAST	YCF9_YEAST	YEV6_YEAST
YG3J_YEAST	YL409_YEAST	YNL0_YEAST	YNN2_YEAST		

Hakes+ (Validation method: crystal structures from the PQS database)

6PGD1_YEAST	ACE2_YEAST	ACS2_YEAST	ACT_YEAST	AP1B1_YEAST	AP1G1_YEAST
ARP2_YEAST	ARP3_YEAST	BDF1_YEAST	BFR2_YEAST	BMH2_YEAST	BRE1_YEAST
BRO1_YEAST	BRX1_YEAST	BUR1_YEAST	C1TC_YEAST	C1TM_YEAST	CAPZA_YEAST
CBF5_YEAST	CBS_YEAST	CCR4_YEAST	CDC16_YEAST	CDC31_YEAST	CDC53_YEAST
CDC6_YEAST	CDC7_YEAST	CDH1_YEAST	CFT2_YEAST	CG22_YEAST	CHD1_YEAST
CNS1_YEAST	COAC_YEAST	COFI_YEAST	COPB2_YEAST	COPB_YEAST	CORO_YEAST
CSK21_YEAST	CSK22_YEAST	CSK2B_YEAST	DBP10_YEAST	DBP4_YEAST	DBP8_YEAST
DED1_YEAST	DLDH_YEAST	DPH1_YEAST	DPOA_YEAST	DPOE_YEAST	DRS1_YEAST
DUN1_YEAST	DUR1_YEAST	EAF3_YEAST	EF1A_YEAST	EF1G1_YEAST	EF1G2_YEAST
EFTU_YEAST	EI2BA_YEAST	EMP24_YEAST	EPL1_YEAST	ERF2_YEAST	ERG25_YEAST
ERG27_YEAST	ERV25_YEAST	ESA1_YEAST	FIP1_YEAST	FRDA_YEAST	GAL1_YEAST
GAR1_YEAST	GAT1_YEAST	GCN20_YEAST	GCN2_YEAST	GCN5_YEAST	GDI1_YEAST
GET3_YEAST	GFA1_YEAST	GLYC_YEAST	H2A2_YEAST	H3_YEAST	H4_YEAST
HAS1_YEAST	HAT2_YEAST	HDA1_YEAST	HEMH_YEAST	HRR25_YEAST	HS104_YEAST
HSC82_YEAST	HSP71_YEAST	HSP77_YEAST	HSP7F_YEAST	HSP82_YEAST	IDH2_YEAST
IF2A_YEAST	IF2G_YEAST	IF4A_YEAST	IF4E_YEAST	IF5_YEAST	IF6_YEAST
IMA1_YEAST	IMDH4_YEAST	IST3_YEAST	ISW1_YEAST	KIN28_YEAST	KKK1_YEAST
KOK0_YEAST	KPR3_YEAST	LAH1_YEAST	LCB1_YEAST	LSM5_YEAST	MAK5_YEAST

MBF1_YEAST	MCM6_YEAST	MDHM_YEAST	MDJ1_YEAST	METL_YEAST	MIA40_YEAST
MMS2_YEAST	MOT1_YEAST	MPG1_YEAST	MPIP_YEAST	MRT4_YEAST	MSH2_YEAST
MSH5_YEAST	MSH6_YEAST	MTR4_YEAST	MYO1_YEAST	NCB5R_YEAST	NCS1_YEAST
NDC80_YEAST	NHP2_YEAST	NOG1_YEAST	NOG2_YEAST	NOP12_YEAST	NOP14_YEAST
NOP4_YEAST	NOP58_YEAST	NUF2_YEAST	NUG1_YEAST	ODO1_YEAST	ODPB_YEAST
OTC_YEAST	PABP_YEAST	PCNA_YEAST	PFS2_YEAST	PHR_YEAST	PHSG_YEAST
PIK1_YEAST	POB3_YEAST	PP11_YEAST	PP12_YEAST	PP2A1_YEAST	PP2A2_YEAST
PRP45_YEAST	PRP4_YEAST	PRS10_YEAST	PRS7_YEAST	PRS8_YEAST	PSA1_YEAST
PSA2_YEAST	PSA4_YEAST	PSA5_YEAST	PSA6_YEAST	PSA7_YEAST	PSB2_YEAST
PSB3_YEAST	PSB4_YEAST	PSB6_YEAST	PSB7_YEAST	PSF2_YEAST	PUB1_YEAST
PUR2_YEAST	PURA_YEAST	PWP1_YEAST	PYR1_YEAST	RAD16_YEAST	RAD50_YEAST
RAD51_YEAST	RFC2_YEAST	RFC3_YEAST	RFC4_YEAST	RFC5_YEAST	RIR1_YEAST
RIX7_YEAST	RL10_YEAST	RL13A_YEAST	RL23_YEAST	RL28_YEAST	RL30_YEAST
RL7A_YEAST	RL8B_YEAST	RLA0_YEAST	RLP24_YEAST	RM08_YEAST	RM09_YEAST
RNA14_YEAST	ROK1_YEAST	RPB10_YEAST	RPB2_YEAST	RPB3_YEAST	RPB5_YEAST
RPC19_YEAST	RPC2_YEAST	RPF1_YEAST	RPN11_YEAST	RPN6_YEAST	RRP44_YEAST
RRP45_YEAST	RRP4_YEAST	RS0A_YEAST	RS3B_YEAST	RS5_YEAST	RSMB_YEAST
RT04_YEAST	RT05_YEAST	RT09_YEAST	RT16_YEAST	RU2A_YEAST	RUVB1_YEAST
SAP1_YEAST	SAR1_YEAST	SEC4_YEAST	SEN1_YEAST	SEN2_YEAST	SERA_YEAST
SGN1_YEAST	SKI2_YEAST	SLT2_YEAST	SMC3_YEAST	SMT3_YEAST	SNU13_YEAST
SODC_YEAST	SOF1_YEAST	SPB1_YEAST	SPT16_YEAST	SSU72_YEAST	STH1_YEAST
STI1_YEAST	SUB2_YEAST	SWD2_YEAST	SYDM_YEAST	SYFA_YEAST	SYFB_YEAST
SYIC_YEAST	SYNC_YEAST	SYV_YEAST	TBA1_YEAST	TBB_YEAST	TBP6_YEAST
TBP_YEAST	TCPB_YEAST	TCPD_YEAST	TCPE_YEAST	TCPQ_YEAST	TCTP_YEAST
TF2B_YEAST	TFB3_YEAST	TIM13_YEAST	TOP2_YEAST	TRX1_YEAST	TWF1_YEAST
UBA1_YEAST	UBC12_YEAST	UBC13_YEAST	UBC2_YEAST	UBP15_YEAST	UBP8_YEAST
UFD4_YEAST	UGPA1_YEAST	ULA1_YEAST	URA7_YEAST	UTP15_YEAST	UTP4_YEAST
UTP7_YEAST	VATB_YEAST	VATD_YEAST	VATG_YEAST	VPS1_YEAST	VPS4_YEAST
WEB1_YEAST	YAK1_YEAST	YB09_YEAST	YBE7_YEAST	YCW2_YEAST	YD036_YEAST
YEM6_YEAST	YEQ8_YEAST	YGA4_YEAST	YGJ9_YEAST	YM71_YEAST	YNL0_YEAST
YP247_YEAST	YPT1_YEAST	YSH1_YEAST	YTH1_YEAST	YTM1_YEAST	ZUO1_YEAST

Tan- (Validation method: one protein found in the mitochondrial membrane and the other protein found in the nuclear membrane)

ADT2_YEAST	ADT3_YEAST	AQY1_YEAST	ATPB_YEAST	ATPG_YEAST	CACM_YEAST
COAC_YEAST	COQ1_YEAST	COQ2_YEAST	CSE1_YEAST	DHSB_YEAST	FET3_YEAST
HEMH_YEAST	HYM1_YEAST	IMB3_YEAST	KAD2_YEAST	MDJ1_YEAST	MPCP_YEAST
OAC1_YEAST	OMA1_YEAST	PLSC_YEAST	PRS10_YEAST	PRS6A_YEAST	PRS6B_YEAST
PRS7_YEAST	PSA3_YEAST	PSB2_YEAST	PSB3_YEAST	PSB5_YEAST	PSD1_YEAST
RL35_YEAST	RPN10_YEAST	RPN12_YEAST	RPN1_YEAST	RPN3_YEAST	RPN5_YEAST
RPN7_YEAST	RPN8_YEAST	UQCR2_YEAST			

GFP- (Validation method: random selection of two proteins from two non-adjacent cellular compartments from the Yeast GFP Localization database)

ACS2_YEAST	AGE2_YEAST	AIP1_YEAST	ALG13_YEAST	ALG1_YEAST	ARL3_YEAST
ATC6_YEAST	ATC7_YEAST	ATPA_YEAST	AZF1_YEAST	BAS1_YEAST	BCS1_YEAST
BFR2_YEAST	BRE1_YEAST	BUD31_YEAST	BZZ1_YEAST	CAPZA_YEAST	CAPZB_YEAST
CASP_YEAST	CBPY_YEAST	CORO_YEAST	COX10_YEAST	CYB5_YEAST	DBP4_YEAST
DID4_YEAST	DLDH_YEAST	DYL1_YEAST	ELF1_YEAST	ERD2_YEAST	ERG24_YEAST
ERG27_YEAST	ERP1_YEAST	ERV25_YEAST	ETFA_YEAST	FAB1_YEAST	FAT1_YEAST
FIMB_YEAST	FKH2_YEAST	GAR1_YEAST	GATH_YEAST	GCN5_YEAST	GPI11_YEAST
H2A2_YEAST	HLJ1_YEAST	IDH1_YEAST	IPL1_YEAST	ISW1_YEAST	KAPB_YEAST
KIN28_YEAST	KKK1_YEAST	LCB1_YEAST	LHS1_YEAST	LONM_YEAST	MCES_YEAST
MCX1_YEAST	MOT1_YEAST	MPD1_YEAST	MPS1_YEAST	MRS3_YEAST	MSH6_YEAST

MSP1_YEAST	MVP1_YEAST	MYO1_YEAST	MYO5_YEAST	NCL1_YEAST	NHP2_YEAST
NOP10_YEAST	NOP12_YEAST	NTG2_YEAST	NUF2_YEAST	ODP2_YEAST	ODPB_YEAST
ORM1_YEAST	PCNA_YEAST	PEX1_YEAST	PEX6_YEAST	PRK1_YEAST	PRP5_YEAST
PRS8_YEAST	PSA2_YEAST	PSB6_YEAST	PUS1_YEAST	PUS3_YEAST	PUT2_YEAST
QRI7_YEAST	RM02_YEAST	RM08_YEAST	RPC19_YEAST	RPC2_YEAST	RSMB_YEAST
RUVB1_YEAST	SCP1_YEAST	SCY1_YEAST	SEN1_YEAST	SMC4_YEAST	SNC2_YEAST
SNF2_YEAST	SNX3_YEAST	SODM_YEAST	SSN8_YEAST	STE6_YEAST	STH1_YEAST
SUCA_YEAST	SYDM_YEAST	SYEM_YEAST	TAL1_YEAST	TIM16_YEAST	TWF1_YEAST
UBC6_YEAST	UTP13_YEAST	UTP4_YEAST	VA0D_YEAST	VAC8_YEAST	VATB_YEAST
VATD_YEAST	VATE_YEAST	VATG_YEAST	VATL1_YEAST	VPH1_YEAST	VPS27_YEAST
VPS34_YEAST	VPS4_YEAST	VPS60_YEAST	WEB1_YEAST	XRN2_YEAST	YB09_YEAST
YB91_YEAST	YBY9_YEAST	YCFI_YEAST	YG5F_YEAST	YHA2_YEAST	YIS4_YEAST
YO246_YEAST	YO7T_YEAST	ZRC1_YEAST			

Appendix 2. Papers for Ligand-Binding Sites

Adrenergic receptor

Reference #	Reference	Protein (PID)	Method(s)
1	(Cavalli *et al.*, 1996)	ADA1B_HUMAN(P35368)	mutation
2	(Chen *et al.*, 1999)	ADA1B_MESAU(P18841)	mutation
3	(Chung *et al.*, 1988)	ADRB2_HUMAN(P07550)	mutation
4	(Dohlman *et al.*, 1988)	ADRB2_HUMAN(P07550)	affinity labelling
5	(Fraser *et al.*, 1988)	ADRB2_HUMAN(P07550)	mutation
6	(Fraser, 1989a)	ADRB2_HUMAN(P07550)	mutation
7	(Gouldson *et al.*, 1997)	ADRB2_HUMAN(P07550)	mutation/CMA
8	(Green *et al.*, 1993)	ADRB2_HUMAN(P07550)	natural SNP
9	(Hamaguchi *et al.*, 1996)	ADA1A_HUMAN(P35348)	mutation
10	(Hieble *et al.*, 1998)	ADA2A_HUMAN(P08913)	mutation
11	(Hwa *et al.*, 1996)	ADA1A_RAT(P43140), ADA1B_MESAU(P18841)	mutation
12	(Isogaya *et al.*, 1998)	ADRB2_HUMAN(P07550)	mutation/chimera
13	(Kikkawa *et al.*, 1998)	ADRB2_HUMAN(P07550), ADRB1_HUMAN(P08588)	mutation/chimera
14	(Lei *et al.*, 2005)	ADA1A_HUMAN(P35348)	SNP
15	(Liapakis *et al.*, 2000)	ADRB2_HUMAN(P07550)	mutation+ligand analogues
16	(Marjamaki *et al.*, 1998)	ADA2A_HUMAN(P08913)	mutation/substituted cysteine accessibility
17	(Pauwels and Colpaert, 2000)	ADA2A_HUMAN(P08913)	mutation/substituted cysteine accessibility
18	(Peltonen *et al.*, 2003)	ADA2A_HUMAN(P08913)	mutation/substituted cysteine accessibility
19	(Perez *et al.*, 1998)	ADA1B_MESAU(P18841), ADA1A_RAT(P43140)	mutation
20	(Porter *et al.*, 1996)	ADA1B_MESAU(P18841)	mutation
21	(Porter *et al.*, 1998)	ADA1B_MESAU(P18841)	mutation
22	(Porter and Perez, 1999)	ADA1B_MESAU(P18841)	mutation
23	(Rudling *et al.*, 1999)	ADA2A_HUMAN(P08913)	mutation/substituted cysteine accessibility
24	(Sato *et al.*, 1999)	ADRB2_HUMAN(P07550)	mutation+ligand analogues
25	(Strader *et al.*, 1987)	ADRB2_MESAU(P04274)	mutation/ligand analogues
26	(Strader *et al.*, 1988)	ADRB2_MESAU(P04274)	mutation/ligand analogues
27	(Strader *et al.*, 1989)	ADRB2_HUMAN(P07550)	mutation
28	(Strader and Dixon, 1991)	ADRB2_MESAU(P04274)	mutation/ligand analogues
29	(Sugimoto *et al.*, 2002)	ADRB1_HUMAN(P08588)	mutation
30	(Suryanarayana *et al.*, 1991)	ADA2A_HUMAN(P08913)	mutation
31	(Suryanarayana and Kobilka, 1993)	ADRB2_HUMAN(P07550)	mutation/CMA (computational)
32	(Wang *et al.*, 1991)	ADA2A_HUMAN(P08913)	mutation/substituted cysteine accessibility
33	(Waugh *et al.*, 2000)	ADA1A_RAT(P43140)	mutation
34	(Waugh *et al.*, 2001)	ADA1A_RAT(P43140)	mutation
35	(Wieland *et al.*, 1996)	ADRB2_HUMAN(P07550)	mutation/modelling
36	(Wong *et al.*, 1988)	ADRB1_MELGA(P07700)	affinity labelling

37	(Zuscik *et al.*, 1998)	ADRB2_HUMAN(P07550)	mutation
38	(Zuurmond *et al.*, 1999)	ADRB2_HUMAN(P07550)	mutation/modelling

Chemokine receptor

Reference #	Reference	Protein (PID)	Method(s)
39	(Ai and Liao, 2002)	CCR6_HUMAN(P51684)	mutation
40	(Berkhout *et al.*, 2003)	CCR2_HUMAN(P41597)	mutation
41	(Blanpain *et al.*, 1999)	CCR5_HUMAN(P51681)	mutation/deletion
42	(Blanpain *et al.*, 2003)	CCR5_HUMAN(P51681)	mutation/chimera
43	(Blanpain *et al.*, 2000)	CCR5_HUMAN(P51681)	mutation
44	(Brelot *et al.*, 2000)	CXCR4_HUMAN(P61073)	mutation
45	(Chen *et al.*, 2006)	CX3C1_HUMAN(P49238)	mutation/surface plasmon resonance
46	(Colvin *et al.*, 2006)	CXCR3_HUMAN(P49682)	mutation/chimera
47	(de Mendonca *et al.*, 2005)	CCR1_HUMAN(P32246)	mutation
48	(Doranz *et al.*, 1999)	CXCR4_HUMAN(P61073)	chimera
49	(Dragic *et al.*, 1998)	CCR5_HUMAN(P51681)	mutation/deletion/chimera
50	(Dragic *et al.*, 2000)	CCR5_HUMAN(P51681)	mutation
51	(Farzan *et al.*, 2002)	CXCR4_HUMAN(P61073)	mutation
52	(Fong *et al.*, 2002)	CX3C1_HUMAN(P49238)	mutation/surface plasmon resonance
53	(Gerlach *et al.*, 2001)	CXCR4_HUMAN(P61073)	mutation
54	(Govaerts *et al.*, 2001)	CCR5_HUMAN(P51681)	mutation
55	(Gutierrez *et al.*, 2004)	CCR8_MOUSE(P56484)	mutation
56	(Hatse *et al.*, 2001)	CXCR4_HUMAN(P61073)	mutation
57	(Howard *et al.*, 1999)	CCR5_HUMAN(P51681)	mutation/deletion
58	(Mirzadegan *et al.*, 2000)	CCR2_HUMAN(P41597)	mutation
59	(Monteclaro and Charo, 1996)	CCR2_HUMAN(P41597)	mutation/chimera
60	(Monteclaro and Charo, 1997)	CCR2_HUMAN(P41597)	mutation/chimera
61	(Preobrazhensky *et al.*, 2000)	CCR2_HUMAN(P41597)	mutation/chimera
62	(Samson *et al.*, 1997)	CCR2_HUMAN(P41597), CCR5_HUMAN(P51681)	mutation/chimera
63	(Wu *et al.*, 1997)	CCR5_HUMAN(P51681)	mutation/chimera
64	(Xanthou *et al.*, 2003)	CXCR3_HUMAN(P49682)	mutation/chimera
65	(Youn *et al.*, 2001)	CCR5_HUMAN(P51681)	mutation
66	(Zhou *et al.*, 2001)	CXCR4_HUMAN(P61073)	mutation
67	(Zoffmann *et al.*, 2002)	CCR1_HUMAN(P32246)	mutation

Interleukin-8 receptor

Reference #	Reference	Protein (PID)	Method(s)
68	(Hebert *et al.*, 1993)	CXCR1_HUMAN(P25024)	mutation
69	(Katancik *et al.*, 2000)	CXCR2_HUMAN(P25025)	mutation
70	(Leong *et al.*, 1994)	CXCR1_HUMAN(P25024)	mutation
71	(Luo *et al.*, 1997)	CXCR2_HUMAN(P25025)	mutation/modelling
72	(Skelton *et al.*, 1999)	CXCR1_HUMAN(P25024)	NMR

Dopamine receptor

Reference	Reference	Protein (PID)	Method(s)

#			
73	(Alberts *et al.*, 1998)	DRD3_HUMAN(P35462)	mutation
74	(Boeckler *et al.*, 2005)	DRD2_HUMAN(P14416)	mutation/modelling
75	(Cho *et al.*, 1995)	DRD2_HUMAN(P14416)	mutation/modelling
76	(Cho *et al.*, 1996)	DRD1_HUMAN(P21728)	mutation
77	(Coley *et al.*, 2000)	DRD2_RAT(P61169)	mutation
78	(Cox *et al.*, 1992)	DRD2_RAT(P61169)	mutation
79	(Cravchik and Gejman, 1999)	DRD5_HUMAN(P21918)	nature SNP
80	(Daniell *et al.*, 1994)	DRD2_RAT(P61169)	mutation
81	(Fu *et al.*, 1996)	DRD2_HUMAN(P14416)	mutation/modelling
82	(Javitch *et al.*, 1995)	DRD2_HUMAN(P14416)	mutation
83	(Javitch *et al.*, 1996)	DRD2_HUMAN(P14416)	mutation/predicted 3D structure/modelling
84	(Javitch *et al.*, 1998)	DRD2_HUMAN(P14416)	mutation/predicted 3D structure/modelling
85	(Javitch *et al.*, 1999)	DRD2_HUMAN(P14416)	mutation
86	(Javitch *et al.*, 2000)	DRD2_HUMAN(P14416)	mutation
87	(Jeanneteau *et al.*, 2004)	DRD3_RAT(P19020)	deletion
88	(Jensen *et al.*, 1995)	DRD1_HUMAN(P21728)	mutation
89	(Kalani *et al.*, 2004)	DRD2_HUMAN(P14416)	mutation/predicted 3D structure/modelling
90	(Lee *et al.*, 2000a)	DRD2_HUMAN(P14416)	mutation/predicted 3D structure
91	(Lundstrom *et al.*, 1998)	DRD3_HUMAN(P35462)	mutation
92	(Mansour *et al.*, 1992)	DRD2_HUMAN(P14416)	mutation/predicted 3D structure
93	(Neve *et al.*, 1991)	DRD2_RAT(P61169)	mutation
94	(Neve *et al.*, 2001)	DRD2_RAT(P61169)	mutation
95	(Pollock *et al.*, 1992)	DRD1_HUMAN(P21728)	mutation
96	(Sartania and Strange, 1999)	DRD3_HUMAN(P35462)	mutation
97	(Schetz *et al.*, 2000)	DRD4_RAT(P30729)	mutation
98	(Simpson *et al.*, 1999)	DRD2_HUMAN(P14416), DRD4_HUMAN(P21917)	mutation/predicted 3D structure
99	(Tomic *et al.*, 1993)	DRD1_HUMAN(P21728)	mutation
100	(Wilcox *et al.*, 2000)	DRD2_RAT(P61169)	mutation
101	(Woodward *et al.*, 1994)	DRD2_RAT(P61169)	mutation
102	(Woodward *et al.*, 1996)	DRD2_RAT(P61169)	mutation

Histamine receptor

Reference #	Reference	Protein (PID)	Method(s)
103	(Gantz *et al.*, 1992)	HRH2_CANFA(P17124)	mutation
104	(Leurs *et al.*, 1994)	HRH1_CAVPO(P31389)	mutation
105	(Leurs *et al.*, 1995)	HRH1_CAVPO(P31389)	mutation
106	(Ligneau *et al.*, 2000)	HRH3_RAT(Q9QYN8)	mutation
107	(Moguilevsky *et al.*, 1995)	HRH1_HUMAN(P35367)	mutation
108	(Nonaka *et al.*, 1998)	HRH1_HUMAN(P35367)	mutation
109	(Ohta *et al.*, 1994)	HRH1_HUMAN(P35367)	mutation
110	(Shin *et al.*, 2002)	HRH4_HUMAN(Q9H3N8)	mutation
111	(Wieland *et al.*, 1999)	HRH1_CAVPO(P31389)	mutation

Muscarinic receptor

Reference #	Reference	Protein (PID)	Method(s)
112	(Allman *et al.*, 2000)	ACM1_RAT(P08482)	mutation
113	(Bluml *et al.*, 1994a)	ACM3_RAT(P08483)	mutation
114	(Fraser *et al.*, 1989b)	ACM1_RAT(P08482)	mutation/affinity labelling
115	(Heitz *et al.*, 1999)	ACM2_HUMAN(P08172)	mutation
116	(Huang *et al.*, 1999a)	ACM1_HUMAN(P11229)	mutation
117	(Kurtenbach *et al.*, 1990)	ACM1_RAT(P08482)	mutation/affinity labelling
118	(Leppik *et al.*, 1994)	ACM2_HUMAN(P08172)	mutation
119	(Lu and Hulme, 1999)	ACM1_RAT(P08482)	mutation/affinity labelling
120	(Lu *et al.*, 2001)	ACM1_RAT(P08482)	mutation
121	(Matsui *et al.*, 1995)	ACM1_RAT(P08482)	mutation
122	(Mosser *et al.*, 2002)	ACM2_RAT(P10980)	mutation
123	(Page *et al.*, 1995)	ACM1_RAT(P08482)	mutation/affinity labelling
124	(Savarese *et al.*, 1992)	ACM1_RAT(P08482)	mutation
125	(Spalding *et al.*, 1994)	ACM1_RAT(P08482)	mutation/affinity labelling
126	(Spalding *et al.*, 1995)	ACM1_RAT(P08482)	mutation/affinity labelling
127	(Vogel *et al.*, 1997)	ACM2_PIG(P06199)	mutation
128	(Vogel *et al.*, 1999)	ACM2_PIG(P06199), ACM2_RAT(P10980)	mutation
129	(Ward *et al.*, 1999)	ACM1_RAT(P08482)	mutation
130	(Wess *et al.*, 1991)	ACM3_RAT(P08483)	mutation
131	(Wess *et al.*, 1992)	ACM3_RAT(P08483)	mutation
132	(Wess *et al.*, 1993)	ACM3_RAT(P08483)	mutation
133	(Wess *et al.*, 1995)	ACM3_RAT(P08483)	mutation

Serotonin receptor

Reference #	Reference	Protein (PID)	Method(s)
134	(Almaula *et al.*, 1996a)	5HT2A_HUMAN(P28223)	mutation
135	(Almaula *et al.*, 1996b)	5HT2C_HUMAN(P28223)	mutation
136	(Boess *et al.*, 1997)	5HT6R_RAT(P31388)	mutation
137	(Boess *et al.*, 1998)	5HT6R_RAT(P31388)	mutation
138	(Chanda *et al.*, 1993)	5HT1A_HUMAN(P08908)	mutation
139	(Choudhary *et al.*, 1993)	5HT2A_HUMAN(P28223)	mutation
140	(Choudhary *et al.*, 1995)	5HT2A_HUMAN(P28223)	mutation
141	(Del Tredici *et al.*, 2004)	5HT1A_HUMAN(P08908)	natural SNP
142	(Glennon *et al.*, 1996)	5HT1B_HUMAN(P28222)	mutation
143	(Granas and Larhammar, 1999)	5HT1B_HUMAN(P28222)	mutation
144	(Granas *et al.*, 1998)	5HT1B_HUMAN(P28222)	mutation
145	(Granas *et al.*, 2001)	5HT1B_HUMAN(P28222)	mutation
146	(Guan *et al.*, 1992)	5HT1A_HUMAN(P08908)	mutation
147	(Herrick-Davis *et al.*, 2005)	5HT2C_HUMAN(P28335)	mutation
148	(Ho *et al.*, 1992)	5HT1A_HUMAN(P08908)	mutation
149	(Johnson *et al.*, 1994)	5HT2A_RAT(P14842)	mutation
150	(Johnson *et al.*, 1997)	5HT2A_RAT(P14842)	mutation
151	(Kao *et al.*, 1992)	5HT2A_HUMAN(P28223)	mutation

152	(Kohen *et al.*, 2001)	5HT6R_MOUSE(Q9R1C8)	mutation
153	(Kristiansen *et al.*, 2000)	5HT2A_HUMAN(P28223)	mutation
154	(Kuipers *et al.*, 1997)	5HT1A_HUMAN(P08908)	mutation
155	(Manivet *et al.*, 2002)	5HT2B_HUMAN(P41595), 5HT2B_RAT(P30994)	mutation
156	(Mialet *et al.*, 2000)	5HT4R_HUMAN(Q13639)	mutation
157	(Obosi *et al.*, 1997)	5HT7R_MOUSE(P32304)	mutation/deletion
158	(Oksenberg *et al.*, 1992)	5HT1B_HUMAN(P28222)	mutation
159	(Parker *et al.*, 1993)	5HT1B_HUMAN(P28222)	mutation
160	(Roth *et al.*, 1993)	5HT2A_HUMAN(P28223)	mutation
161	(Roth *et al.*, 1997)	5HT2A_HUMAN(P28223)	mutation
162	(Shapiro *et al.*, 2000)	5HT2A_RAT(P14842)	mutation
163	(Wang *et al.*, 1993)	5HT2A_HUMAN(P28223)	mutation

Appendix 3. Papers for G Protein-Coupling Sites

Adrenergic receptor

Reference #	Reference	Protein (PID)	Method(s)
164	(Cotecchia *et al.*, 1990)	ADA1B_MESAU(P18841)	mutation/chimera
165	(Eason and Liggett, 1995)	ADA2A_HUMAN(P08913)	deletion/chimera
166	(Eason and Liggett, 1996)	ADA2A_HUMAN(P08913)	mutation/deletion/chimera
167	(Greasley *et al.*, 2001)	ADA1B_MESAU(P18841)	mutation/chimera
168	(Hawes *et al.*, 1994)	ADA1B_HUMAN(P35368)	mutation/peptide
169	*(Liggett et al., 1991)*	ADRB2_HUMAN(P07550)	deletion/chimera
170	(Moro *et al.*, 1993)	ADRB2_MESAU(P04274)	mutation
171	(Nasman *et al.*, 1997)	ADA2A_MOUSE(Q01338)	chimera
172	(O'Dowd *et al.*, 1988)	ADRB2_HUMAN(P07550)	deletion/chimera
173	(Valiquette *et al.*, 1993)	ADRB2_HUMAN(P07550)	mutation
174	(Wade *et al.*, 1999)	ADA2A_HUMAN(P08913)	mutation/chimera
175	(Wong *et al.*, 1990)	ADRB1_MELGA(P07700/P11229)	chimera
176	(Wurch *et al.*, 2003)	ADA1B_HUMAN(P35368)	mutation/peptide

Chemokine receptor

Reference #	Reference	Protein (PID)	Method(s)
177	(Arai *et al.*, 1997)	CCR2_HUMAN(P41597)	deletion
178	(Auger *et al.*, 2002)	CCR3_HUMAN(P51677)	mutation
179	(Brelot *et al.*, 2000)	CXCR4_HUMAN(P61073)	deletion
180	(Gosling *et al.*, 1997)	CCR5_HUMAN(P51681)	mutation/deletion
181	(Kraft *et al.*, 2001)	CCR5_HUMAN(P51681)	mutation/deletion

Interleukin-8 receptor

Reference #	Reference	Protein (PID)	Method(s)
182	(Ben-Baruch *et al.*, 1995)	CXCR2_HUMAN(P25025)	deletion
183	(Damaj *et al.*, 1996)	CXCR1_HUMAN(P25024)	mutation
184	(Xie *et al.*, 1997)	CXCR2_HUMAN(P25025)	mutation

Dopamine receptor

Reference #	Reference	Protein (PID)	Method(s)
185	(Chaar *et al.*, 2001)	DRD1_HUMAN(P21728)	deletion
186	(Filteau *et al.*, 1999)	DRD2_HUMAN(P14416), DRD3_HUMAN(P35462)	chimera/peptide
187	(Ilani *et al.*, 2002)	DRD2_MOUSE(P61168), DRD3_MOUSE(P30728)	chimera
188	(Malek *et al.*, 1993)	DRD2_HUMAN(P14416)	chimera/peptide
189	(Woodward *et al.*, 1996)	DRD2_RAT(P61169)	mutation

Histamine receptor

Reference #	Reference	Protein (PID)	Method(s)
NA			

Muscarinic receptor

Reference #	Reference	Protein (PID)	Method(s)
190	(Blin *et al.*, 1995)	ACM3_RAT(P08483)	mutation/chimera
191	(Bluml *et al.*, 1994b)	ACM3_RAT(P08483)	mutation/chimera
192	(Bluml *et al.*, 1994c)	ACM3_RAT(P08483)	chimera/insertion
193	(Burstein *et al.*, 1995)	ACM5_HUMAN(P08912)	mutation
194	(Burstein *et al.*, 1996)	ACM5_HUMAN(P08912)	mutation
195	(Burstein *et al.*, 1998)	ACM5_HUMAN(P08912)	mutation
196	(Duerson *et al.*, 1993)	ACM3_HUMAN(P20309)	mutation
197	(Fraser *et al.*, 1989b)	ACM1_RAT(P08482)	mutation
198	(Hawes *et al.*, 1994)	ACM1_HUMAN(P11229), ACM2_HUMAN(P08172)	mutation/chimera/peptide
199	(Hill-Eubanks *et al.*, 1996)	ACM5_HUMAN(P08912)	mutation
200	(Huang *et al.*, 1999b)	ACM1_HUMAN(P11229)	mutation
201	(Kostenis *et al.*, 1997a)	ACM2_HUMAN(P08172)	mutation/chimera/peptide
202	(Kostenis *et al.*, 1997b)	ACM3_RAT(P08483)	mutation/chimera/insertion
203	(Liu *et al.*, 1995)	ACM2_HUMAN(P08172)	mutation/chimera/peptide
204	(Liu *et al.*, 1996)	ACM2_HUMAN(P08172)	deletion
205	(Moro *et al.*, 1993)	ACM1_HUMAN(P11229)	mutation
206	(Wess *et al.*, 1995)	ACM3_RAT(P08483)	mutation/chimera
207	(Wu *et al.*, 2000)	ACM3_RAT(P08483)	mutation

Serotonin receptor

Reference #	Reference	Protein (PID)	Method(s)
208	(Albert *et al.*, 1998)	5HT1A_HUMAN(P08908)	mutation
209	(Kushwaha and Albert, 2005)	5HT1A_RAT(P19327)	mutation
210	(Kushwaha *et al.*, 2006)	5HT1A_RAT(P19327)	mutation
211	(Lembo *et al.*, 1997)	5HT1A_HUMAN(P08908), 5HT1A_RAT(P19327)	mutation
212	(Malmberg and Strange, 2000)	5HT1A_HUMAN(P08908)	mutation
213	(Obosi *et al.*, 1997)	5HT7R_MOUSE(P32304)	mutation
214	(Oksenberg *et al.*, 1995)	5HT2A_HUMAN(P28223)	chimera
215	(Papoucheva *et al.*, 2004)	5HT1A_MOUSE(Q64264)	mutation
216	(Price *et al.*, 2001)	5HT2C_HUMAN(P28335)	RNA editing

Appendix 4. Papers for Oligomerization Sites

Adrenergic receptor

Reference #	Reference	Protein (PID)	Method(s)
217	(Carrillo *et al.*, 2003)	ADA1B_MESAU(P18841)	co-immunoprecipitation+TR-FRET
218	(Carrillo *et al.*, 2004)	ADA1B_MESAU(P18841)	mutation/co-immunoprecipitation+TR-FRET
219	(Hebert *et al.*, 1996)	ADRB2_HUMAN(P07550)	immoprecipitation
220	(Lopez-Gimenez *et al.*, 2007)	ADA1B_MESAU(P18841)	mutation/co-immunoprecipitation+TR-FRET
221	(Stanasila *et al.*, 2003)	ADA1B_MESAU(P18841)	co-immunoprecipitation+TR-FRET
222	(Xu *et al.*, 2003)	ADA2A_HUMAN(P08913), ADRB1_HUMAN(P08588)	co-immunoprecipitation

Chemokine receptor

Reference #	Reference	Protein (PID)	Method(s)
223	(de Juan *et al.*, 2005)	CCR5_HUMAN(P51681)	mutation+FRET/co-immunoprecipitation+FRET
224	(Gouldson *et al.*, 2001)	CXCR4_HUMAN(P61073)	CMA
225	(Hernanz-Falcon *et al.*, 2004)	CCR5_HUMAN(P51681)	mutation+FRET/co-immunoprecipitation+FRET

Interleukin-8 receptor

Reference #	Reference	Protein (PID)	Method(s)
NA			

Dopamine receptor

Reference #	Reference	Protein (PID)	Method(s)
226	(Canals *et al.*, 2003)	DRD2_HUMAN(P14416)	pull-down+mass spectrometry/FRET+BRET/cross-linking
227	(Ciruela *et al.*, 2004)	DRD2_HUMAN(P14416)	pull-down+mass spectrometry/FRET+BRET
228	(Guo *et al.*, 2003)	DRD2_HUMAN(P14416)	mutation/immunoblot
229	(Lee *et al.*, 2002)	DRD2_HUMAN(P14416)	immunoblot
230	(Ng *et al.*, 1996)	DRD2_HUMAN(P14416)	FRET+BRET/cross-linking
231	(Woods *et al.*, 2005)	DRD1_HUMAN(P21728)	unknown

Histamine receptor

Reference #	Reference	Protein (PID)	Method(s)
232	(Carrillo *et al.*, 2003)	HRH1_HUMAN(P35367)	co-immunoprecipitation+TR-

			FRET

Muscarinic receptor

Reference #	Reference	Protein (PID)	Method(s)
233	(Zeng and Wess, 1999)	ACM3_RAT(P08483)	mutation

Serotonin receptor

Reference #	Reference	Protein (PID)	Method(s)
NA			

Appendix 5. Papers for Protein-Protein Interaction Sites

Adrenergic receptor

Reference #	Reference	Protein (PID)	Method(s)
234	(Cao *et al.*, 1999)	ADRB2_HUMAN(P07550)	mutation
235	(DeGraff *et al.*, 2002)	ADA2B_HUMAN(P18089)	mutation/deletion
236	(Diviani *et al.*, 2003)	ADA1B_MESAU(P18841)	deletion
237	(Hall *et al.*, 1998)	ADRB2_HUMAN(P07550)	mutation
238	(He *et al.*, 2004)	ADRB1_HUMAN(P08588)	mutation/peptide array/Y2H + co-immunoprecipitation
239	(He *et al.*, 2006)	ADRB1_HUMAN(P08588)	mutation/peptide array/Y2H + co-immunoprecipitation
240	(Hu *et al.*, 2000)	ADRB1_HUMAN(P08588)	mutation/peptide array/Y2H + co-immunoprecipitation
241	(Hu *et al.*, 2003)	ADRB1_HUMAN(P08588)	mutation/peptide array/Y2H + co-immunoprecipitation
242	(Javitch *et al.*, 1997)	ADRB2_HUMAN(P07550)	mutation
243	(Klein *et al.*, 1997)	ADRB2_MOUSE(P18762)	Y2H + co-immunoprecipitation
244	(Marion *et al.*, 2006)	ADRB2_HUMAN(P07550)	mutation
245	(Pak *et al.*, 2002)	ADRB1_HUMAN(P08588)	mutation/peptide array/Y2H + co-immunoprecipitation
246	(Wang and Limbird, 2002)	ADA2A_PIG(P18871)	deletion
247	(Xu *et al.*, 2001)	ADRB1_HUMAN(P08588)	mutation/peptide array/Y2H + co-immunoprecipitation
248	(Zhang *et al.*, 2004)	ADA1A_HUMAN(P35348), ADA1B_HUMAN(P35368), ADA1D_HUMAN(P25100)	Y2H

Chemokine receptor

Reference #	Reference	Protein (PID)	Method(s)
249	(Bieniasz *et al.*, 1997)	CCR5_HUMAN(P51681)	mutation/chimera
250	(Blanpain *et al.*, 1999)	CCR5_HUMAN(P51681)	mutation/chimera/expression
251	(Brelot *et al.*, 1997)	CXCR4_HUMAN(P61073)	mutation/chimera/deletion
252	(Brelot *et al.*, 1999)	CXCR4_HUMAN(P61073)	mutation/chimera/deletion
253	(Brelot *et al.*, 2000)	CXCR4_HUMAN(P61073)	mutation/chimera/deletion
254	(Chabot *et al.*, 1999)	CXCR4_HUMAN(P61073)	mutation/chimera/deletion
255	(Cheng *et al.*, 2000)	CXCR4_HUMAN(P61073)	pull-down assay/deletion/cross-linking assay
256	(Doranz *et al.*, 1997)	CCR5_HUMAN(P51681)	mutation/chimera/expression
257	(Dragic *et al.*, 1998)	CCR5_HUMAN(P51681)	mutation/chimera/expression
258	(Farzan *et al.*, 1998)	CCR5_HUMAN(P51681)	mutation/chimera/expression
259	(Cormier *et al.*, 2000)	CCR5_HUMAN(P51681)	mutation/chimera/expression
260	(Hill *et al.*, 1998)	CCR5_HUMAN(P51681)	mutation/chimera/expression
261	(Huttenrauch *et al.*, 2002)	CCR5_HUMAN(P51681)	mutation/Y2H/pull-down assay

262	(Ko *et al.*, 2004)	CCR1_HUMAN(P32246), CCR5_HUMAN(P51681)	mutation/Y2H/pull-down assay
263	(Kraft *et al.*, 2001)	CCR5_HUMAN(P51681)	mutation/Y2H/pull-down assay
264	(Kuhmann *et al.*, 1997)	CCR5_HUMAN(P51681)	mutation/chimera/expression
265	(Lee *et al.*, 2004)	CCR1_HUMAN(P32246)	Y2H
266	(Lin *et al.*, 2003)	CXCR4_HUMAN(P61073)	mutation/chimera/deletion
267	(Lu *et al.*, 1997)	CCR5_HUMAN(P51681), CXCR4_HUMAN(P61073)	mutation/chimera/expression
268	(Papin and Subramaniam, 2004)	CXCR5_MOUSE(Q04683)	Y2H
269	(Rabut *et al.*, 1998)	CCR5_HUMAN(P51681)	mutation/chimera/expression
270	(Rey *et al.*, 2002)	CCR5_HUMAN(P51681), CXCR4_HUMAN(P61073)	mutation/Y2H/pull-down assay/deletion
271	(Ross *et al.*, 1998)	CCR5_HUMAN(P51681)	mutation/chimera/expression
272	(Siciliano *et al.*, 1999)	CCR5_HUMAN(P51681)	mutation/chimera
273	(Wu *et al.*, 1997)	CCR5_HUMAN(P51681)	mutation/chimera/expression

Interleukin-8 receptor

Reference #	Reference	Protein (PID)	Method(s)
274	(Fan *et al.*, 2001a)	CXCR2_HUMAN(P25025)	mutation/GST pull-down assay
275	(Fan *et al.*, 2001b)	CXCR2_HUMAN(P25025)	mutation/GST pull-down assay
276	(Fan *et al.*, 2002)	CXCR2_HUMAN(P25025)	mutation

Dopamine receptor

Reference #	Reference	Protein (PID)	Method(s)
277	(Bermak *et al.*, 2001)	DRD1_RAT(P18901)	mutation
278	(Bofill-Cardona *et al.*, 2000)	DRD2_HUMAN(P14416)	mass spectrometry/co-immunoprecipitation
279	(Jeanneteau *et al.*, 2004)	DRD2_RAT(P61169), DRD3_RAT(P19020)	mutation/deletion
280	(Lee *et al.*, 2002)	DRD1_RAT(P18901)	co-immunoprecipitation
281	(Li *et al.*, 2000)	DRD2_HUMAN(P14416)	Y2H+mutation
282	(Liu *et al.*, 2000)	DRD5_HUMAN(P25115)	co-immunoprecipitation
283	(Liu *et al.*, 2007)	DRD2_RAT(P61169)	mutation
284	(Woods *et al.*, 2008)	DRD2_HUMAN(P14416)	mass spectrometry/co-immunoprecipitation

Histamine receptor

Reference #	Reference	Protein (PID)	Method(s)
NA			

Muscarinic receptor

Reference #	Reference	Protein (PID)	Method(s)
285	(Lee *et al.*, 2000b)	ACM2_HUMAN(P08172)	mutation/deletion
286	(Lucas *et al.*, 2006)	ACM1_HUMAN(P11229)	mutation
287	(Wu *et al.*, 1997)	ACM3_RAT(P08483)	immunoblotting

Serotonin receptor

Reference #	Reference	Protein (PID)	Method(s)
288	(Becamel *et al.*, 2001)	5HT2C_HUMAN(P28335)	mutation
289	(Gelber *et al.*, 1999)	5HT2A_RAT(P14842)	mutation
290	(Parker *et al.*, 2003)	5HT2C_RAT(P08909)	recombinant peptides

Appendix 6. Proximity to Family- and Subfamily-Level Motifs for Ligand-Binding Sites

Publications for the ligand-binding regions listed below can be found in Appendix 2. All positions were determined based on bovine rhodopsin structure 1F88.

Since accessibility was determined according to 1F88, positions not mapped to 1F88 sequence in the same MSA were denoted as NON_STRUCTURE_RES.

Proximity scores:
0 = in motif; -1 = one residue left to motif; +1 = one residue right to motif; -3 = within three residues left to motif; +3 = within three residues right to motif; -5 = within five residues left to motif; +5 = within five residues right to motif

Adrenergic receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
3, 25, 32	2.50	BURIED		
29	2.56	ACCESSIBLE	-3	
10	2.61	BURIED	0	
29	2.63	ACCESSIBLE	0	
4, 9	2.64	BURIED	0	
1, 20, 25, 26, 28, 32	3.32	ACCESSIBLE		
37	3.35	BURIED		
5, 32	3.49	ACCESSIBLE		
8	4.56	BURIED		
33	4.62	ACCESSIBLE		
14	4.62(+3)	BURIED		
6	5.34(-5)	ACCESSIBLE		
6	5.34(-4)	ACCESSIBLE		
1	5.38	BURIED		
11, 16, 19	5.39	BURIED		+3
33	5.41	ACCESSIBLE		
1, 15, 16, 17, 19, 23, 24	5.42	ACCESSIBLE		
15, 16, 18, 24, 27	5.43	ACCESSIBLE		
15, 16, 17, 18, 19, 24, 27, 32	5.46	BURIED		
14	5.54	ACCESSIBLE	-5	
2	6.51	ACCESSIBLE		
11, 35, 38	6.55	BURIED		-1
12, 13, 29, 34	7.35	BURIED		0
12, 21, 22	7.36	ACCESSIBLE		0
14	7.38	ACCESSIBLE		

7, 30, 31, 34	7.39	ACCESSIBLE		+3
36	7.40	ACCESSIBLE		0
1	7.43	ACCESSIBLE		
25	7.45	ACCESSIBLE		
10, 27	7.46	BURIED		

Chemokine receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
46, 64	1.28(-48)-1.28(-38)	NON_STRUCTURE_RES		0
46, 59, 62, 64	1.28(-37)	NON_STRUCTURE_RES		0, -1
46, 59, 60, 62, 64	1.28(-36)-1.28(-33)	NON_STRUCTURE_RES		0
44, 59, 60, 62, 64	1.28(-32)-1.28(-25)	PARTIALLY_ACCESSIBLE		-5, 0, -1
41, 59, 60, 62, 64	1.28(-24)	ACCESSIBLE		-5, 0, +3
41, 59, 60, 62, 64	1.28(-23)	ACCESSIBLE		+5, 0, -3
41, 46, 59, 60, 62, 64	1.28(-22)	BURIED		+5, -5, 0, -3
41, 59, 60, 62, 64	1.28(-21)	ACCESSIBLE		-5, 0, -1
41, 44, 46, 59, 60, 62, 64	1.28(-20)	ACCESSIBLE		0, -3
41, 44, 59, 60, 62, 64	1.28(-19)	ACCESSIBLE		+1, 0, -3
41, 59, 60, 62, 64	1.28(-18)-1.28(-17)	ACCESSIBLE		0, +3
41, 59, 60, 62, 64	1.28(-16)	ACCESSIBLE		+5, 0
41, 49, 55, 59, 60, 62, 64	1.28(-15)	ACCESSIBLE		+5, 0
41, 55, 57, 59, 60, 61, 62, 64, 66	1.28(-14)	ACCESSIBLE		-5, 0
41, 44, 45, 51, 52, 59, 60, 62, 64, 66	1.28(-13)	ACCESSIBLE		-5, 0
59, 60, 61, 62, 64	1.28(-12)	ACCESSIBLE		+1, 0, -3
59, 60, 62, 64	1.28(-11)-1.28(-9)	PARTIALLY_ACCESSIBLE	-3	0, +3
41, 59, 60, 62, 64	1.28(-8)	ACCESSIBLE	-3	+5, 0
59, 60, 62, 64	1.28(-7)	ACCESSIBLE	-1	0
39, 57, 59, 60, 62, 64	1.28(-6)	ACCESSIBLE	0	+1, 0
48, 59, 60, 62, 64	1.28(-5)	BURIED	0	-5, 0, +3
48, 59, 60, 62, 64	1.28(-4)	ACCESSIBLE	0	0, +3, -3
48, 59, 62, 64	1.28(-3)	ACCESSIBLE	0	+5, 0, -3
45, 48, 59, 62, 64	1.28(-2)	ACCESSIBLE	0	+5, 0, -1
48, 59, 62, 64	1.28(-1)-1.30	ACCESSIBLE	0	0
43, 48, 62, 64	1.31	ACCESSIBLE	0	+5, 0
48, 62, 64	1.32	BURIED	0	+5, 0
48, 49, 50	1.33	ACCESSIBLE	0	0
50	1.35	ACCESSIBLE	0	
47, 50	1.39	BURIED	0	
50	2.56	ACCESSIBLE	0	
54	2.58	BURIED	0	

42	2.59	ACCESSIBLE	0	
50	2.6	ACCESSIBLE	0	
64	2.62	ACCESSIBLE	0	
44, 46, 64	2.63	ACCESSIBLE	0	-3
64	2.64-3.24	PARTIALLY_ACCESSIBLE	+1, -5	
39, 64	3.25	BURIED	-5	
64	3.26	ACCESSIBLE	-3	
42	3.28	BURIED	-1	
40, 47, 50	3.32	ACCESSIBLE	0	
40, 42	3.33	ACCESSIBLE	0	
42	3.36	ACCESSIBLE	0	
53, 56	4.60	ACCESSIBLE		
62, 63, 64	4.62(+2)	ACCESSIBLE		0
42, 62, 63, 64	4.62(+3)	BURIED		0
62, 63, 64	4.62(+4)-4.62(+6)	PARTIALLY_ACCESSIBLE		+1, 0
42, 62, 63, 64	4.62(+7)	BURIED		+5, 0
62, 63, 64, 67	4.62(+8)	BURIED		+5, 0
46, 62, 63, 64, 67	4.62(+9)	BURIED		0
62, 63, 64, 67	4.62(+10)-5.32(-13)	PARTIALLY_ACCESSIBLE		0, -3
39, 62, 63, 64, 67	5.32(-12)	ACCESSIBLE		0, +3, -1
44, 62, 63, 64, 67	5.32(-11)	BURIED		+5, 0
62, 63, 64, 67	5.32(-10)-5.32(-4)	PARTIALLY_ACCESSIBLE		+5, 0, -3
62, 63, 67	5.32(-3)-5.33	ACCESSIBLE		0
62, 63	5.34	NON_STRUCTURE_RES		0, +3
42, 46, 49, 62, 63, 64	5.35	BURIED		+5, 0
49, 62, 63	5.36	ACCESSIBLE		+5, 0
62, 63	5.37-5.42	PARTIALLY_ACCESSIBLE		0
45, 46, 53, 56, 64	6.58	BURIED		
64	6.59-6.61(+2)	ACCESSIBLE		
64, 66	6.61(+3)	ACCESSIBLE		-3
64	7.22(-2)-7.24	ACCESSIBLE		-3
39, 60, 64	7.25	NON_STRUCTURE_RES		-1
64	7.26-7.31	NON_STRUCTURE_RES	0	0
42, 58, 64	7.32	ACCESSIBLE	0	0
64	7.33	ACCESSIBLE	0	0
40, 44, 47, 50, 58	7.39	ACCESSIBLE	0	+5
40	7.40	ACCESSIBLE	0	
65	7.43	ACCESSIBLE	0	

Interleukin-8 receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs

69, 71	1.28(-37)	NON_STRUCTURE_RES		0
69	1.28(-35)	NON_STRUCTURE_RES		0
69	1.28(-32)	ACCESSIBLE		0
68	1.28(-24)	ACCESSIBLE		0
72	1.28(-14)	ACCESSIBLE	-5	-3
72	1.28(-11)	ACCESSIBLE	-3	0
72	1.28(-10)	BURIED	-1	0
72	1.28(-9)	ACCESSIBLE	0	0
72	1.28(-8)	ACCESSIBLE	0	0
72	1.28(-6)	ACCESSIBLE	0	0
69	2.64	BURIED		
69	3.26	ACCESSIBLE	0	
70	5.35	BURIED	0	
70	5.39	BURIED	0	
70	6.58	BURIED	0	
68	7.23	NON_STRUCTURE_RES	0	
68	7.28	NON_STRUCTURE_RES	0	

Dopamine receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
85, 93, 99	2.50	BURIED		
97, 98	2.61	BURIED	0	
89, 98	3.28	BURIED		
97	3.29	BURIED		
74, 89, 90, 92	3.32	ACCESSIBLE		
82	3.33	ACCESSIBLE		
99	3.35	BURIED		
73, 74, 83, 89, 99	3.36	ACCESSIBLE		
94	3.39	ACCESSIBLE		
94	3.42	BURIED		
86	4.50	ACCESSIBLE		
86	4.60	ACCESSIBLE		
86	4.62	ACCESSIBLE		
82, 89, 98	5.38	BURIED		
77, 78, 95, 96, 100, 102	5.42	ACCESSIBLE		
77, 78, 90, 95, 99, 100, 102	5.43	ACCESSIBLE		
77, 82, 89, 90, 92, 95, 99, 100, 102	5.46	BURIED		
82	5.47	ACCESSIBLE		
82	5.50	BURIED		
74, 84, 89	6.48	ACCESSIBLE		
75, 76	6.49	ACCESSIBLE		
74, 75, 84	6.51	ACCESSIBLE		

74, 84, 91, 101	6.55	BURIED		-5, -3
91	7.39	ACCESSIBLE	0	+3
74, 80, 81	7.43	ACCESSIBLE	+3	
79, 81	7.45	ACCESSIBLE	+5	
78, 81, 94	7.46	BURIED	+5	
87	7.65-7.67	ACCESSIBLE	0	
88	7.69(+1)	ACCESSIBLE	+1	-5

Histamine receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
108, 109, 110	3.32	ACCESSIBLE		
106	3.40	BURIED		
111	4.56	BURIED		0
105, 111	5.39	BURIED		0
103, 107	5.42	ACCESSIBLE		0
103, 104, 107, 109, 110	5.46	BURIED		0
111	6.52	ACCESSIBLE		-3
111	6.55	BURIED		0

Muscarinic receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
128	2.50	BURIED		
114, 128	3.26	ACCESSIBLE		
115, 119, 121	3.28	BURIED	-5	
119	3.29	BURIED	-5	
114, 115, 117, 119, 122, 123, 125, 126, 128	3.32	ACCESSIBLE	-1	
119, 130	3.33	ACCESSIBLE	0	
119	3.36	ACCESSIBLE	0	
119	3.37	ACCESSIBLE	0	
119	3.40	BURIED	0	
119	3.46	BURIED	+5	
120, 132	4.5	ACCESSIBLE		
120	4.53	BURIED		
120	4.56	BURIED		
120	4.57	BURIED	-5	
120, 132	4.59	ACCESSIBLE	-3	
120	4.61	ACCESSIBLE	-1	
120	4.62	ACCESSIBLE	0	
120	4.62(+2)	ACCESSIBLE	0	
118	4.62(+12)-4.62(+15)	PARTIALLY_ACCESSIBLE	+3	
112	5.38	BURIED	0	

112, 115, 130	5.39	BURIED	0	
112, 115, 116, 130, 131	5.42	ACCESSIBLE	0	
112	5.46	BURIED	+3	
132	5.50	BURIED		
115, 132	6.48	ACCESSIBLE		
115, 122, 127, 129, 130, 131	6.51	ACCESSIBLE		
113, 115, 116, 129, 133	6.52	ACCESSIBLE		
121	7.35	BURIED	0	
120, 121, 130	7.39	ACCESSIBLE	0	
120, 124	7.42	BURIED	0	
120, 130	7.43	ACCESSIBLE	0	

Serotonin receptor

Reference #	Position(s)	Accessibility	Proximity_to family motifs	Proximity to subfamily motifs
141	1.46	BURIED		
148, 155, 156, 163	2.50	BURIED		
160	2.64	BURIED		
137, 148, 152, 153, 155, 156, 163	3.32	ACCESSIBLE		
134, 147, 155	3.36	ACCESSIBLE		
163	3.49	ACCESSIBLE		
161	4.50	ACCESSIBLE		
156	4.53	BURIED		-5
145	4.61	ACCESSIBLE		
144, 148	5.42	ACCESSIBLE	0	
148, 150, 156, 162	5.43	ACCESSIBLE	0	+5
134, 135, 136, 149, 151	5.46	BURIED	0	0
162	5.47	ACCESSIBLE	0	
162	5.48	ACCESSIBLE	+1	
155	5.49	ACCESSIBLE	+3	
155	5.52	ACCESSIBLE	+5	
157	5.59-5.61(+1)	PARTIALLY_ACCESSIBLE		0
157	5.61(+2)	ACCESSIBLE		0
157	5.61(+3)-5.61(+4)	ACCESSIBLE		0
157	5.61(+5)	ACCESSIBLE		0
157	5.61(+6)	ACCESSIBLE		0
157	5.61(+7)-5.61(+10)	NON_STRUCTURE_RES		0
143, 155, 161	6.48	ACCESSIBLE	-3	
139, 140, 156, 160	6.51	ACCESSIBLE	0	
139, 140, 144, 145, 155, 156, 160, 161	6.52	ACCESSIBLE	0	
144, 155, 156	6.55	BURIED	0	-3
141	7.34	ACCESSIBLE		
143	7.36	ACCESSIBLE		

142, 146, 154, 158, 159	7.39	ACCESSIBLE		
161	7.40	ACCESSIBLE		
156, 161	7.43	ACCESSIBLE		+5
138	7.46	BURIED		
155	7.49	ACCESSIBLE	-5	

Appendix 7. Proximity to Family- and Subfamily-Level Motifs for G Protein-Coupling Sites

Publications for the G protein-coupling regions listed below can be found in Appendix 3. All positions were determined based on bovine rhodopsin structure 1F88.

Since accessibility was determined according to 1F88, positions not mapped to 1F88 sequence in the same MSA were denoted as NON_STRUCTURE_RES.

Proximity scores:
0 = in motif; -1 = one residue left to motif; +1 = one residue right to motif; -3 = within three residues left to motif; +3 = within three residues right to motif; -5 = within five residues left to motif; +5 = within five residues right to motif

Adrenergic receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
171	3.49-3.51	PARTIALLY_ACCESSIBLE		
166, 171	3.52-3.55(+2)	ACCESSIBLE	-3	
166, 167, 170, 171	3.55(+3)	ACCESSIBLE	-3	
166, 171	3.55(+4)-4.41	ACCESSIBLE	0	
171	4.42	BURIED	+1	
169	5.55-5.60	ACCESSIBLE	0	
165, 166, 168, 169	5.61-5.61(+2)	PARTIALLY_ACCESSIBLE	0	
165, 166, 168, 169	5.61(+3)-5.61(+4)	ACCESSIBLE	0	
165, 166, 168, 169, 175	5.61(+5)-5.61(+10)	ACCESSIBLE	0	-3
165, 166, 168, 169, 175	5.61(+11)-5.61(+13)	NON_STRUCTURE_RES		0
166, 168, 169, 175	5.61(+14)-5.61(+16)	NON_STRUCTURE_RES		0
166, 168	5.61(+17)	NON_STRUCTURE_RES		0
168	5.61(+18)-5.61(+26)	NON_STRUCTURE_RES		
168	5.61(+27)-5.61(+28)	NON_STRUCTURE_RES		
164, 167, 168	5.61(+29)	NON_STRUCTURE_RES		-5
164, 168	5.61(+30)-5.61(+32)	NON_STRUCTURE_RES		-3
164, 167, 168	5.61(+33)	NON_STRUCTURE_RES		-1
164, 168	5.61(+34)	NON_STRUCTURE_RES		0
168	5.61(+35)-6.28(-33)	NON_STRUCTURE_RES		0
168	6.28(-32)-6.28(-7)	ACCESSIBLE		0
166, 168, 174	6.28(-6)	ACCESSIBLE		+5

166, 168	6.28(-5)	ACCESSIBLE		
166, 168, 174	6.28(-4)	ACCESSIBLE		
166, 168, 169	6.28(-3)	ACCESSIBLE		0
166, 168, 169, 174	6.28(-2)	ACCESSIBLE		0
166, 168, 169	6.28(-1)-6.28	ACCESSIBLE		0
166, 168, 169, 172	6.29-6.30	ACCESSIBLE		0
164, 166, 168, 169, 172, 174	6.31	ACCESSIBLE		+1
166, 168, 169, 172, 174	6.32	ACCESSIBLE		+3
168, 169, 172	6.33	ACCESSIBLE		+3
164, 168, 169, 172, 176	6.34	BURIED		+5
168, 169, 172	6.35	ACCESSIBLE		+5
168	6.36	BURIED		
169, 172	7.54-7.57	PARTIALLY_ACCESSIBLE	-3	
164	7.58	ACCESSIBLE	-1	
164, 169	7.59-7.67	PARTIALLY_ACCESSIBLE	0	-3
164	7.68-7.69(+1)	ACCESSIBLE	0	
173	7.69(+9)	ACCESSIBLE		0

Chemokine receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
180	3.49	ACCESSIBLE		
180	3.5	BURIED		
180	3.51	ACCESSIBLE	-5	
180	3.55(+4)	ACCESSIBLE	0	
178	5.61(+2)-5.61(+4)	ACCESSIBLE		0
178, 179	5.61(+5)-6.28(-6)	ACCESSIBLE		0
178	6.28(-5)-6.35	PARTIALLY_ACCESSIBLE		0
181	7.64	ACCESSIBLE	0	
180, 181	7.65-7.69(+7)	ACCESSIBLE	0	-3
180	7.69(+8)-7.69(+39)	ACCESSIBLE		0

Interleukin-8 receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
183	3.51	ACCESSIBLE		
183	3.52	ACCESSIBLE		
183	3.54	ACCESSIBLE		
183	3.55	ACCESSIBLE		
184	4.38	ACCESSIBLE		
184	4.39	ACCESSIBLE		
183	6.34	BURIED		

182	7.56-7.63	PARTIALLY_ACCESSIBLE	0	

Dopamine receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
189	5.46	BURIED		
188	5.57	ACCESSIBLE		
187, 188	5.58	ACCESSIBLE		
187, 188	5.59-5.61(+14)	PARTIALLY_ACCESSIBLE		0
187	5.61(+15)-6.28(-11)	NON_STRUCTURE_RES		0
186, 187	6.28(-10)-6.28(-7)	ACCESSIBLE		0
186, 187, 188	6.28(-6)-6.29	ACCESSIBLE		0
187, 188	6.30-6.35	PARTIALLY_ACCESSIBLE		+3
187, 188	6.36	BURIED		
188	6.37-6.39	PARTIALLY_ACCESSIBLE		
185	7.69(+6)-7.69(+33)	ACCESSIBLE		0

Muscarinic receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
197	2.5	BURIED		
197, 202	3.49	ACCESSIBLE	-3	
202	3.50-3.52	PARTIALLY_ACCESSIBLE	0	
190, 202	3.53	ACCESSIBLE	0	
202	3.54-3.55	ACCESSIBLE	0	
190, 202	3.55(+1)	ACCESSIBLE	0	
202	3.55(+2)	ACCESSIBLE	0	
202, 205	3.55(+3)	ACCESSIBLE	0	
202	3.55(+4)-4.38(-5)	ACCESSIBLE	0	
190, 202	4.38(-4)	ACCESSIBLE	0	
202	4.38(-3)-4.40	ACCESSIBLE	0	
190, 202	4.41	ACCESSIBLE	+3	
192, 198, 202	5.6	ACCESSIBLE		
191, 198, 202	5.61	BURIED		
190, 191, 194, 195, 198, 199, 202, 206	5.61(+1)	ACCESSIBLE		
198, 202	5.61(+2)	ACCESSIBLE		
196, 198, 202	5.61(+3)	ACCESSIBLE		
198, 199, 202	5.61(+4)	ACCESSIBLE		
198, 202	5.61(+5)-5.61(+6)	ACCESSIBLE		
194, 195, 198, 202	5.61(+7)	NON_STRUCTURE_RES		

198, 202	5.61(+8)-5.61(+19)	NON_STRUCTURE_RES		0
198	5.61(+20)-5.61(+58)	NON_STRUCTURE_RES		-5, 0
198, 207	5.61(+59)	NON_STRUCTURE_RES		-5, 0
198	5.61(+60)	NON_STRUCTURE_RES		0, -3
198, 207	5.61(+61)	NON_STRUCTURE_RES		0, -3
198	5.61(+62)-5.61(+70)	NON_STRUCTURE_RES		0, -3
198, 207	5.61(+71)-5.61(+76)	NON_STRUCTURE_RES		+5, 0, +3
198	5.61(+77)-6.31	ACCESSIBLE		0
198	6.32	ACCESSIBLE		
190, 195, 198, 199, 201, 202, 203	6.33	ACCESSIBLE		
190, 193, 195, 198, 199, 201, 202, 203	6.34	BURIED		
198	6.35	ACCESSIBLE		
190, 201, 202, 203	6.37	BURIED		
190, 202, 203	6.38	ACCESSIBLE		
204	6.39	ACCESSIBLE		
200	6.59	ACCESSIBLE	-1	

Serotonin receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
216	3.54	ACCESSIBLE		
216	3.55(+1)	ACCESSIBLE		
216	3.55(+4)	ACCESSIBLE		
210	3.55(+5)-4.38(-1)	ACCESSIBLE		
208, 209, 211	4.38	ACCESSIBLE		
210	4.39-4.41	ACCESSIBLE		
214	5.59-6.29	PARTIALLY_ACCESSIBLE		0
213, 214	6.3	ACCESSIBLE		
214	6.31	ACCESSIBLE		
214	6.32	ACCESSIBLE		
212, 214	6.33	ACCESSIBLE		
212, 214	6.34	BURIED		
214	6.35-6.36	PARTIALLY_ACCESSIBLE		
215	7.69(+1)	ACCESSIBLE		
215	7.69(+4)	ACCESSIBLE		

Appendix 8. Proximity to Family- and Subfamily-Level Motifs for Oligomerization Sites

Publications for the oligomerization regions listed below can be found in Appendix 4. All positions were determined based on bovine rhodopsin structure 1F88.

Since accessibility was determined according to 1F88, positions not mapped to 1F88 sequence in the same MSA were denoted as NON_STRUCTURE_RES.

Proximity scores:
0 = in motif; -1 = one residue left to motif; +1 = one residue right to motif; -3 = within three residues left to motif; +3 = within three residues right to motif; -5 = within five residues left to motif; +5 = within five residues right to motif

Adrenergic receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
222	1.28(-39)	NON_STRUCTURE_RES		0
222	1.28(-19)	ACCESSIBLE		0
222	1.28(-15)	ACCESSIBLE		0
217, 218, 221	1.32-1.58	PARTIALLY_ACCESSIBLE		+1
217	3.55(+3)	ACCESSIBLE	-3	
218	4.42-4.45	PARTIALLY_ACCESSIBLE	+1	
218, 220	4.46	ACCESSIBLE	+5	
218, 220	4.47	ACCESSIBLE		
218	4.48-4.62	PARTIALLY_ACCESSIBLE		
219	6.38	ACCESSIBLE		
219	6.42	ACCESSIBLE		
219	6.46	ACCESSIBLE		
221	7.34-7.67	PARTIALLY_ACCESSIBLE	0	

Chemokine receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
223	1.33-1.53	PARTIALLY_ACCESSIBLE	-5, 0	
223, 225	1.54	ACCESSIBLE	-3	
223	1.55-1.60	PARTIALLY_ACCESSIBLE	0	
224	2.46-2.65	PARTIALLY_ACCESSIBLE	+5, 0	+1, -1

223	4.39-4.46	PARTIALLY_ACCESSIBLE	0	
223, 225	4.47	ACCESSIBLE	+1	
223	4.48-4.62(+1)	PARTIALLY_ACCESSIBLE	+3	-1
224	5.40-5.59	PARTIALLY_ACCESSIBLE		-5, 0
224	6.37-6.57	PARTIALLY_ACCESSIBLE		0

Dopamine receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
229	4.42-4.57	PARTIALLY_ACCESSIBLE		
228, 229	4.58	ACCESSIBLE		
229	4.60-4.62(+3)	PARTIALLY_ACCESSIBLE		
226	5.36-5.59	PARTIALLY_ACCESSIBLE		
226	5.60-5.61(+4)	PARTIALLY_ACCESSIBLE		
226, 227	5.61(+5)-5.61(+10)	ACCESSIBLE		
226	5.61(+11)-6.35	PARTIALLY_ACCESSIBLE		0
226, 230	6.36-6.59	PARTIALLY_ACCESSIBLE		0
226	6.60-7.32	ACCESSIBLE	0	0
230	7.33-7.56	PARTIALLY_ACCESSIBLE	0	0
231	7.69(+51)	NON_STRUCTURE_RES		+5
231	7.69(+58)-7.69(+59)	NON_STRUCTURE_RES		-5

Histamine receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
232	3.55(+3)	ACCESSIBLE		

Muscarinic receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
233	3.25	BURIED		
233	5.34(-12)	ACCESSIBLE	-5	

Appendix 9. Proximity to Family- and Subfamily-Level Motifs for Protein-Protein Interaction Sites

Publications for the protein-protein interaction regions listed below can be found in Appendix 5. All positions were determined based on bovine rhodopsin structure 1F88.

Since accessibility was determined according to 1F88, positions not mapped to 1F88 sequence in the same MSA were denoted as NON_STRUCTURE_RES.

Proximity scores:
0 = in motif; -1 = one residue left to motif; +1 = one residue right to motif; -3 = within three residues left to motif; +3 = within three residues right to motif; -5 = within five residues left to motif; +5 = within five residues right to motif

Adrenergic receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
244	3.49-3.55(+3)	PARTIALLY_ACCESSIBLE	-3	
235, 246	5.61-5.61(+6)	PARTIALLY_ACCESSIBLE	0	
235, 246	5.61(+7)	ACCESSIBLE	+3	
235, 246	5.61(+8)-5.61(+9)	NON_STRUCTURE_RES	+3	-5
235, 246	5.61(+10)	NON_STRUCTURE_RES	+5	-3
235, 246	5.61(+11)	NON_STRUCTURE_RES		-3
235, 246	5.61(+12)-5.61(+13)	NON_STRUCTURE_RES		0
235	5.61(+14)-5.61(+20)	NON_STRUCTURE_RES		
246	5.61(+77)-6.28(-21)	NON_STRUCTURE_RES		0
235, 246	6.28(-20)-6.28(-8)	ACCESSIBLE		0
235, 246	6.28(-7)	ACCESSIBLE		0
235, 246	6.28(-6)	ACCESSIBLE		0
235, 246	6.28(-5)	ACCESSIBLE		0
235, 246	6.28(-4)	ACCESSIBLE		+1
235, 246	6.28(-3)-6.28	ACCESSIBLE		+3
235, 246	6.29	ACCESSIBLE		
235, 246	6.3	ACCESSIBLE		
235, 246	6.31	ACCESSIBLE		
235, 246	6.32	ACCESSIBLE		
246	6.33-6.38	PARTIALLY_ACCESSIBLE		
242	6.47	BURIED		
248	7.49-7.56	PARTIALLY_ACCESSIBLE	-3	
1997, 243, 248	7.57-7.69(+8)	PARTIALLY_ACCES	0	0

		SIBLE		
1997, 236, 243, 248	7.69(+9)	ACCESSIBLE		-5, 0, +3
1997, 236, 243, 248	7.69(+10)-7.69(+16)	ACCESSIBLE		0, -1
1997, 243, 248	7.69(+17)-7.69(+56)	ACCESSIBLE		0, -3
1997, 243, 248	7.69(+57)-7.69(+73)	NON_STRUCTURE_RES		+1, +5, 0
1997, 237, 243, 248	7.69(+74)	NON_STRUCTURE_RES		+5, 0
234, 237, 243, 248	7.69(+75)	NON_STRUCTURE_RES		+5, -5, 0
Klen et al, 1997, 248	7.69(+76)	NON_STRUCTURE_RES		-5, 0
Klen et al, 1997, 237, 248	7.69(+77)	NON_STRUCTURE_RES		0, -3
248	7.69(+78)-7.69(+82)	NON_STRUCTURE_RES		+1, 0
238, 239, 240, 247, 248	7.69(+83)	NON_STRUCTURE_RES		0
238, 239, 240, 241, 245, 247, 248	7.69(+84)	NON_STRUCTURE_RES		0
239, 248	7.69(+85)	NON_STRUCTURE_RES		+1, 0
238, 239, 240, 245, 247, 248	7.69(+86)	NON_STRUCTURE_RES		0, +3
248	7.69(+87)-7.69(+124)	NON_STRUCTURE_RES		-5, 0, +3
248	7.69(+125)-7.69(+154)	NON_STRUCTURE_RES		0, +3
248	7.69(+155)	NON_STRUCTURE_RES		+5

Chemokine receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
261	3.49	ACCESSIBLE		
261	3.50	BURIED		
261	3.51	ACCESSIBLE	-5	
255	5.60-6.36	PARTIALLY_ACCESSIBLE		0
268	7.44-7.56	PARTIALLY_ACCESSIBLE	0	
262, 268	7.57	ACCESSIBLE	0	
262, 265, 268, 270	7.58	ACCESSIBLE	0	
262, 265, 268, 270	7.59-7.69	PARTIALLY_ACCESSIBLE	0	-5, 0
255, 262, 265, 268, 270	7.69(+1)-7.69(+22)	ACCESSIBLE	+1	0
255, 261, 262, 263, 265, 268, 270	7.69(+23)	NON_STRUCTURE_RES		+1, 0
255, 261, 262, 263, 265, 268, 270	7.69(+24)	NON_STRUCTURE_RES		+1, 0, +3
255, 262, 265, 268, 270	7.69(+25)-7.69(+28)	NON_STRUCTURE_RES		0, +3
255, 261, 262, 263, 265, 268, 270	7.69(+29)	NON_STRUCTURE_RES		0
255, 262, 265, 268, 270	7.69(+30)-7.69(+35)	NON_STRUCTURE_RES		0
261, 262, 263, 265, 268, 270	7.69(+36)	NON_STRUCTURE_RES		+3

262, 265, 270	7.69(+37)-7.69(+39)	NON_STRUCTURE_RES		
262, 265	7.69(+40)-7.69(+41)	NON_STRUCTURE_RES		

Interleukin-8 receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
274	7.59-7.63	PARTIALLY_ACCESSIBLE	0	
274, 275	7.64	ACCESSIBLE	0	
275	7.65	ACCESSIBLE	0	
275, 276	7.66-7.68	ACCESSIBLE	0	
276	7.69-7.69(+5)	ACCESSIBLE	0	-5

Dopamine receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
278	5.57-5.58	ACCESSIBLE		
278, 283	5.59	ACCESSIBLE		
278, 283	5.6	ACCESSIBLE		
278, 283	5.61	BURIED		
278	5.61(+1)-5.61(+2)	ACCESSIBLE		
278, 284	5.61(+3)-5.61(+12)	ACCESSIBLE		
278	5.61(+13)-5.61(+14)	NON_STRUCTURE_RES		
281	6.28(-12)	NON_STRUCTURE_RES		+5
282	7.55	ACCESSIBLE	0	
277, 282	7.56	ACCESSIBLE	0	
282	7.57-7.58	ACCESSIBLE	0	
279, 282	7.59	ACCESSIBLE	0	
277, 279, 282	7.6	BURIED	0	
279, 282	7.61	ACCESSIBLE	0	
282	7.62-7.63	PARTIALLY_ACCESSIBLE	0	
277, 282	7.64	ACCESSIBLE	0	
282	7.65	ACCESSIBLE	0	
279, 282	7.66	ACCESSIBLE	0	
282	7.67-7.69	ACCESSIBLE	0	
279, 282	7.69(+1)	ACCESSIBLE	+1	
282	7.69(+2)-7.69(+40)	ACCESSIBLE	+3	0
280, 282	7.69(+41)-7.69(+70)	NON_STRUCTURE_RES		0
280, 282	7.69(+71)-7.69(+100)	NON_STRUCTURE_RES		0
282	7.69(+101)-7.69(+103)	NON_STRUCTURE_RES		

Muscarinic receptor

Reference #	Position(s)	Accessibility	Proximity to	Proximity

			family motifs	to subfamily motifs
287	5.61(+55)-5.61(+115)	NON_STRUCTURE_RES		0
285	6.28(-73)	NON_STRUCTURE_RES		0
285	6.28(-71)-6.28(-69)	NON_STRUCTURE_RES		0, -3
287	6.28(-58)-6.28(-5)	ACCESSIBLE		0
286, 287	6.28(-4)-6.28	ACCESSIBLE		+5
287	6.29	ACCESSIBLE		
286, 287	6.3	ACCESSIBLE		
287	6.31-6.42	PARTIALLY_ACCESSIBLE		

Serotonin receptor

Reference #	Position(s)	Accessibility	Proximity to family motifs	Proximity to subfamily motifs
289	5.59-6.36	PARTIALLY_ACCESSIBLE		0
288, 290	7.69(+72)	NON_STRUCTURE_RES		0
288	7.69(+74)	NON_STRUCTURE_RES		+3