UPTEC X 06 031
JUN 2006

ISSN 1401-2138

HANNES BRÅBERG

Alignment of protein sequences/structures and its application to predicting protein complex compositions

Master's degree project

UPPSALA
UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 06 031

Date of issue 2006-06

Author

Hannes Bråberg

Title (English)

Alignment of protein sequences/structures and its application to predicting protein complex compositions

Title (Swedish)

Abstract

The SALIGN module of MODELLER is a newly developed general protein structure/sequence alignment tool. Described in the first half of this thesis is a web server that accesses SALIGN, to calculate pairwise and multiple alignments of the users' protein structures and/or sequences. The SALIGN server is available at http://salilab.org/salign.

The second half of this thesis presents structure-based predictions of 3,213 binary and 1,234 higher order protein complexes in *S. cerevisiae* involving 750 and 195 proteins, respectively. To generate candidate complexes, comparative models of individual proteins were built and combined together using complexes of known structure as templates. These candidate complexes were then assessed using a specialized statistical potential. Moreover, the predicted complexes were also filtered using functional annotation and sub-cellular localization data. Through integration with MODBASE, the application of the method to proteomes that are less well characterized than that of *S. cerevisiae* will contribute to expansion of the structural and functional coverage of protein interaction space.

Keywords
protein complexes, protein interaction prediction, complex structure assessment
sequence alignment, structure alignment, web interface, web server

Supervisors

M. S. Madhusudhan
University of California at San Francisco

Scientific reviewer

Gerard Kleywegt
Uppsala University

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **41**

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

Alignment of protein sequences/structures and its application to predicting protein complex compositions

Sammanfattning

Proteiner är de mest mångsidiga makromolekylerna i biologiska system och deltar i alla cellulära processer. Ett proteins funktionella egenskaper bestäms av dess tre-dimensionella struktur, vilken i sin tur dikteras av sekvensen av aminosyror som utgör proteinet. Härav följer att noggranna metoder för proteinstrukturbestämning är av yttersta vikt. Homologimodellering är en metod som effektivt predikterar okända proteinstrukturer genom att huvudsakligen förlita sig på deras “alignments”[1] till liknande proteiner med kända strukturer. Sekvens/struktur “alignments” är även viktiga i flera andra avseenden. SALIGN är en sekvens/struktur “alignment” modul som tillhanda-håller en stor mängd funktioner. Det första delprojektet i examens-arbetet bestod av att skapa ett web-baserat användargränssnitt till SALIGN, vilket torde underlätta kategoriseringen och studierna av proteinfamiljer.

Proteiner fungerar genom interaktioner med andra molekyler. Av detta inses att nätverket av fysiska interaktioner, proteiner emellan, är av stort intresse för biologer. I det andra delprojektet konstruerades en metod för att prediktera proteinkomplex-sammansättningar genom att generera homologimodeller av kandidatkomplex, baserat på sekvenslikhet till strukturellt kända komplex, följt av modellutvärdering. Metoden applicerades på *Saccharomyces cerevisiae* proteomet, vilket resulterade i struktur-baserade prediktioner av 3213 binära proteinkomplex och 1234 proteinkomplex av högre ordning, involverande 750 och 195 proteiner, respektive. Metodens applicering på mindre välkarakteriserade proteom kommer att bidra till expansionen av den strukturella och funktionella kartläggningen av proteininteraktioner.

[1] En sekvens/struktur alignment är en beskrivning av vilka aminosyror som motsvarar varandra i två eller flera proteiner (eller delar därav), baserat på sekvens, struktur, eller en kombination av de två.

Hannes Bråberg
Examensarbete, Molekylär Bioteknik, 2006
Uppsala Universitet

1 General background

1.1 Protein structure

Proteins carry out a wide variety of tasks in the cells and participate in all the cellular processes. They are the most versatile macromolecules in biological systems and the numerous roles of proteins include acting as enzymes, transmitting nerve impulses, controlling cell growth and differentiation and providing mechanical support and immune protection. They also transport and store other molecules, and generate movement of cells.

The functional properties of proteins are determined by their three-dimensional (3D) structures. The 3D structures are in turn dictated by the sequences of amino acids comprising the proteins. This ability to spontaneously fold into precise, complex structures serves as a direct link between the one-dimensional (1D) world of sequences and the 3D world of structure and function. It is an important feature that is crucial to the central role of proteins in biochemistry. Proteins are built up of linear chains of amino acid residues and can be described in four levels of structure (Berg *et al.*, 2002):

- Primary structure refers to the sequences of the polypeptide chains consisting of L-amino acids linked by peptide bonds. The polypeptide chains are linear and the peptide bonds are actually amide bonds formed between the carboxyl group of residue *n* and the amino group of residue *n+1* in the sequence. Peptide bonds possess a number of features that are essential to the structure and function of proteins. First, they are uncharged which allows the chains to pack tightly, forming compact structures. Second, the peptide bonds have significant double-bond character, which imposes some rigidity on the chains. Third, each peptide bond has a hydrogen bond donor as well as a hydrogen bond acceptor; this is an important feature for stabilizing the regular 3D structures of proteins. Finally, peptide bonds do not hydrolyze spontaneously, which results in proteins being kinetically stable under physiological conditions.

- Secondary structure refers to the local, regular structures of the polypeptide chain, such as alpha helices and beta strands. Alpha helices are sections where the polypeptide chain is tightly coiled and residue *n* is hydrogen bonded to residues *n-3* and *n+4* in the sequence. An alpha helix can be either right-handed or left-handed, depending on the direction of the coil. In general, L-amino acids cannot form left-handed alpha helices, due to steric hindrance. Consequently, alpha helices in proteins are almost always right handed. In contrast to the compact alpha helices, beta strands are sections where the chains are more or less fully extended. Beta sheets consist of two or more beta strands, alongside each other, with hydrogen bonds between them. A beta sheet can be either parallel or antiparallel. In a parallel sheet the residues in successive strands run in the same biochemical direction, and in an antiparallel sheet the residues in successive strands run in alternating directions.

- Tertiary structure describes the complete folding of one polypeptide unit, consisting of arranged sections of secondary structure. In aqueous milieus this folding usually results in compact structures with hydrophobic residues buried in the interior and hydrophilic residues on the surface. This arrangement is governed by the hydrophobic effect and allows the hydrophilic side chains to interact with the environment. Proteins in hydrophobic environments (membranes) usually display the inverse arrangement with hydrophilic residues sheltered in the core and hydrophobic residues on the surface. Besides the hydrophobic effect, salt links, hydrogen bonds and covalent disulfide links (between cysteine residues) stabilize the tertiary structure.

- Quaternary structure describes the arrangement of multiple polypeptide chains that form multi-subunit structures. Proteins can consist of one or many subunits. Subunits can be identical or different and are usually held together by non-covalent forces.

In this thesis, the term “structure” refers to three-dimensional structure (i.e., not primary structure) unless otherwise noted.

1.2 Protein structure modeling

1.2.1 Introduction

In light of the crucial roles played by proteins in biology, it is evident that developing methods for functional annotation of proteins is tremendously important. Accurate 3D structures of proteins are very useful for such processes, due to the strong connection between protein structure and function. Protein structures are best determined by experimental methods, such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. Experimental methods can, however, only be applied to a fraction of all proteins, for a number of reasons. Some proteins are especially difficult to analyze experimentally due to factors such as inability to crystallize etc., but furthermore the number of known proteins is far too large for it to be feasible to determine all structures experimentally. One of the prime motivations for developing protein structure modeling methods is the fact that the sequence databases are growing at a much higher rate than the database of experimentally determined structures. The number of experimentally determined structures deposited in the Protein Data Bank (PDB) increased from 23 096 to 31 823 over the last 2 years (August 2005) (Westbrook *et al.*, 2002). Over the same period, the number of sequences in comprehensive sequence databases, such as UniProt (Bairoch *et al.*, 2005) and GenPept (Benson *et al.*, 2005), increased from 1.2 to 2 million.

These issues emphasize the need for computational methods for predicting protein structures. There are two major classes of methods for computational modeling of protein structures (Madhusudhan *et al.*, 2005; Baker and Sali, 2001; Fiser *et al.*, 2002). Comparative methods, including comparative (or homology) modeling and threading, predict the structure of a protein by relying primarily on its alignment to at least one similar protein with known structure. *Ab initio* (or *de novo*) modeling methods model protein structures based on sequence information alone, but do not utilize any sequence similarity to known protein structures. *Ab initio* modeling is based on the assumption that the native state of a protein corresponds to the global free energy minimum in conformational space. These methods are based on the laws of physics and attempt to find the tertiary structure with the lowest possible free energy for a given sequence of amino acids. Such a procedure consists of two major components: an algorithm that efficiently carries out the conformational search and a free energy function used for evaluating the possible conformations. The accuracy and reliability of *ab initio* models are significantly lower than those of comparative models based on 30% or higher sequence identity (Madhusudhan *et al.*, 2005; Baker and Sali, 2001). Since 1994, a meeting on Critical Assessment of techniques for protein Structure Prediction (CASP, http://predictioncenter.gc.ucdavis.edu/) has been held every second year. Well in advance of each meeting, the participating groups are presented with a number of target proteins whose structures are about to be solved experimentally. Prior to the public release of the structures, predictions are collected from the participating groups. The categories include comparative modeling, threading and *ab initio* modeling. Independent assessors evaluate all predictions, and the results are released shortly before the meeting, during which the results and successful methods are presented and discussed. The aim of CASP is to establish the current state of the art in protein structure prediction, to identify what progress has been made, and to locate the areas in which future improvement efforts may be most profitable.

This thesis focuses on methods used in conjunction with comparative modeling.

1.2.2 Comparative modeling

Comparative modeling is based on statistical learning and utilizes the fact that evolutionary changes are gradual in order to preserve important functional features, which in turn requires the conservation of structure and, to a lesser extent, sequence. This process has resulted in families of related proteins that have similar sequences and structures, and sometimes even share functional features (Fiser *et al.*, 2002). The 3D structures of proteins within a family are more conserved than their sequences (Lesk and Chothia, 1980). Hence, if there is a significant degree of similarity between two proteins at the sequence level, this implies that they have similar 3D structures as well. The aim of comparative modeling is to generate a 3D model for a protein of unknown structure (the target), based on its sequence alignment to at least one similar protein of known structure (the template) (Marti-Renom *et al.*, 2000). Two conditions have to be met in order for such a process to be feasible. First, the target sequence must have detectable similarity to at least one protein of known structure, which will be used as a template. Second, it must be possible to compute a substantially correct alignment between the target sequence and the template structure. In general, comparative modeling consists of four steps: fold assignment and template selection, alignment of the target to the template(s), model building and model assessment (Fig. 1) (Madhusudhan *et al.*, 2005). The quality of a model is strongly related to the level of sequence identity between the target and the template, partly because higher sequence identity implies higher 3D structural similarity, partly because alignment accuracy increases with increasing sequence identity. High-accuracy models are based on templates to which they have more than 50% sequence identity. The root mean square (RMS) error for the main-chain atoms of these models is generally about 1 Å, which is comparable to that of low-resolution X-ray structures and medium-resolution NMR structures (Baker and Sali, 2001). Medium-accuracy models have 30-50% sequence identity to their templates and usually have 90% of the main-chain modeled with a RMS error of 1.5 Å. Finally, the low-accuracy models are those that have less than 30% sequence identity to their templates.

There exists a plethora of applications for comparative protein structure models. In general, modeling errors are relatively rare in functionally important regions of proteins, such as active sites and binding sites, since these regions are usually more conserved than the rest of the fold (Sanchez and Sali, 1998). Thus, from a perspective of function prediction, a comparative model can often provide more accurate information than its overall RMS error would suggest. The accuracy of a model determines the applications for which it is suitable (Fig. 2). Low-accuracy models are mostly used for fold assignment of proteins, and rarely provide any detailed information. Nevertheless, function can sometimes be predicted from only rough

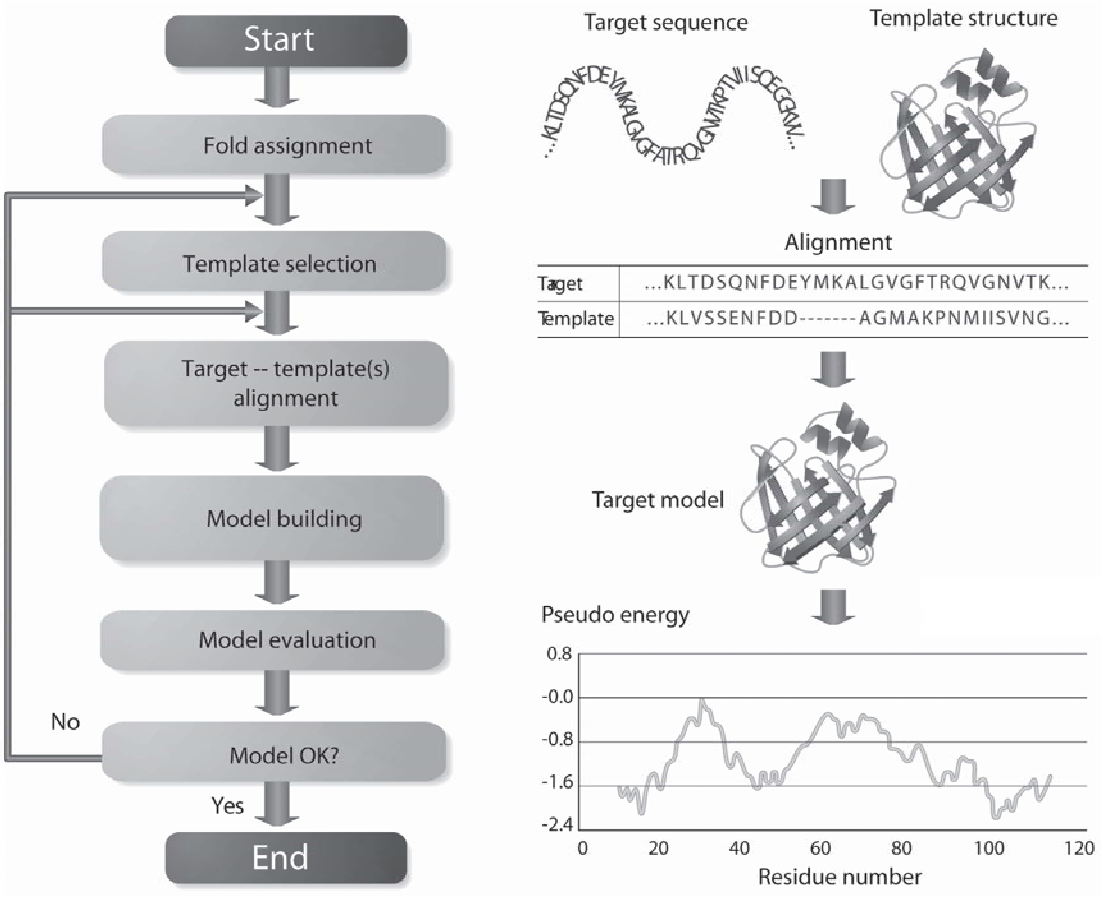


Fig. 1. *A flow chart of the steps involved in comparative protein structure modeling. First, all protein structures that are related to the target are identified (fold assignment) and the ones that are appropriate for the given modeling problem are selected as templates (template selection). The target sequence is then aligned to the selected templates (target—template(s) alignment) and a 3D model of the target is constructed (model building). Finally, the model is evaluated (model evaluation) and a decision is made whether to keep the model or start over from the template selection or alignment steps. (Adapted from Madhusudhan et al., 2005).*

structural features. Medium- and high-accuracy models are often used for improving functional predictions derived from sequence alone, since ligand binding is more directly determined by the structure of the binding site than by its sequence (Baker and Sali, 2001). It is often possible to predict features of a target protein that do not exist in its template. For example, the existence and location of a binding site can be predicted by searching for clusters of charged residues (for binding charged ligands) (Matsumoto *et al.*, 1995), and the volume of the binding site cleft provides information about the size of the corresponding ligand (Xu *et al.*, 1996). Medium- and high-accuracy models can also be used to design proteins with specific features and purposes. Examples of these are proteins with compact structures – lacking long tails, loops and exposed hydrophobic residues – for improved crystallization, and proteins containing extra disulphide bonds for enhanced stability. High-accuracy models are often of such good quality that they can be used for docking experiments, where small ligands (Ring *et al.*, 1993) or whole proteins (Vakser, 1995) are docked onto the modeled protein. Combining comparative modeling with other methods, such as electron microscopy, extends its use. For example, molecular models of large macromolecular assemblies can be produced by fitting comparative models of the constituent proteins into electron microscopy maps of the whole assemblies.

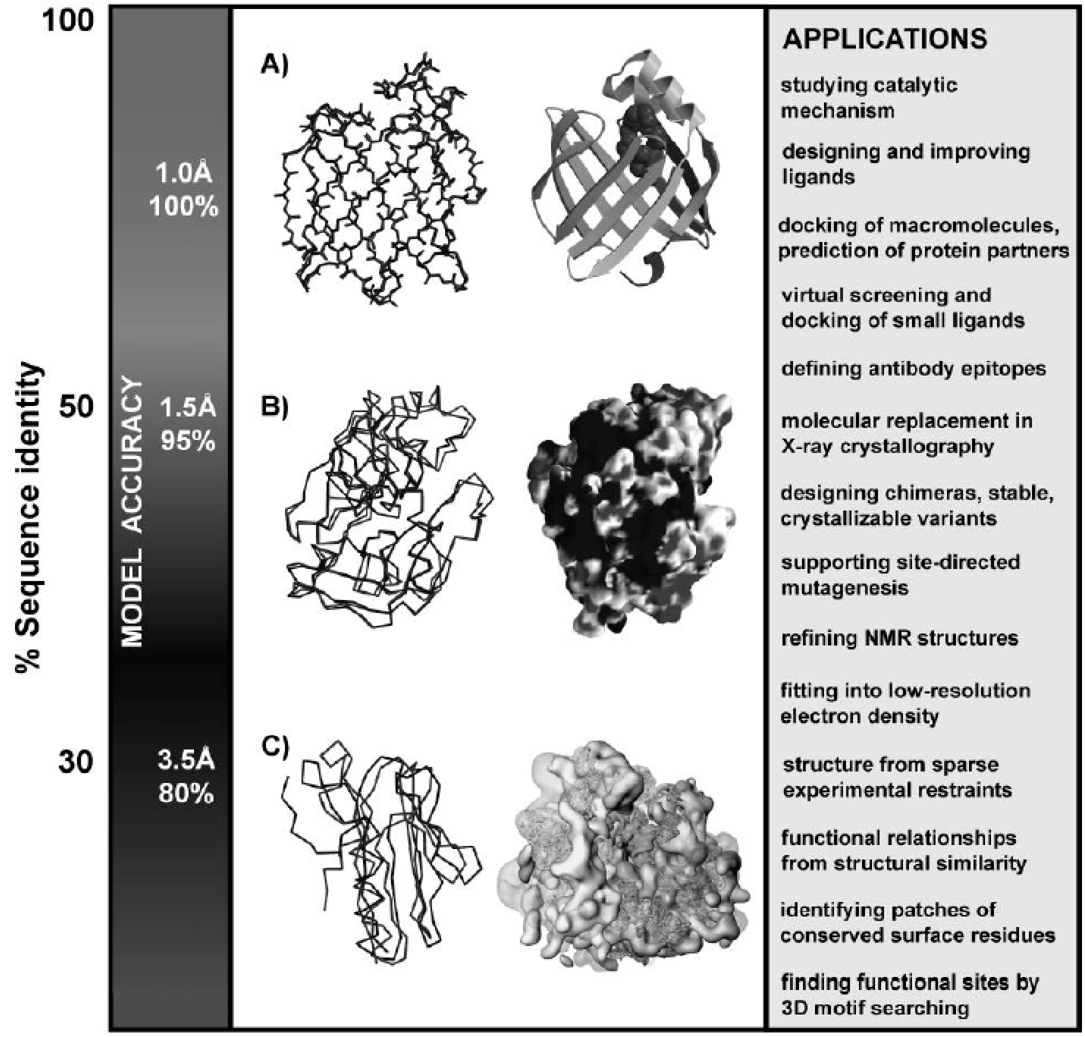

100
50
30
% Sequence identity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
A)
B)
C)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
structure from sparse experimental restraints
functional relationships from structural similarity
identifying patches of conserved surface residues
finding functional sites by 3D motif searching

Fig. 2. *Applications of comparative models. The applications of a comparative model depend on its accuracy, which is strongly correlated with the sequence identity between the model and its template(s). The vertical axis indicates the different ranges of applicability of comparative protein structure modeling, the corresponding accuracy of protein structure models, and examples of applications. (A) The docosahexanoic fatty acid ligand was docked into a high-accuracy model of brain lipid-binding protein (right) based on 62% sequence identity to the structure of adipocyte lipid-binding protein (Xu et al., 1996). A number of fatty acids were ranked for their affinity to brain lipid-binding protein, and the results were consistent with experimental methods, even though the ligand specificity profiles differ between this protein and its template (left). (B) A medium accuracy model of mouse mast cell protease 7 (right), modeled based on 39% sequence identity to the structure of bovine pancreatic trypsin. A putative proteoglycan binding patch was identified on the model, even though its template does not bind proteoglycans (Matsumoto et al., 1995). The prediction was confirmed by experimental methods. (C) A molecular model of the complete yeast ribosome (right) was constructed by fitting atomic rRNA and protein models into the electron density of the 80S ribosomal particle, obtained by electron microscopy at 15 Å resolution (Spahn et al., 2001). (Adapted from Fiser et al., 2000).*

Currently, automated comparative modeling, generating reliable models, is possible for domains in about 60% of the approximately 1.8 million unique protein sequences in the Universal Protein Resource (UniProt) database (July 5, 2005) (Pieper *et al.,* 2006; Bairoch *et al.,* 2005). However, roughly two thirds of these models have less than 30% sequence identity to their best template and are likely to contain significant errors. At such low sequence identities, target-template alignment errors are common and they constitute the major error source in low-accuracy models. At present, there is no comparative modeling program that can recover from an incorrect target-template alignment. A substantial effort is thus invested in constructing more sophisticated structure-sequence alignment methods and making modeling less dependent on the input alignments.

A factor contributing greatly to the importance of comparative modeling is its role in structural genomics, which aims to structurally annotate most protein sequences utilizing a combination of experiment and prediction (Baker and Sali, 2001). The first step of structural genomics is to carefully select a set of target proteins that will be structurally characterized by X-ray crystallography or NMR spectroscopy. There are a number of target selection schemes, ranging from studying only proteins that are likely to have novel folds to selecting all the proteins of a model genome. In a model-centric view, the targets for experimental structure determination should be selected such that most remaining protein sequences are closely related to at least one of the solved structures. In this way, accurate comparative models can be built for a majority of all proteins, based on a relatively small number of experimentally solved structures. It is desirable that all of these model-template pairs pass a 30% sequence identity cutoff, due to the rapid decrease in model accuracy below it. It has been estimated that this cutoff requires a minimum of 16,000 experimental targets in order to cover 90% of all protein domain families (Vitkup *et al.*, 2001). The experimental characterization of these 16,000 structures will allow the modeling of a very much larger number of proteins. For example, the New York Structural Genomics Research Consortium (http://www.nysgxrc.org/) found that each of their new solved structures on average allowed roughly 100 proteins, with previously unknown structures, to be modeled at least at the fold level. This illustrates the importance of comparative modeling in large-scale structure characterization efforts.

2 Designing a web interface to the MODELLER sequence/structure alignment module SALIGN

2.1 Introduction

2.1.1 Protein sequence/structure alignments

As discussed above, determining the structure of a protein and characterizing its function are crucial steps for obtaining a better understanding of cellular processes. To achieve these aims, it is important that robust methods are employed to compare or align protein sequences and structures with one another. Such methods are frequently used for inferring the function of a newly sequenced protein by analogy to previously characterized proteins (Koehl, 2001). Classifying proteins into structural families often requires pairwise and multiple structural superimpositions (Andreeva *et al.*, 2004; Holm and Sander, 1999). To build models of a protein (target) based on homology to other proteins of known structures (templates), it is vital to correctly align the sequence of the target protein to those of the templates (Marti-Renom *et al.*, 2000) (see section 1.2.2). Conserved and variable regions of sequences can be identified by studying the corresponding segments of many aligned proteins. These are but some examples of the applicability of protein sequence/structure alignment methods. Methods for aligning sequences or structures follow the same general principles, and the alignments are constructed in analogous manners.

Sequence/structure alignment refers to the assignment of residue-residue correspondences between two or more proteins (or sections thereof), based on sequence alone, structure alone, or a combination of sequence and structure. Any such assignment, where the sequential order of residues within each protein is preserved, is an alignment. The objective of an alignment program is to find the best possible alignment for a given set of sequences/structures. In such a process, a system for scoring the alignments is crucial. A variety of scoring schemes have been invented and implemented for different types of alignments.

2.1.2 Sequence-sequence alignments

A simple type of scoring scheme is that used for pairwise sequence-sequence alignments. Such a scoring scheme reflects the similarity between the aligned sequences, based on the number and types of editing operations required to transform one sequence into the other. The rationale behind the use of such a measure lies in the fact that these editing operations mimic the natural events that take place during evolution and cause sequences of common ancestry to diverge. There are two distinct types of events – substitutions and deletions/insertions. A scoring function should punish rare substitutions and reward those that are likely (as well as conservations) and correspondingly favor some identities more than others. This is implemented by introducing a substitution matrix, which contains the substitution and match scores for all possible residue-residue combinations. Insertions and deletions are accounted for by introducing a gap penalty; a cost for matching a residue in one sequence with a gap in another. The simplest gap penalty functions are directly proportional to the gap lengths, whereas affine functions penalize the opening of a gap more than its elongation. Given a substitution matrix and a gap penalty function, a score can be calculated for any pair of aligned sequences. The similarity of two sequences, **X** and **Y**, comprised of residues $x_1,\ldots,x_{N1}$ and $y_1,\ldots,y_{N2}$ respectively, is defined as:

$$sim(\mathbf{X},\mathbf{Y}) = \max_{\substack{\textit{all alignments}\\ \textit{between } \mathbf{X}\&\mathbf{Y}}} score(\mathbf{X},\mathbf{Y}) \qquad \mathbf{X} = x_1,\ldots,x_{N_1},\ \mathbf{Y} = y_1,\ldots,y_{N_2}$$

An alignment that produces the maximum score is called an optimal alignment. The original, and still widely used, method for finding an optimal alignment is based on a mathematical technique called dynamic programming (Needleman and Wunsch, 1970; Sellers, 1974). The dynamic programming algorithm guarantees to find the global optimum, and thus the best alignment, with respect to the utilized scoring function. It should, however, be noted that many alignments can have the same "optimal" score and that none of these necessarily have to correspond to the evolutionarily correct alignment. The dynamic programming algorithm calculates the optimal alignment score recursively, utilizing the fact that the total alignment score is a sum of the scores for all positions. With time, the scoring function and its optimization have been improved, resulting in increased accuracy and speed (Marti-Renom *et al.*, 2004). Furthermore, they have been extended and applied to a variety of alignment problems. Most of these improved methods are based on the same general principles as the simple approach described above, even though specific steps of the procedures vary greatly.

One of the most significant improvements in alignment accuracy was achieved through the use of sequence profiles (Gribskov *et al.*, 1987, 1990; Gribskov, 1994). A sequence profile is calculated from a multiple sequence alignment (MSA) of related sequences and specifies a preference for each of the 20 standard amino acid residue types at each position in the alignment. A MSA may, however, not contain enough homologs to calculate a statistically robust profile solely from the distribution of residue types in the MSA. In order to circumvent this problem, a number of estimation schemes have been suggested, most of which depend on *prior* or expected probabilities of residue occurrences and/or residue-residue substitutions. Profiles are valuable for detecting remote homologs in the so-called "twilight zone", where the sequence identity between the proteins is lower than 30% (Sadreyev *et al.*, 2003). Furthermore, the use of profiles increases the accuracy of "twilight zone" alignments significantly. This is of great importance for comparative modeling and is reflected in the accuracy and extent of the resulting models. Today, methods exist for sequence-profile alignments as well as profile-profile alignments, which have been shown to be more sensitive than the former (Madhusudhan *et al.*, 2005).

2.1.3 Sequence-structure alignments

Another approach that increases the accuracy of alignment methods significantly is the incorporation of structural information about one of the sequences in a pairwise comparison. One such method is threading (Torda, 1997), where fold assignment and alignment are attained by threading a sequence through each of the structures in a library of all known folds. Each such sequence-structure alignment is assessed by the energy of a corresponding coarse model, without taking sequence similarity into account.

Yet another approach, which lies between purely sequence-based methods and threading methods, is to incorporate structural information into profile alignment methods. This is implemented by making the substitution scores depend on solvent accessible surface area, secondary structure type, hydrogen bonding properties etc. (Luthy *et al.*, 1992). Further enhancement of this approach is possible by extending the use of structural data to the sequence side of the structure-sequence pair. This can be achieved by making use of the predicted local structure of the sequence (Tang *et al.*, 2003). Further improvement of the accuracy can be achieved by adjusting gap penalties according to the local environment in which the gaps occur (Zhu *et al.*, 1992).

2.1.4 Structure-structure alignments

Structure-structure alignment methods can usually align proteins in the "twilight zone" much more accurately than sequence based methods. This is due to the fact that 3D structures of proteins in the "twilight zone" are more conserved than their sequences. The most direct approach for comparing two structures is to superimpose them as rigid bodies and look for equivalent residues (Koehl, 2001). This approach is however limited to structures that are relatively similar, as it will not be able to detect local similarities between structures that differ on the global level. Breaking the structures into fragments solves this problem, but can lead to situations where the global alignment is missed instead. Recent work has been focused on methods satisfying both the global and local criteria (Koehl, 2001). A majority of the structure-structure scoring schemes are based solely on the geometrical properties of the sets of points that represent the structures, ignoring information about the local environment of the residues. Even though most of these are far more complicated than the root mean square (RMS) deviation, this remains the general measure for describing the similarity of two protein structures. Two types of RMS measures have been proposed, *cRMS* and *dRMS*. The *cRMS* provides a measure for the distance between the coordinate sets of two superimposed structures:

$$cRMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\|\mathbf{x}(i) - \mathbf{y}(i)\|^2\right)}$$

where N is the number of atoms to be compared, **x**(*i*) is the coordinate vector for atom *i* in one of the structures, and **y**(*i*) is the corresponding coordinate vector for the other structure. *dRMS*, on the other hand, compares the intramolecular distances between two structures:

$$dRMS = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(d_{ij}^{A} - d_{ij}^{B}\right)^2}$$

where d_{ij}^{A} is the distance between atoms *i* and *j* in one of the structures, and d_{ij}^{B} is the distance between the corresponding atoms in the other structure. Both RMS measures are based on the Euclidian norm and thus very sensitive to outliers, which limits their efficacy to closely related structures. For example, consider two distantly related proteins with similar structures of the core regions, but major differences in their loop geometries. In such a case, a RMS measure could favor a poor alignment, where all regions of the proteins were relatively close to each other, rather than one where the core regions were well aligned and the loops were far away from each other. An important complement to the RMS measure is the structural overlap, or equivalent positions measure. This estimates the number of equivalent residue atoms (*e.g.* Cα) that lie within a certain cut-off distance. A number of other methods, some less sensitive to outliers than others, have been proposed, but none of them appears to be ideal for all scenarios. Koehl (Koehl, 2001) argues that the problem of structure comparison is ill posed and that additional information is required to characterize a problem with a well-defined solution. He exemplifies this by fold recognition applications, which focus more on the conserved core regions of the proteins than loop geometry. For such situations, he suggests defining a similarity score that only includes atoms in the core.

2.1.5 SALIGN

The multi-purpose alignment module of MODELLER (Sali and Blundell, 1993), SALIGN (Madhusudhan *et al.*, in preparation), is capable of aligning sequences, structures, or a combination of the two. It is loosely based on the algorithms used by the program COMPARER (Sali and Blundell, 1990). All pair-wise alignments are calculated using global or local dynamic programming methods. The weight matrix used in the dynamic programming consists of a combination of weighted scores contributed from 6 different sequence and structure features (Fig. 3). The features include 1) residue-residue substitution score, 2) root mean square deviation (RMSD) of chosen atoms of residues, 3) fractional side chain solvent accessibility, 4) secondary structure type, 5) local similarity as reflected in the distance RMSD, and 6) any user created input matrix. Features 2-5 are useful in structure alignments while feature 1 is useful to align sequences. SALIGN provides two distinct methods, “tree alignment” and “progressive alignment”, for generating multiple alignments. The tree algorithm first creates a dendrogram of the structures/sequences from a matrix of all pairwise alignment scores. Guided by the dendrogram, the tree multiple alignment is then constructed, by aligning the closest linked branches to each other (Fig. 3). The progressive alignment algorithm is simpler and less computationally expensive. This approach begins with the alignment of two arbitrary sequences to each other, followed by the alignment of a third sequence to the first two; and in *n-1* steps, a multiple alignment of *n* sequences is created. If two pre-aligned blocks of sequences are to be aligned, the profile-profile alignment method is used. To align a block of sequences to a block of structures, the Align2D algorithm (Madhusudhan *et al.*, 2006) is used. Align2D uses local or global dynamic programming but replaces the affine gap penalties with an environment-dependent gap penalty function. SALIGN is extremely flexible, and the user can manipulate most features described above.

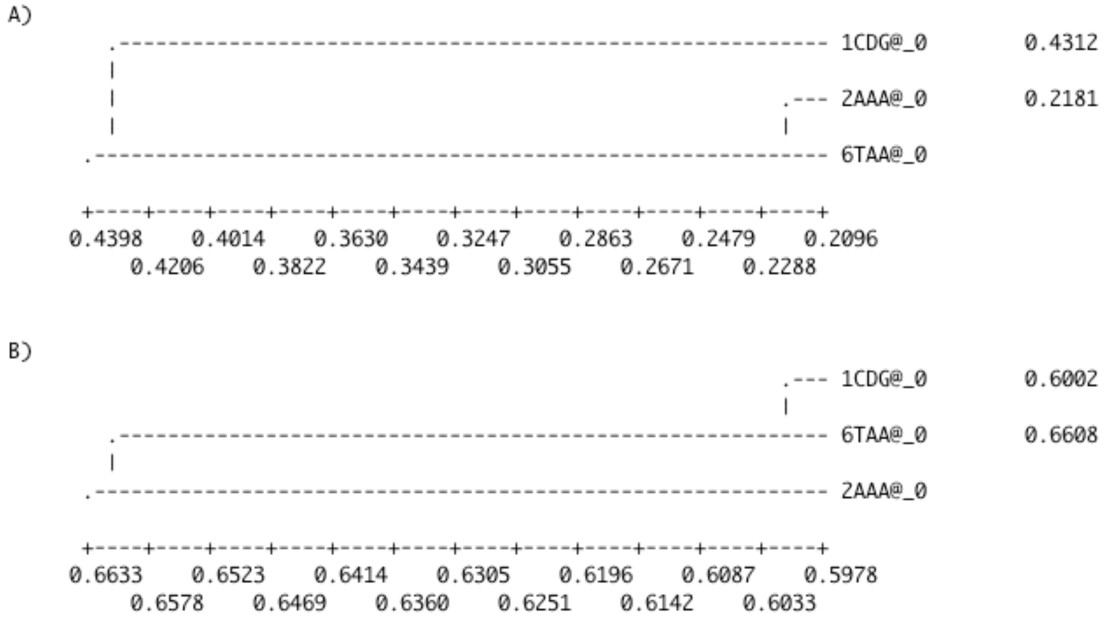
The current project consisted of creating a web-based user interface to SALIGN. Such a utility should be vastly helpful in categorizing and studying families of proteins, by making SALIGN available to non-experts. The web server is available at http://salilab.org/salign/ (password protected during an evaluation period). The methodology is first described, followed by a brief section covering implementation details. Finally, an attempt is made to describe how the server decides on a course of action based on the input information.

2.2 Methodology

The main user interface is an input page that allows the user to upload arbitrary numbers of structure (in PDB format) and alignment files (in PIR or FASTA format) (Fig. 4). The alignment files may contain sequence entries, structure entries, or a combination of the two. For each structure entry, the SALIGN server searches the PDB library as well as the uploaded files for the corresponding structure file. If no match is found the entry is treated as a sequence instead. In case the user wants to align structures that are not represented in any alignment file, the segments to be aligned can be specified manually on the web page. This option is available for uploaded structure files as well as those that can be fetched from the PDB. Furthermore, an option for pasting sequences is provided.

To simplify usage, the server processes the input information and decides on a course of action that is likely to result in the most accurate alignment. The proposed action is presented to the user who can choose to submit the job or switch to an advanced view. The advanced view offers the option to override the default action and furthermore allows a number of advanced parameters to be set (Fig. 5). The advanced features displayed depend on the input. For example, the user will not be given the option to ask for a structural alignment if the input only consists of sequences.

After successful completion of an alignment task, the results package contains the resulting alignment file, superimposed coordinate files if structures were aligned, a dendrogram file if a tree was constructed, the MODELLER log file, which gives details pertaining to the alignment process, and the MODELLER input file(s). The MODELLER input file can be used with any stand-alone version of MODELLER, version 8 and higher. The log file contains information about RMSD, number of equivalent positions, number of residues etc. The results package is retrievable via a web page, which is reachable through a hyperlink that is emailed to the user. On the results web page, the user can either download or view the output files. If structures have been aligned, the page also features a link that opens aligned structure files in the molecular graphics viewer CHIMERA (Pettersen *et al.*, 2004), which provides instant visualizations of the alignments. If errors are encountered during the run, the user is notified by email as well. This email contains a link to a web page that allows



```
A)
       .------------------------------------------------------- 1CDG@_0        0.4312
       |
       |                                                    .--- 2AAA@_0        0.2181
       |                                                    |
     .---------------------------------------------------------- 6TAA@_0

     +----+----+----+----+----+----+----+----+----+----+----+----+
   0.4398    0.4014    0.3630    0.3247    0.2863    0.2479    0.2096
        0.4206    0.3822    0.3439    0.3055    0.2671    0.2288


B)
                                                            .--- 1CDG@_0        0.6002
                                                            |
       .------------------------------------------------------- 6TAA@_0        0.6608
       |
     .---------------------------------------------------------- 2AAA@_0

     +----+----+----+----+----+----+----+----+----+----+----+----+
   0.6633    0.6523    0.6414    0.6305    0.6196    0.6087    0.5978
        0.6578    0.6469    0.6360    0.6251    0.6142    0.6033
```

Fig. 3. *Multiple structure tree alignments. PDBs 1cdg, 2aaa and 6taa were multiply aligned by SALIGN, using the tree algorithm, based on two different sets of feature weights. The feature weights dictate the influence of different sequence and structure features on the alignment (section 2.1.5). A) Feature weights: 1 1 1 1 1 0 (quality score: 88.4%) B) Feature weights: 0.1 1 0 0 0 0 (quality score 81.3%).*

the user to view or download the log file. In such a case, it may be instructive to peruse the log file, since errors are generally reported there.

2.3 Technical details

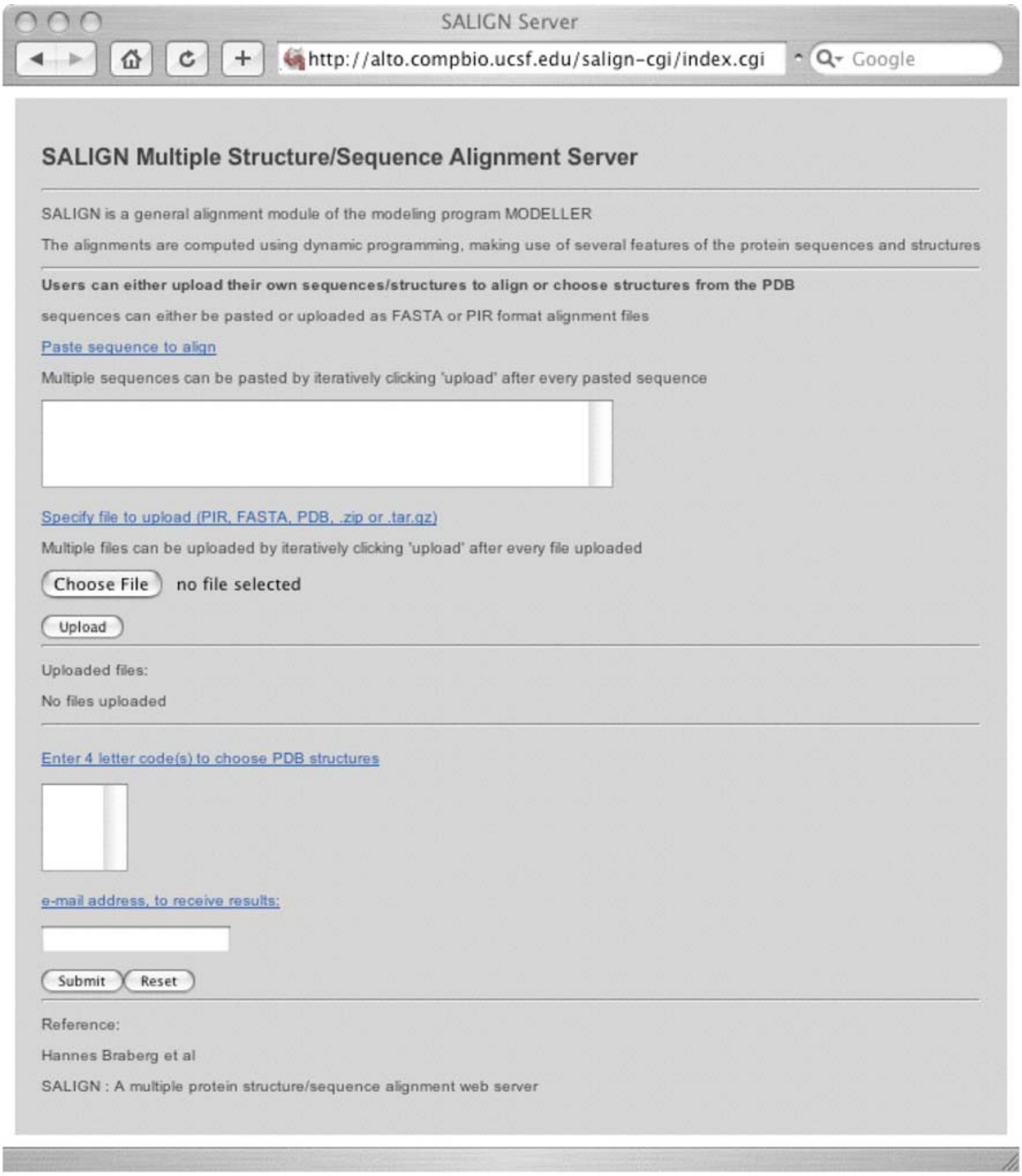
2.3.1 Implementation

The web server was implemented as a set of Perl and Perl/CGI scripts. As a job is submitted, a script creates the required MODELLER input files. In the next step, the job is added to a Linux cluster queue by a daemon that checks for new jobs every minute. SALIGN is then run on the cluster, computing the appropriate alignment(s). When a run is finished, the daemon executes a script that processes the results. This script checks for errors and emails the user a link to the results web page.

2.3.2 Decision process

This section describes how the server decides on a course of action based on the input information. Additionally, a set of flowcharts, which may clarify the decision process, is provided in the appendix. Note that the user can choose to override this default procedure in the advanced options.

Given a set of structures, the server will opt to construct a tree-based multiple alignment. The same is true for a set of sequences. There is no limit on the number of structures or sequences that the server can handle but some practical limits are enforced to optimize run time. Progressive alignment is used when the number of sequences exceeds 500 or when the number of structures exceeds 50. If two sets of sequences are input a two-step approach is performed. In the first step, each set of sequences is aligned using a substitution matrix. Sets of more than 500 sequences are, however, not aligned and should thus be prealigned upon submission. In the second



SALIGN Multiple Structure/Sequence Alignment Server

SALIGN is a general alignment module of the modeling program MODELLER

The alignments are computed using dynamic programming, making use of several features of the protein sequences and structures

Users can either upload their own sequences/structures to align or choose structures from the PDB

sequences can either be pasted or uploaded as FASTA or PIR format alignment files

Paste sequence to align

Multiple sequences can be pasted by iteratively clicking 'upload' after every pasted sequence

Specify file to upload (PIR, FASTA, PDB, .zip or .tar.gz)

Multiple files can be uploaded by iteratively clicking 'upload' after every file uploaded

no file selected

Uploaded files:

No files uploaded

Enter 4 letter code(s) to choose PDB structures

e-mail address, to receive results:

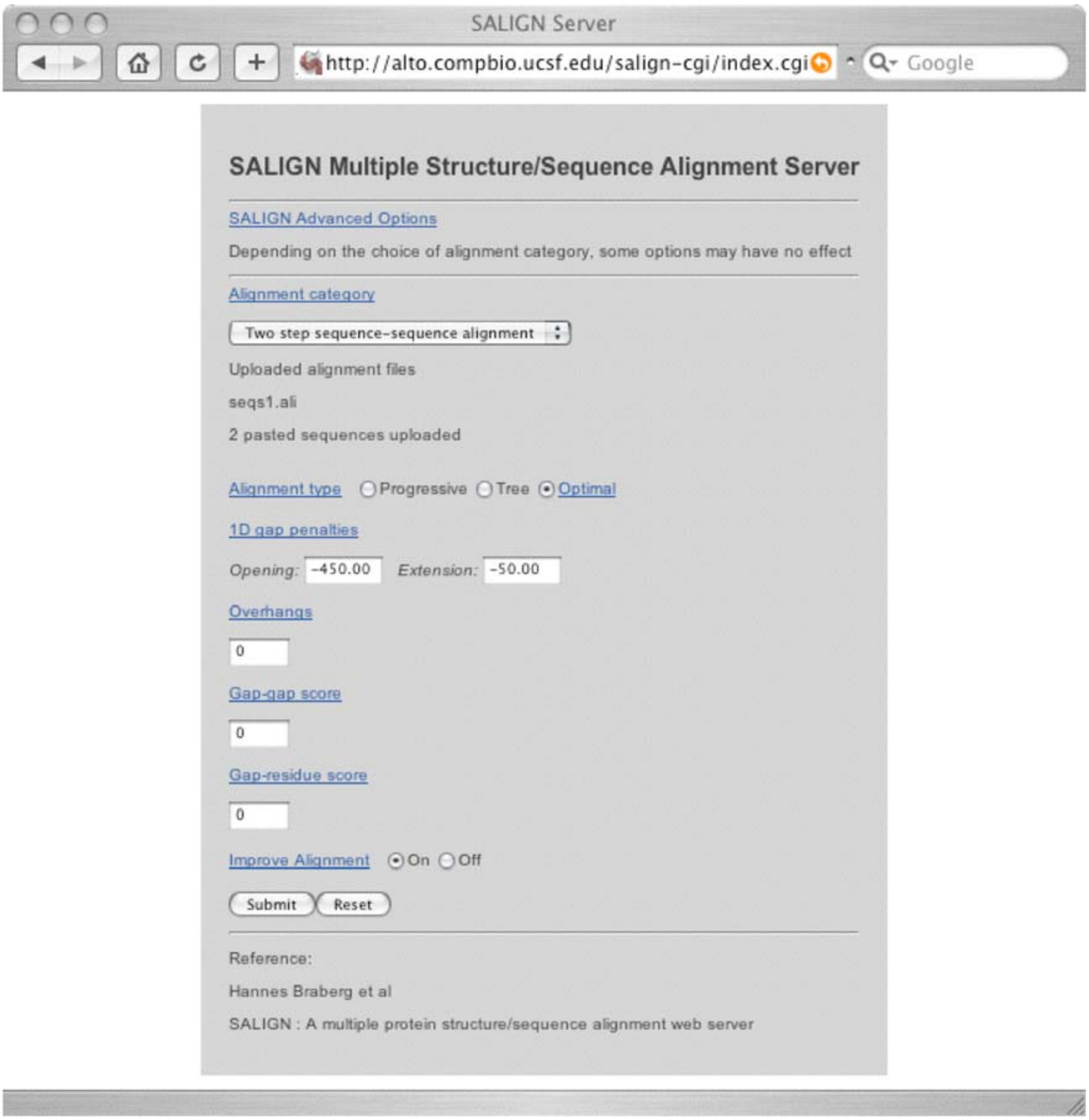
Reference:

Hannes Braberg et al

SALIGN : A multiple protein structure/sequence alignment web server

Fig. 4. *SALIGN web server input page. The upper text input field provides the user with the option to paste sequences to be aligned. By clicking the "Choose File" button, the user can upload sequence/structure alignment files, as well as PDB structure files. Clicking "Upload" for a chosen file enables the user to select a new file for upload. Pasted sequences and uploaded files are listed in the area below the "Upload" button. Further down a text field is provided for specifying structure files to be fetched from the PDB library.*

step the two sets are aligned to each other by matching their profiles. The same procedure is carried out even if one or both files consist of mixtures of structure and sequence entries. In this case, only sequence information is used for the structure entries as well. If one of the sets consists of only structure entries, it is aligned using the structure-structure feature instead. Step two is then performed as a structure-sequence alignment if the sets contain no more than 100 sequences and structures respectively. For larger sets a profile-profile alignment is performed. If the input consists of a mixture of structures and sequences, not arranged in two distinct sets,



SALIGN Multiple Structure/Sequence Alignment Server

SALIGN Advanced Options

Depending on the choice of alignment category, some options may have no effect

Alignment category

Two step sequence-sequence alignment

Uploaded alignment files

seqs1.ali

2 pasted sequences uploaded

Alignment type ○ Progressive ○ Tree ● Optimal

1D gap penalties

Opening: -450.00 *Extension:* -50.00

Overhangs

0

Gap-gap score

0

Gap-residue score

0

Improve Alignment ● On ○ Off

Reference:

Hannes Braberg et al

SALIGN : A multiple protein structure/sequence alignment web server

Fig. 5. *Example of an advanced view page of the SALIGN web server. The SALIGN web server customizes the advanced view according to the inputs. The options presented in this figure are based on the uploading of two distinct sets of sequences and no structures. In the advanced view, the user is also provided with the option to override the default alignment category (see section 2.2.3.2 and Appendix).*

independent multiple alignments of sequences and structures are performed, regardless of the distributions in the uploaded files. The multiple sequence and structure alignments are then aligned to each other by a structure-sequence pairwise alignment if neither contains more than 50 entries. If either is larger than 50 entries the two blocks are aligned using a profile-profile alignment instead.

3 Protein complex compositions predicted by structural similarity

3.1 Introduction

As discussed in section 1, accurate protein structures may provide essential information about cellular processes. The structural characterization of isolated proteins alone is, however, often not sufficient for deducing biological function. This is partly due to the fact that biologically functional units often are large, complex assemblies of several macromolecules (Russell *et al.*, 2004). These assemblies vary widely in size and shape, and play a number of key roles in the cellular processes. Examples include the ribosome, which is responsible for protein synthesis, and the nuclear pore complex, which controls the trafficking of macromolecules through the nuclear envelope. The structural characterization of macromolecular assemblies is an important component of the mapping of biochemical and cellular processes.

Recent developments in high-throughput screening have generated large data sets identifying protein complexes. The *Saccharomyces cerevisiae* proteome has been especially well characterized through yeast-two-hybrid (Y2H) (Uetz *et al.*, 2000; Ito *et al.*, 2001) and tandem affinity purification (TAP) experiments (Gavin *et al.*, 2006; Ho *et al.*, 2002; Gavin *et al.*, 2002). Experimentally observed interactions, resulting from both high-throughput and traditional low-throughput methodologies, are deposited in databases such as the Biomolecular Interaction Network Database (BIND, Bader *et al.*, 2003) and the Database of Interacting Proteins (DIP, Salwinski *et al.*, 2004).

Concomitant with these experimental advances, a spate of computational techniques to predict protein-protein interactions have also been developed. Several approaches based on protein sequence, structure, function, and genomic features have been described (Salwinski and Eisenberg, 2003). In an effort to reduce the prediction errors, several methods integrate multiple types of experimentally determined information and theoretical considerations (Jansen *et al.*, 2003; Lee *et al.*, 2004; Lu *et al.*, 2005).

Structure-based methods have been developed for the prediction of binary protein interactions. InterPreTS (Aloy and Russell, 2002) uses a statistical potential derived from known hetero-dimer structures and MULTIPROSPECTOR (Lu *et al.*, 2002) relies on threading to score pairs of proteins that are similar to binary interactions of known structure. In addition to predicting new interactions, structure-based methods can also annotate interactions that have been previously observed experimentally. A recent study used computational methods in conjunction with experimentally determined complex compositions and electron density maps from negative-stain electron cryo-microscopy to generate structural models of yeast complexes (Aloy *et al.*, 2004). In a similar vein, structural knowledge has been used to predict the domains that are most likely to mediate binary protein interactions (Nye *et al.*, 2005).

In this study (Davis *et al.*, 2006) we predicted proteins that form complexes in *S. cerevisiae* based on similarity to complexes whose atomic structures have been solved experimentally. First, comparative models of conceivable complexes are built and then assessed by a specialized statistical potential. The high-confidence interactions can be additionally filtered by examining orthogonal sources of information including sub-cellular localization and functional annotation.

The current study is unique primarily in its prediction of structural models for higher-order complexes as well as homomeric complexes. Computational methods have been developed to infer higher-order complexes from binary protein interaction networks (Bader and Hogue, 2003; Spirin and Mirny, 2003), but they do not explicitly use structural knowledge. Previous studies have also focused primarily on the prediction of heterodimers, though homodimerization is biologically prevalent and functionally significant (Marianayagam *et al.*, 2004). The multiple structure-based assessment steps, from the initial fold assignment, to the interaction prediction, enables our method to achieve a higher coverage, and presumably accuracy, than methods based solely on sequence similarity (section 3.4.2).

First, the approach and benchmarking of the method are described. Predictions are then presented for proteins in *S. cerevisiae* and validated against experimentally observed complexes. The performance of the protocol is highlighted in the selection of the correct binding mode when multiple template interface structures are available and newly predicted co-complexed superfamilies are discussed. Finally, section 3 of this thesis is concluded with a brief discussion of potential applications of the method in light of the ultimate goal of full structural coverage of interaction space.

3.2 Methods

3.2.1 Prediction algorithm

Candidate complexes are first generated, then assessed, and finally filtered by orthogonal biological information (Fig. 6(a)).

3.2.1.1 Candidate complex generation

Pairs of *S. cerevisiae* proteins were identified as potential interaction partners if they were assigned SCOP domains belonging to superfamilies for which an interaction structure exists in PIBASE (Fig. 6(b)) (Davis and Sali, 2005). In some superfamilies, such as the ARM superfamily (SCOP a.118.1), the lengths of the member domains vary widely. Because alignments between structures of different lengths are difficult, a threshold was placed on the relative sizes of the target and template domains – the shorter of the two domains must be at least 60% of the length of the longer domain. In addition, the target interface was required to have residue pairs aligned to at least 50% of the template interface contacts.

Protein Data Bank (PDB) (Berman *et al.*, 2000) structures that contained more than two domains were used as templates for the prediction of higher-order complexes with more than two proteins. Target domains that were assessed to interact through the interface modes in a given PDB structure were listed as candidate complex members. Each complex was then scored with the worst of the Z-scores for the interacting domain pairs it contained, as described below. Predicted complexes were merged if they contained different domains of a single target protein. In effect, the covalent link between the domains served as a “bridge” between predicted complexes that were based on different templates (Fig. 6(c)).

3.2.1.2 Assessment of candidate complexes

Each candidate interaction pair was scored by assessing the agreement between the target sequences and the template interface structure using a statistical potential derived from binary interface structures in PIBASE.

First, residue contacts across the interface were calculated for the template interface and grouped into classes based on the main chain or side chain participation of each residue. Next, the MODBASE models of each candidate interaction partner

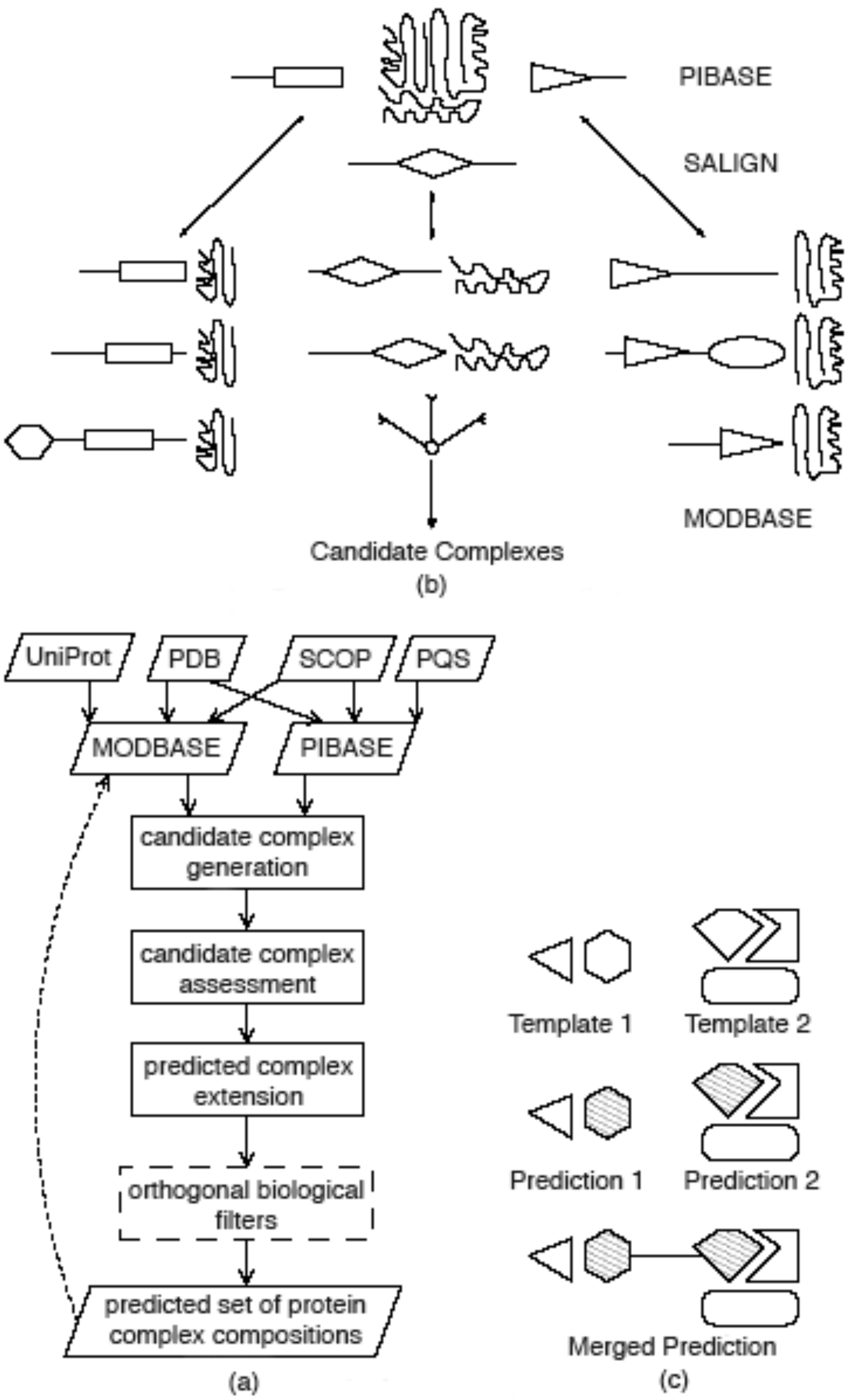

PIBASE
SALIGN
MODBASE
Candidate Complexes
(b)
UniProt
PDB
SCOP
PQS
MODBASE
PIBASE
candidate complex generation
candidate complex assessment
predicted complex extension
orthogonal biological filters
predicted set of protein complex compositions
(a)
Template 1
Template 2
Prediction 1
Prediction 2
Merged Prediction
(c)

Fig. 6. *Prediction Logic Overview. (a) Prediction Flowchart. Groups of protein sequences modeled with SCOP domains observed to form a complex in PIBASE are listed as candidate complexes. These candidate complexes are then assessed by a statistical potential. Interactions that score above a Z-score threshold are filtered using sub-cellular localization and functional annotation. The resultant predictions are deposited in MODBASE. (b) Candidate Complex Generation. Comparative models of target domains are structurally aligned to templates of known structure in PIBASE using the SALIGN module of MODELLER. Putative interface residues are identified from the alignment. (c) Predicted complexes are merged if they contain different domains of a single target protein.*

were structurally aligned against the corresponding domains in the template interface using the SALIGN module of MODELLER (Sali and Blundell, 1993). Finally, the residue correspondences defined by the alignments were used to score the candidate partner sequences against the template interface contacts using the statistical potential, as described below.

A Z-score was calculated to assess the significance of the raw statistical potential score, by consideration of the mean and standard deviation of the statistical potential scores for 1000 shuffled target sequences. Sequence randomization has been previously shown to perform comparably to a more physical model involving structural sampling in the context of fold assessment (Melo *et al.*, 2002).

3.2.1.3 Orthogonal biological information

Orthogonal biological support for each predicted complex was provided by sub-cellular localization and gene ontology functional annotation of their components, obtained from the YeastGFP (Ghaemmaghami *et al.*, 2003) and SGD databases (Dwight *et al.*, 2002), respectively. The numbers of shared localization and function terms were computed for both experimental and predicted complexes. If all pairs of proteins in a complex shared at least one function or localization term, the complex was flagged as co-functioning or co-localized, respectively.

3.2.2 Construction of statistical potentials

A series of statistical potentials was built using the binary domain interfaces in PIBASE extracted from structures at or below 2.5 Å resolution, randomly excluding 100 benchmark interfaces. Twenty-four statistical potentials were built using different values of three parameters: the contacting atom types (main chain - main chain, main chain - side chain, side chain - side chain, or all), the relative location of the contacting residues (inter- or intra- domain), and the distance threshold for contact participation (4, 6, or 8 Å):

$$g_{ij} = \frac{\sum_{p=1}^{N} \sum_{c=1}^{\Delta n_{ij}^{(p)}(R_o)} \mathrm{cifa}_{ci,cj} n_p}{\sum_{p=1}^{N} n_{ij}^{(p)} \max(\mathrm{cifa}_{i,j})} \tag{1}$$

$$\mathrm{cifa}_{x,y} = \min\left(\frac{\text{interacting atoms}_x}{\text{atoms}_x}, \frac{\text{interacting atoms}_y}{\text{atoms}_y} \right)$$

$$n_{ij}^{(p)} = \begin{cases} n_i^{(p)} n_j^{(p)} & \text{intra - domain potential,} \\ n_i^{(d1)} n_j^{(d2)} + n_i^{(d2)} n_j^{(d1)} & \text{inter - domain potential.} \end{cases}$$

$$w_{ij} = -\ln\left[\frac{g_{ij}}{\frac{1}{400}\sum_{k=1}^{20}\sum_{l=1}^{20} g_{kl}} \right] \tag{2}$$

Each of the $\Delta n_{ij}^{(p)}(R_o)$ residue pairs of type *i* and *j* in protein *p* that occurred within the distance threshold R_o was weighted by cifa, the minimum of the fraction of total atoms (of the type specified in the potential) in each residue that fell within the distance threshold (Eqn. 1), and n_p, the number of residues in the protein. This count for each residue type pair was normalized by $n_{ij}^{(p)}$, the total number of possible contacts of that type in each protein, weighted by max(cifa$_{ij}$). In the case of the inter-domain potential, $n_{ij}^{(p)}$ was computed by taking into account the occurrence of each residue type in each domain individually. Finally, the score for each residue type pair was normalized by the sum of the scores observed for all residue type pairs (Eqn. 2).

3.2.3 Benchmarking of statistical potentials

Performance on the benchmark set of 100 interfaces was used to compare the 24 statistical potentials. The sequences of these interfaces were scored against their structures and a Z-score was calculated, as described above. Receiver-operator curves (ROC) were built to describe the observed false-positive and true-positive rates at different Z-score thresholds. The ROC curves were then integrated to calculate the area under the curve (AUC). The AUC represents the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance, with 0.5 corresponding to a random prediction, and 1 to a perfect classifier (Fawcett, 2003).

3.2.4 Validation of complex prediction

The predicted interactions were validated in two ways. First, the predicted *S. cerevisiae* complexes were compared to the experimentally determined complexes in the BIND database (Bader *et al.*, 2003) and those recently reported by Gavin *et al.*, referred to as Cellzome (Gavin *et al.*, 2006). The binary interactions were compared by counting the overlap of the predictions with the interactions in the BIND and Cellzome sets. The Cellzome set consisted of pairs of proteins that were deemed highly reliable in forming partnerships based on their computed ‘socio-affinity’ score (Gavin *et al.*, 2006).

Second, the higher order complexes were compared between the predicted and experimental sets by counting how many of the predicted complexes were equivalent to, or were subcomplexes of, experimentally determined complexes. Since the predictions are based on known structures, the sizes of the predicted complexes are far smaller than those obtained by biochemical methods such as tandem affinity purification methods. For this reason, we elected not to use a metric that explicitly penalizes size differences (*e.g.*, the metric defined in Bader and Hogue, 2003).

3.2.5 Binding mode selection

The ability of the potential to select the proper binding mode when multiple template interfaces of different orientation are available was assessed. The test cases used were the structures of camelid VHH domains AMB7, AMD10, and AMD9 bound to porcine pancreatic α-amylase (PPA) (PDB codes 1kxt, 1kxv, and 1kxq, respectively). All three modes were evaluated for each VHH-PPA complex using the interface statistical potential.

3.2.6 Data sources

The prediction algorithm uses three types of data: (i) target protein sequences among which complexes are to be predicted, (ii) structures of protein complexes to be used as templates, and (iii) a list of the locations and types of structural domains in the target and template proteins (Fig. 6(a)).

3.2.6.1 Target proteins

S. cerevisiae protein sequences were obtained from MODBASE, a relational database of annotated comparative protein structure models for all available protein sequences matched to at least one known protein structure (Pieper *et al.*, 2006). The models were calculated by MODPIPE (Eswar *et al.*, 2003), an automated modeling pipeline that relies on MODELLER for fold assignment, sequence-structure alignment, model building, and model assessment (Sali and Blundell, 1993). 6,600 *S. cerevisiae* proteins were processed, resulting in 9,464 models for 3,440 sequences. 2,659 sequences had at least one reliable model (5,387 reliable models in total). A model is considered reliable when the model score, derived from statistical potentials, is higher than a cutoff of 0.7 (Melo *et al.*, 2002). A reliable model has a greater than 95% probability of having at least 30% of Cα atoms within 3.5 Å of their correct positions. 3,376 sequences had at least one reliable fold assignment (8,935 reliable folds in total). A fold assignment is considered reliable when the model is based on a PSI-BLAST match to a template with an e-value smaller than 0.0001.

3.2.6.2 Structural domain annotation

The domain definitions for PDB structures were obtained from the SCOP database (ver. 1.69) that classifies each domain using a four level hierarchy, class, fold, superfamily, and family (Murzin *et al.*, 1995). The location and types of domains in the target protein sequences were then predicted using the SCOP annotation of their MODBASE templates, as follows. Domain boundaries were first assigned based on the MODBASE alignment of each target protein to its structural template. Each target domain was required to have at least 70% of the residues in its template domain to receive the domain assignment. Next, if the target domain had greater than 30% sequence identity to the template domain and the MODBASE structural model was assessed to be reliable, the target domain received the template's SCOP classification at the family level. If the sequence identity was less than 30% and a reliable model was built or if the sequence identity was greater than 30% but MODBASE deemed only a reliable fold assignment, the superfamily was assigned. The remaining domains received the template domain's SCOP classification at the fold level, and were not used in the interaction prediction.

For those target proteins for which multiple models were available in MODBASE, a tiling procedure combined the domain assignments for each model into a non-overlapping set of domain boundaries that maximized the coverage length and classification detail in the SCOP hierarchy.

3.2.6.3 Template complexes

Structures of template complexes were retrieved from PIBASE, a comprehensive relational database of structurally defined protein interfaces (Davis and Sali, 2005). It currently includes 209,961 structures of interactions between 2,613 SCOP domain families. The ASTEROIDS component of the SCOP ASTRAL compendium was used to cluster the interfaces, reducing the computational expense of the predictions (Chandonia *et al.*, 2004). The ASTEROIDS alignments, available for SCOP classes a-g, were used together with the interface contacts stored in PIBASE to cluster all interface structures that shared pairs of SCOP families. When two interfaces shared at least 75% equivalent interface contact positions, they were merged into a single cluster. The clustering reduced the 79,428 domain interfaces between pairs of domains in the SCOP classes a-g to 21,791 representative interfaces. These interfaces were filtered using a threshold of at least 1,000 interatomic contacts resulting in a set of interfaces of significant size. The final set of template binary interfaces contained 5,275 structures, including both intermolecular and intramolecular interfaces.

3.2.7 Technology

The prediction system was implemented as a Perl module and an integrated set of Perl scripts, except for the inter-atomic contacts calculator written in ANSI C (Davis and Sali, 2005). The SALIGN module of MODELLER (Sali and Blundell, 1993) was used to generate model template alignments. The Perl DBI interface was used to access the MODBASE and PIBASE MySQL databases (http://www.mysql.com). The calculations were done in a parallel fashion on 50 3.0 GHz Pentium IV processors, taking 20 hours for the yeast genome. The predictions are accessible *via* the MODBASE web interface (http://salilab.org/modbase).

3.3 Results

3.3.1 Benchmark

The statistical potentials were tested using the benchmark set of 100 complexes, and their performance compared using receiver operator curves (ROC) (Methods). The highest power of discriminating between the native and non-native interfaces was achieved by the statistical potential built from side chain - side chain contacts across the interfaces at a threshold of 8 Å, corresponding to the extent of the first residue shell (Fig. 7). The ROC curve for this potential had an area under the curve (AUC) of 0.993, and at the optimal Z-score threshold of −1.7 had true positive and false positive rates of 97% and 3%, respectively. Clear performance trends were observed for the parameters sampled in the potential construction. The inter-domain potential always performed better than the corresponding intra-domain potential, when all other parameters were equivalent (data not shown). The side chain - side chain (SS) potential performed better than the corresponding main chain - side chain (MS) potential, which in turn performed better than the corresponding main chain - main chain (MM) potential. At 6 Å and 8 Å, the all atom-type potential performed better than only the MM potential. At 4 Å, the all atom-type potential performed better than both MS and MM potentials. The range of performances, generated by varying the other parameters (*i.e.*, atom type, inter- or intra-domain), was widest at the 4 Å distance threshold and least at 8 Å.

3.3.2 Predictions

The best statistical potential, as determined above, was then used to assess candidate interactions between *S. cerevisiae* proteins. 12,867 binary interactions that scored at or below a Z-score threshold of −1.7 were predicted between 1,390 *S. cerevisiae* proteins (Fig. 8(a)). Next, the co-function and co-localization filters were separately applied, reducing the original 12,867 interactions to 6,808 and 4,432, respectively. The combined co-localization and co-function filter resulted in 3,213 predictions. 12,702 higher-order complexes were also predicted at a Z-score threshold of −1.7 between 589 proteins. Similar to the binary predictions, the orthogonal filters reduced this number to 1,234 complexes between 195 proteins.

The predictions spanned the entire spectrum of target-template sequence similarity (Fig. 8(b)). This distribution reflects both the comparative modeling procedure used to build models of the individual proteins and the procedure used to identify potential interaction templates. The mean target-template sequence identity of the reliable

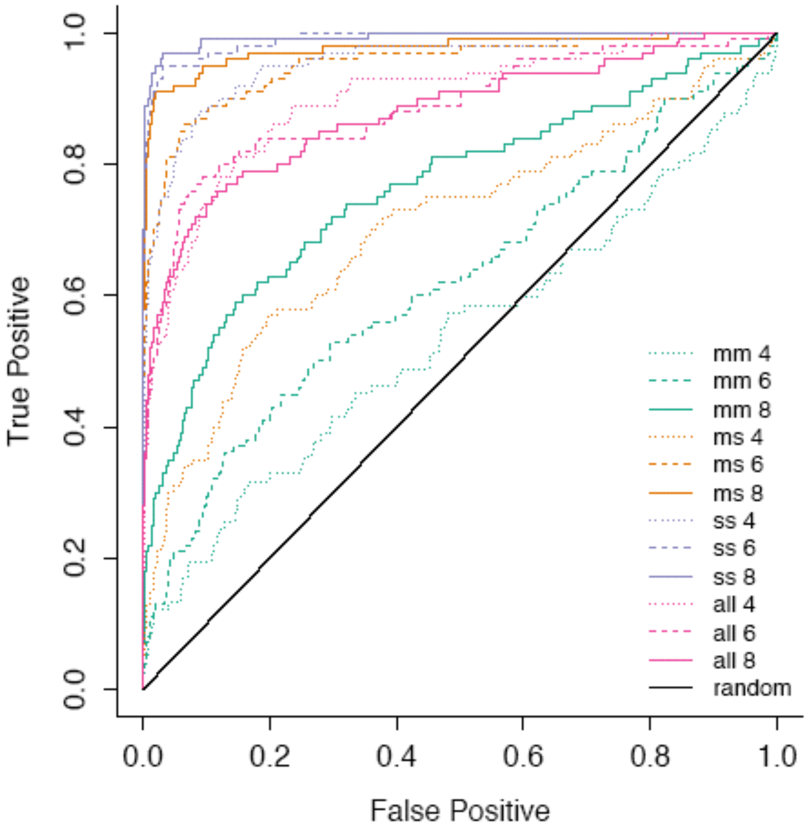


Fig. 7. *Assessment of statistical potentials. Receiver operator curves (ROC) are shown for the inter-domain potential performance on the benchmark set of complexes.*

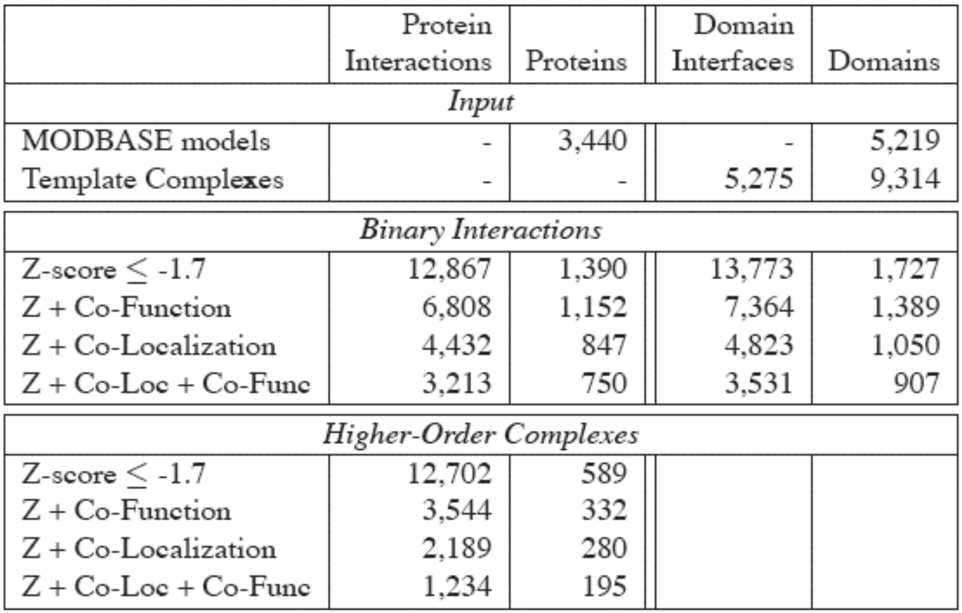
models built for *S. cerevisiae* proteins is 31%. Domains from different families within the same superfamily, the SCOP level used to identify potential interaction templates, often share less than 30% sequence identity. Both of these factors influence the distribution of target-template identities observed for the predicted interactions.

The fractions of predicted binary interactions that passed the co-function (53%), co-localization (34%), and both co-function and co-localization (25%) filters were similar to the fractions for BIND interactions (39%, 33%, and 21%, respectively). The Cellzome set more readily passed these filters (85%, 58%, and 52%, respectively).

3.3.3 Validation

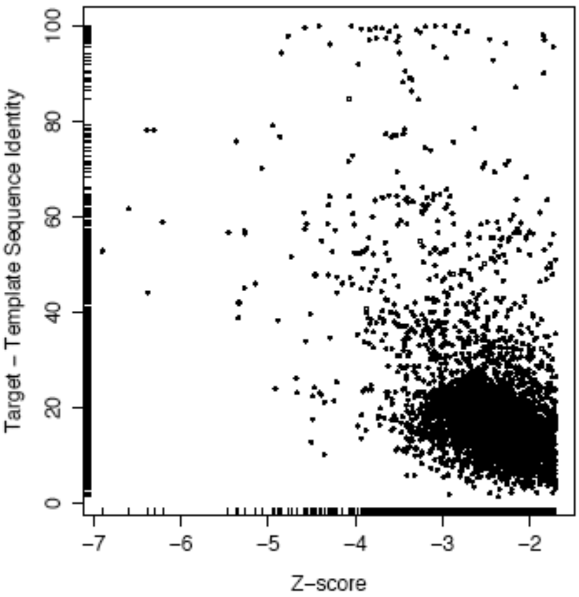
The predictions were then compared with known experimental interactions, as deposited in the BIND database. 248 of the 3,213 predicted binary interactions that passed the combined co-localization and co-function filter overlapped with known binary interactions. 8 of the 1,234 predicted higher-order complexes were also found as subcomplexes of experimental complexes.

The enrichment of the unfiltered predictions with known binary interactions begins to plateau at 0.2 around a Z-score threshold of −3.5, with an enrichment value of 0.03 at the Z-score of −1.7 (Fig. 9(a)). The predictions that passed the separate localization and function filters both reached a peak of 0.28 at a Z-score of −3.6. Both filters produced enrichment values of 0.06 at the Z-score threshold of −1.7. The enrichment of the predictions that passed the combined co-localization and co-function filter exhibited a higher peak of 0.36 at the Z-score of −3.6. At the Z-score threshold of −1.7, the combined filter produced an enrichment of 0.08, a more than two-fold



	Protein Interactions	Proteins	Domain Interfaces	Domains
Input				
MODBASE models	-	3,440	-	5,219
Template Complexes	-	-	5,275	9,314
Binary Interactions				
Z-score $\leq$ -1.7	12,867	1,390	13,773	1,727
Z + Co-Function	6,808	1,152	7,364	1,389
Z + Co-Localization	4,432	847	4,823	1,050
Z + Co-Loc + Co-Func	3,213	750	3,531	907
Higher-Order Complexes				
Z-score $\leq$ -1.7	12,702	589		
Z + Co-Function	3,544	332		
Z + Co-Localization	2,189	280		
Z + Co-Loc + Co-Func	1,234	195		

(a) *S. cerevisiae* complex predictions



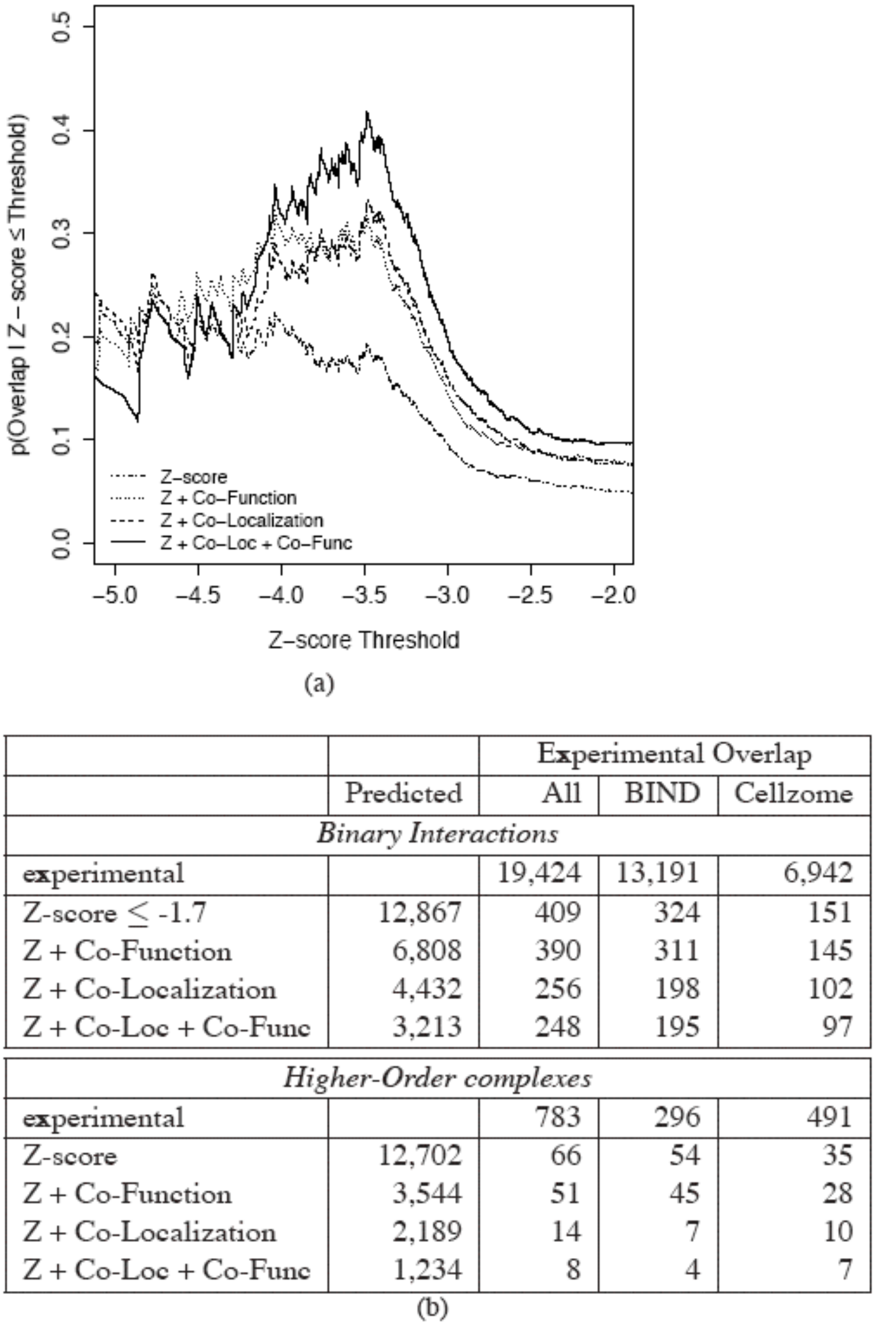
(b)

Fig. 8. *S. cerevisiae predictions. (a) Predictions of binary and higher-order complexes filtered by sub-cellular localization and annotated function. (b) Average sequence identity of predicted interaction partners to template interacting domains vs. Z-score. The predictions shown were scored with Z-score ≤ −1.7, and passed the combined co-localization and co-function filter.*

increase compared to the unfiltered predictions.

3.3.4 Comparison to other computational methods

The performance of the method in predicting binary interactions is comparable to similar structure-based methods that have been previously applied to *S. cerevisiae* on a genomic scale. Here, an overlap of 248 binary interactions is observed between the set of 3,213 (7.7%) predictions and 19,424 (1.3%) experimental observed binary interactions. 374 of 7,321 (5%) interactions predicted by threading occurred in a set of



		Experimental Overlap		
	Predicted	All	BIND	Cellzome
Binary Interactions				
experimental		19,424	13,191	6,942
Z-score ≤ -1.7	12,867	409	324	151
Z + Co-Function	6,808	390	311	145
Z + Co-Localization	4,432	256	198	102
Z + Co-Loc + Co-Func	3,213	248	195	97
Higher-Order complexes				
experimental		783	296	491
Z-score	12,702	66	54	35
Z + Co-Function	3,544	51	45	28
Z + Co-Localization	2,189	14	7	10
Z + Co-Loc + Co-Func	1,234	8	4	7

Fig. 9. *Experimental overlap of S. cerevisiae predictions. (a) The probability of finding an experimentally observed interaction in the predicted set, as a function of the statistical potential Z-score. The unfiltered predictions are represented by dotted-dashed, the co-function filtered by dotted, the co-localization by dashed, and the combined co-localization and co-function filtered set by solid lines. The curves are only shown to a Z-score threshold of −5.0, because of the sparseness of predictions below this level. (b) Experimental overlap of the binary and higher-order predictions filtered by sub-cellular localization and annotated function.*

78,930 (0.4%) experimentally determined yeast interactions (Lu *et al.*, 2003). An overlap of 59 predicted interactions with an experimental set of 2,590 (2.3%) interactions was obtained by interface model assessment (Aloy and Russell, 2002).

3.3.5 Alternate binding modes

The ability of the algorithm to correctly select the native binding mode when alternate templates are available was tested. The native binding mode was correctly selected for all three VHH domains interacting with porcine pancreatic α-amylase (Fig. 10). In addition, the statistical potential scores that were computed for the native binding modes exhibit the same rank-order as the affinity measured experimentally by total internal reflectance (Lauwereys *et al.*, 1998).

3.3.6 Co-complexed domains

An extension process merged predicted complexes containing different domains of a single target protein (Fig. 6(c)). This process predicted 279 pairs of co-complexed SCOP domain families that were not present in the structures of template complexes. The comparison to experimental complexes was done at the superfamily level, as many of the domains in the experimental complexes were assigned domains that were classified only to this level in the SCOP hierarchy (Fig. 11).

3.4 Discussion

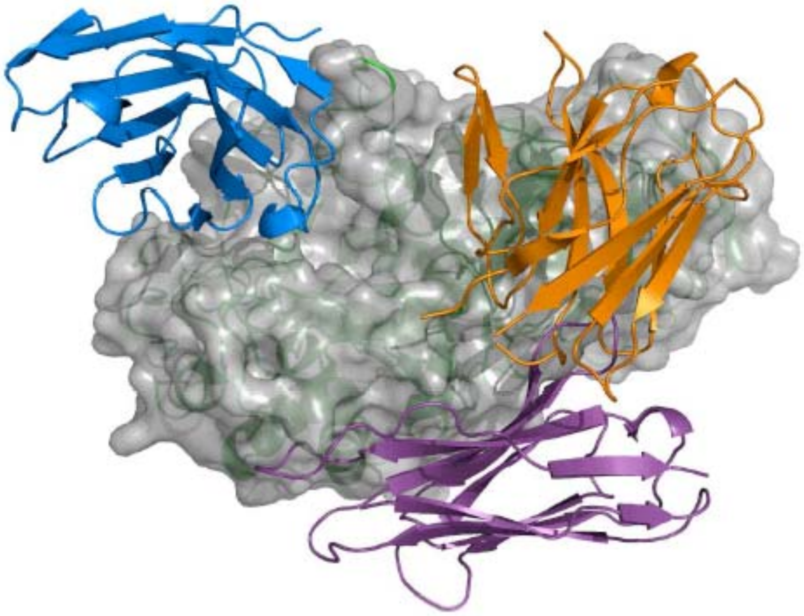
Sections 3.2 and 3.3 described our method to predict protein complex compositions by generating comparative models of candidate complexes based on sequence similarity to structurally known complexes followed by model assessment (Fig. 6). We applied the method to the *S. cerevisiae* proteome (Fig. 8) and compared the predicted complexes with experimental data (Fig. 9, Fig. 11). We further tested the method by distinguishing between multiple template binding modes (Fig. 10). The observed performance is now discussed and the limitations of the algorithm are described. Finally, the information gained in the present study, and its applications to increasing structural description of protein interactions, is discussed.

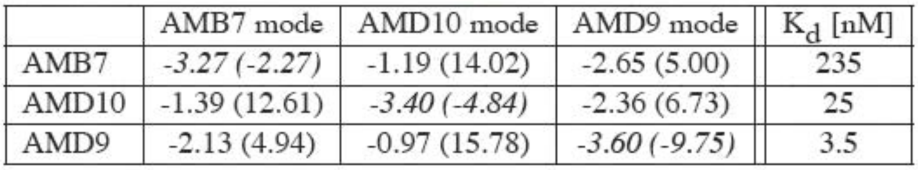
3.4.1 Accuracy

Because a large set of true negative interactions is not available, only the positives, or predicted interactions, can be compared between experiment and predictions. This limitation restricts the validation of the predictions because if the Z-score threshold is loosened, maximal overlap can be achieved at the expense of the false positive rate. However, the false positive rate can not be counted with certainty, as false positives can not be distinguished from false negatives in the experimental data sets, which can be quite high (von Mering *et al.*, 2002). Similar validation problems are encountered when testing protein ligand docking algorithms. Here, a measure related to the enrichment factor used in protein ligand docking was applied (Fig. 9(a)).

The overlap observed between the predicted and experimentally observed complexes is comparable to that between different experimental procedures (von Mering *et al.*, 2002). 248 of the 3,213 predicted binary interactions and 8 of the 1,234 predicted higher-order complexes were present in the BIND or Cellzome datasets (Fig. 9).

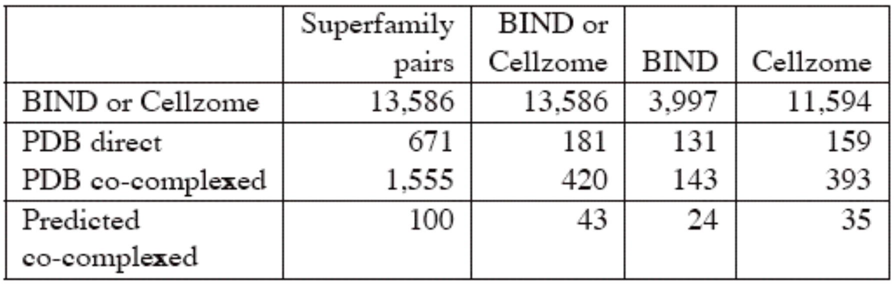
This overlap is a result of several factors. First, by construction our method is restricted to protein interactions for which structural templates exist. For this reason, our method is also biased towards complexes that are stable enough to be amenable to structure determination, whereas the yeast-two-hybrid method that populates most of the high-throughput entries in BIND, is biased towards transient interactions (von Mering *et al.*, 2002). Secondly, many PDB entries do not contain complete domains for both partners (e.g., SH3 domain - peptide complexes) and were thus not





	AMB7 mode	AMD10 mode	AMD9 mode	K_d [nM]
AMB7	*-3.27 (-2.27)*	-1.19 (14.02)	-2.65 (5.00)	235
AMD10	-1.39 (12.61)	*-3.40 (-4.84)*	-2.36 (6.73)	25
AMD9	-2.13 (4.94)	-0.97 (15.78)	*-3.60 (-9.75)*	3.5

Fig. 10. *Selection among alternate binding modes. Camelid VHH domains AMB7 (orange), AMD10 (magenta), and AMD9 (blue) bind to porcine pancreatic α-amylase (PPA, grey surface) through three distinct binding modes (PDB codes 1kxt, 1kxv, and 1kxq, respectively). All three modes were evaluated for each VHH-PPA complex using the interface statistical potential. The Z-scores are presented along with the raw score in parenthesis. Dissociation constants measured by total internal reflectance (IAsys) were obtained from literature (Lauwereys et al., 1998). Image created by PyMOL (Delano Scientific, 2002).*



	Superfamily pairs	BIND or Cellzome	BIND	Cellzome
BIND or Cellzome	13,586	13,586	3,997	11,594
PDB direct	671	181	131	159
PDB co-complexed	1,555	420	143	393
Predicted co-complexed	100	43	24	35

Fig. 11. *Co-complexed domain superfamilies. The pairs of co-complexed superfamilies observed in the BIND and Cellzome complexes are compared to the direct interactions in the PDB, co-complexed pairs in the PDB, and the predicted co-complexed pairs resulting from the complex extension procedure.*

considered as templates in the current prediction protocol. Finally, the challenge faced in predicting binary interactions increases combinatorially for higher-order complexes.

The use of sub-cellular localization data and functional annotation as filters for the predictions increased their overlap with experimental complexes, as compared to the unfiltered predictions. This finding is in agreement with previous observations that combining multiple sources of information improves the accuracy of function annotation as well as interaction prediction (Jansen *et al.*, 2003; Lee *et al.*, 2004; Lu *et al.*, 2005). Our method easily allows for the use of additional biological filters when other types of data are available, such as synthetic gene lethality (Tong *et al.*, 2001), co-expression (Tirosh and Barkai, 2005), *etc.* This incremental addition of orthogonal information is also necessary to more accurately represent the conditions in the cellular milieu, where the propensity of two protein structures to interact is not limited only by the physical chemistry of the interaction, but also by higher levels of biological regulation, including compartmentalization, expression, degradation, abundance, *etc*. Depending on the application, the user may decide to apply different biological filters.

3.4.2 Importance of structure

The majority (98.6%) of the filtered binary interactions as well as the subset that overlapped with experimentally observed interactions (86.9%) were based on templates sharing less than 80% sequence identity, a threshold previously established for reliable transfer of a known interaction to a putative interaction between homologous proteins (Fig. 8(b)) (Yu *et al.*, 2004). This distribution highlights the advantage garnered by the use of structure and the importance of a structure-based assessment.

One such example is the experimentally observed interaction between LSM2 and LSM7 that was predicted here based on structural similarity to the 14-mer complex of SmAP3, an Sm-like protein from the archae *Pyrobaculum aerophilum* (PDB 1m5q). The sequence identities of LSM2 and LSM7 to SmAP3 are 23% and 2.4%, respectively. While interface templates with higher sequence identities were available (highest identities of 20.7% for LSM2 and 32.1% for LSM7 to chains G and A of PDB 1jbm, respectively), the 1m5q-based model was scored most favorably by the statistical potential. Another example of a known interaction predicted using a distantly related template interaction is that between the delta (GCD2) and beta (GCD7) subunits of the translation initiation factor eIF2B, predicted based on similarity to the structure of Ypr118w, a methylthioribose-1-phosphate isomerase related to regulatory eIF2B subunits. The prediction was made based on sequence similarities of 16% and 15%, respectively.

3.4.3 Alternative binding modes

The ability of the algorithm to choose the correct binding mode when multiple templates are available was illustrated by evaluation of three alternative binding modes that have been structurally characterized between porcine pancreatic α-amylase and camelid VHH domains (Fig. 10). The algorithm successfully chose the native binding mode for all three VHH domains. In addition, the statistical potential scores that were computed for the native binding modes exhibit the same rank-order as the affinity of the interactions measured by total internal reflectance (Lauwereys *et al.*, 1998).

However, this example is also cautionary in that each VHH domain had one non-native mode that scored below the optimal Z-score threshold, though only the native modes produced negative raw scores (Results). In a large-scale predictive setting, if the native binding mode was not available as a template, the VHH domain would have been predicted to interact with PPA, but through an incorrect binding mode.

3.4.4 Network specificities

A more difficult test of the method is the prediction of specificities within interaction networks between homologous proteins. To address this problem, the method was applied to predict the specificities within the Epidermal Growth Factor Receptor (EGFR) and Tumor Necrosis Factor β (TNFβ) networks of ligand receptor interactions (data not shown). In both networks the method failed to recapitulate known binding preferences. Specifically, the rank order of the Z-scores for the assessed pairs did not correlate with known binding preferences.

This error was not surprising. The randomization scheme employed in the Z-score assessment of the raw statistical potential score simulated alternative binding modes. In contrast, it was not designed or tested to determine specificities. This task is difficult as large training data sets of this type are not available.

Rather than predicting specificities, the method presented here is applicable as a first pass for genome-wide predictions of protein complexes. The resulting predictions are then suitable for a follow up with more accurate computational methods, which on their own are not feasible on a large-scale.

3.4.5 Extension of known co-complexed domain superfamilies

Large protein complexes present unique challenges to structural characterization. Direct physical interactions have been experimentally observed between domains from 671 pairs of different SCOP superfamilies (excluding homo-family interactions). Domains from 1,555 pairs of different superfamilies have been observed to co-complex in the same PDB entry. 420 of these pairs have also been observed in biochemical complexes. Through an extension process that merged predicted complexes containing different domains of a single target protein, an additional 100 pairs of superfamilies were predicted to be co-complexed (Fig. 6(c), Fig. 11). 43 of these newly predicted pairs were also found in the experimental complexes. This extension procedure will be especially informative when applied to proteins from higher organisms with greater domain architecture complexity than *S. cerevisiae* (Bornberg-Bauer *et al.*, 2005).

3.4.6 Future directions

Section 3 of this thesis presented a tool for the prediction and assessment of the composition and structure of protein complexes. The results suggest that the algorithm may in practice be useful in conjunction with additional biological data, such as protein localization and functional annotation. Through its integration with MODBASE, the method is applicable, in an automated fashion, to all genomes with sequences that are amenable to comparative protein structure modeling. The method will be especially informative for proteomes of species that have not been characterized to the extent of *S. cerevisiae*, either because the genomes have only recently been sequenced or because the organisms are difficult to analyze experimentally.

In addition to proposing new protein complexes that have not previously been observed, the present study also enables a more rigorous, structure-based, analysis of experimental protein interaction data. For instance, the system could be used to distinguish complexes from temporally distinct interactions by assessing whether the interactions are sterically compatible or exclusive (Han *et al.*, 2004). The predictions may also prove useful in guiding experiments that aim to probe the interactions, such as various site-directed mutagenesis and interaction design studies.

Comparative protein structure modeling is increasingly used to help bridge the resolution gap between electron cryo-microscopy (cryo-EM) density maps and atomic protein structures (Topf and Sali, 2005). Fitting of protein and protein domain models into density maps of large assemblies is already common, but depending on the resolution, the information encoded in the map is often insufficient for an unambiguous determination of the positions and orientations of the individual proteins (Fabiola and Chapman, 2005). Models of the complexes predicted here may provide additional restraints for a more accurate fitting of proteins into large complexes studied by cryo-EM and electron cryo-tomography (Sali *et al.*, 2003; Aloy *et al.*, 2004).

As the number and size of experimentally determined structures of protein complexes increase, the number of complexes that can be predicted and modeled using these structures as templates increases correspondingly, expanding the structural coverage of protein interaction space (Aloy and Russell, 2004). In combination with other computational methods, the presented method will allow biologists to harness interaction information that has been experimentally determined for similar systems to inform their hypotheses or experiments.

Acknowledgements

I would like to thank my supervisor, M. S. Madhusudhan, for always being there when needed, Andrej Sali for having me at his lab, and the members of the Sali laboratory for bearing with me, in particular Fred P. Davis and Ursula Pieper. I am also grateful to my reviewer, Gerard Kleywegt, for his comments and suggestions.

References

Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., *et al.* (2004) Structure-based assembly of protein complexes in yeast. *Science,* **303**, 2026–2029.

Aloy, P. and Russell, R. B. (2002) Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A*, **99**, 5896–5901.

Aloy, P. and Russell, R. B. (2004) Ten thousand interactions for the molecular biologist. *Nat Biotechnol,* **22**, 1317–1321.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., Murzin, A. G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, **32**, D226-229.

Bader, G. D., Betel, D. and Hogue, C. W. (2003) Bind: the biomolecular interaction network database. *Nucleic Acids Res,* **31**, 248–250.

Bader, G. D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Baker, D. and Sali, A. (2001) Protein structure prediction and structural genomics. *Science,* **294**, 93-96.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.

Berg, J., Tymoczko, J. L., Stryer, L. (2002) *Biochemistry, 5th edition, W. H. Freeman, New York.*

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235–242.

Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S. K., Teichmann, S. A. and r. d. Weiner J. (2005) The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci*, **62**, 435–445.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004) The Gene Ontology Annotation (GOA) database: sharing knowledge in uniprot with gene ontology. *Nucleic Acids Res*, **32**, D262–D266.

Chandonia, J. M., Hon, G., Walker, N. S., Conte, L. L., Koehl, P., Levitt, M. and Brenner, S. E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res*, **32**, D189–D192.

Davis, F. P., Braberg, H., Shen, M. Y., Pieper, U., Sali, A., Madhusudhan, M. S. (2006) Protein complex compositions predicted by structural similarity. *Nucleic Acids Res*, **34**, 2943–2952.

Davis, F. P. and Sali, A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.

Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., *et al.* (2002) Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go). *Nucleic Acids Res*, **30**, 69–72.

Eswar, N., John, B., Mirkovic, N., Fiser, A., Ilyin, V. A., *et al.* (2003) Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res*, **31**, 3375–3380.

Fabiola, F. and Chapman, M. S. (2005) Fitting of high-resolution structures into electron microscopy reconstruction images. *Structure*, **13**, 389–400.

Fawcett, T. (2003) ROC graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Labs, HP Labs, Palo Alto, CA, USA. URL www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf.

Fiser, A., Feig, M., BrooksIII, C. L., Sali, A. (2002) Evolution and physics in comparative protein structure modeling. *Accounts of Chemical Research*, **35**, 413-421.

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*. (ENG).

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K. and Weissman, J. S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.

Gribskov, M. (1994) Profile analysis. *Methods Mol Biol*, **25**, 247–266.

Gribskov, M., McLachlan, A. D., Eisenberg, D. (1987) Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci U S A*, **84**, 4355-4358.

Gribskov, M., Luthy, R., and Eisenberg, D. (1990) Profile analysis. *Methods Enzymol*, **183**, 146–159.

Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., *et al.* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

Holm, L. and Sander, C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res*, **27**, 244-247.

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

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. and Gerstein, M. (2003) A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.

Koehl, P. (2001) Protein structure similarities. *Curr Opin Struct Biol*, **11**, 348-353.

Lauwereys, M., Ghahroudi, M. A., Desmyter, A., Kinne, J., Holzer, W., Genst, E. D., Wyns, L. and Muyldermans, S. (1998) Potent enzyme inhibitors derived from dromedary heavy-chain antibodies. *EMBO J*, **17**, 3512–3520.

Lee, I., Date, S. V., Adai, A. T. and Marcotte, E. M. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.

Lesk, A. M. and Chothia, C. (1980) How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J. Mol. Biol.*, **136**, 225–270.

Lu, L., Arakaki, A. K., Lu, H. and Skolnick, J. (2003) Multimeric threading-based prediction of protein-protein inter-actions on a genomic scale: application to the saccharomyces cerevisiae proteome. *Genome Res*, **13**, 1146–1154.

Lu, L., Lu, H. and Skolnick, J. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, **49**, 350–364.

Lu, L. J., Xia, Y., Paccanaro, A., Yu, H. and Gerstein, M. (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res*, in press.

Luthy, R., Bowie, J. U., Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

Madhusudhan, M. S., Marti-Renom, M. A., Sanchez, R., Sali, A. (2006) Variable gap penalty for protein sequence-structure alignment. *Protein Engineering Design & Selection*, **19**, 129-133.

Madhusudhan, M. S., Marti-Renom, M. A., Eswar, N., John, B., Pieper, U., Karchin, R., Shen, M. Y., Sali, A. (2005) Comparative Protein Structure Modeling. In: *The Proteomics Protocols Handbook*. Ed: J.M. Walker. *Humana Press Inc., Totowa, NJ*, 831-860.

Marianayagam, N. J., Sunde, M. and Matthews, J. M. (2004) The power of two: protein dimerization in biology. *Trends Biochem Sci*, **29**, 618–625.

Marti-Renom, M. A., Stuart, A., Fiser, A., Sanchez, R., Melo, F., Sali, A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct*, **29**, 291-325.

Marti-Renom, M. A., Madhusudhan, M. S., Sali, A. (2004) Alignment of protein sequences by their profiles. *Protein Science*, **13**, 1071-1087.

Matsumoto, R., Sali, A., Ghildyal, N., Karplus, M., Stevens, R. L. (1995) Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells. A cluster of histidines on mouse mast cell protease 7 regulates its binding to heparin serglycin proteoglycans. *J Biol Chem*, **270,** 19524–19531.

Melo, F., Sanchez, R. and Sali, A. (2002) Statistical potentials for fold assessment. *Protein Sci*, **11**, 430–448.

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**, 536–540.

Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**, 443-453.

Nye, T. M., Berzuini, C., Gilks, W. R., Babu, M. M. and Teichmann, S. A. (2005) Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, **21**, 993–1001.

Pawson, T. and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., Ferrin, T. E. (2004) UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J Comput Chem*, **25**, 1605-1612

Pieper, U., Eswar, N., Davis, F. P., Braberg, H., Madhusudhan, M. S., *et al.* (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res*, **34**, D291–D295.

Ring, C. S., Sun, E., McKerrow, J. H., *et al.* (1993) Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc Natl Acad Sci. U S A*, **90**, 3583–3587.

Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, M., Sali, A. (2004) A structural perspective on protein-protein interactions. *Curr Opin Struct Biol*, **14**, 313-324.

Sadreyev, R., Grishin, N. (2003) COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol*, **326**, 317-336.

Sali, A. and Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, **234**, 779–815.

Sali, A., Glaeser, R., Earnest, T. and Baumeister, W. (2003) From words to literature in structural proteomics. *Nature*, **422**, 216–225.

Sali, A. and Blundell, T. L. (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol*, **212**, 403-428

Salwinski, L. and Eisenberg, D. (2003) Computational methods of analysis of protein-protein interactions. *Curr Opin Struct Biol*, **13**, 377–382.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, **32**, D449–D451.

Sanchez, R. and Sali, A. (1998) Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc Natl Acad Sci U S A*, **95**, 13597–13602.

Sellers, P. H. (1974) Theory and computation of evolutionary distances. *Siam J Appl Math,* **26**, 787-793.

Spahn, C. M., Beckmann, R., Eswar, N., Penczek, P.A., Sali, A., Blobel, G., Frank, J. (2001) Structure of the 80S ribosome from Saccharomyces cerevisiae--tRNA-ribosome and subunit-subunit interactions. *Cell,* **107**, 373-386.

Spirin, V. and Mirny, L. A. (2003) Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, **100**, 12123–12128.

Tang, K. S., Fersht, A. R., Itzhaki, L. S. (2003) Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure (Camb.),* **11**, 67–73.

Tirosh, I. and Barkai, N. (2005) Computational verification of protein-protein interactions by orthologous co-expression. *BMC Bioinformatics*, **6**, 40.

Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364–2368.

Topf, M. and Sali, A. (2005) Combining electron microscopy and comparative protein structure modeling. *Curr Opin Struct Biol,* **15**, 578–585.

Torda, A. E. (1997) Perspectives on Protein Fold Recognition. *Curr Opin Struct Biol*, **7**, 200-205.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

Vakser, I. A. (1995) Protein docking for low-resolution structures. *Protein Eng,* **8**, 371–377.

Vitkup, D., Melamud, E., Moult, J., Sander, C. (2001) Completeness in Structural Genomics. *Nature Struct. Biol*, **8**, 559-566.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V., Gilliland, G. L., Bluhm, W., Weissig, H., Greer, D. S. *et al.* (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.

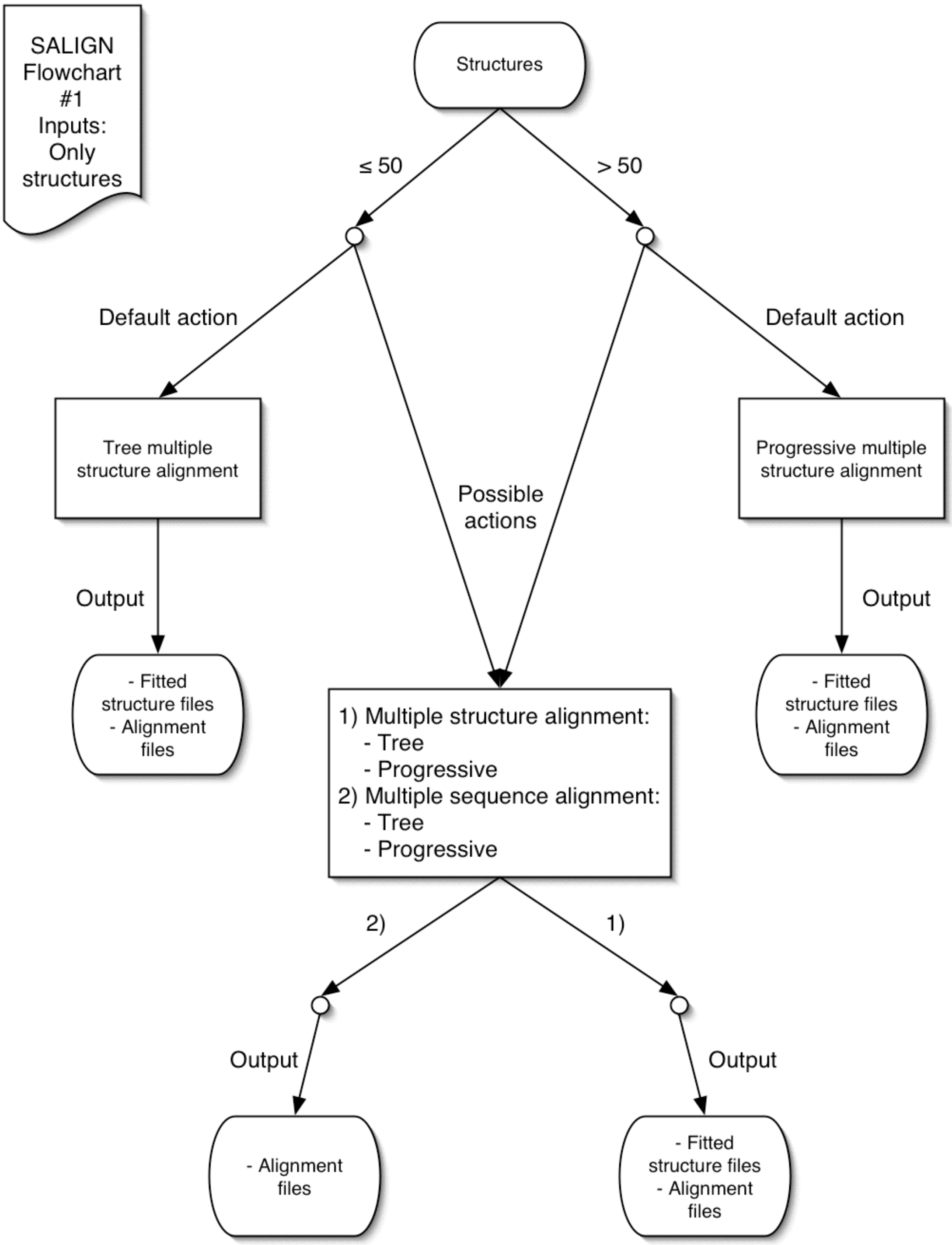
Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., et al (2006) The universal protein resource (Uniprot): an expanding universe of protein information. *Nucleic Acids Res*, **34**, D187–D191.

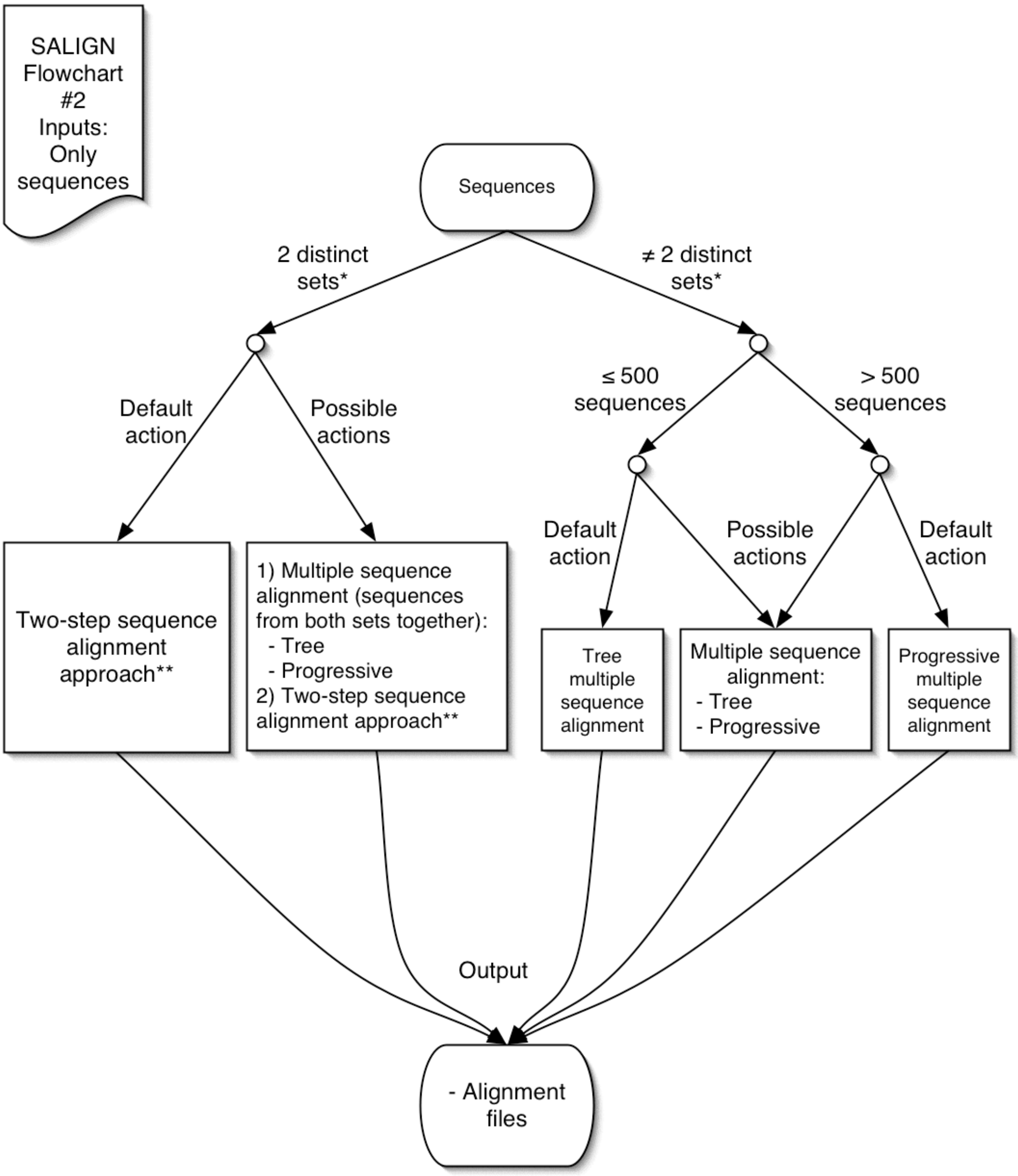
Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., *et al.* (2004) Annotation transfer between genomes: protein-protein interologs and protein-dna regulogs. *Genome Res*, **14**, 1107–1118.

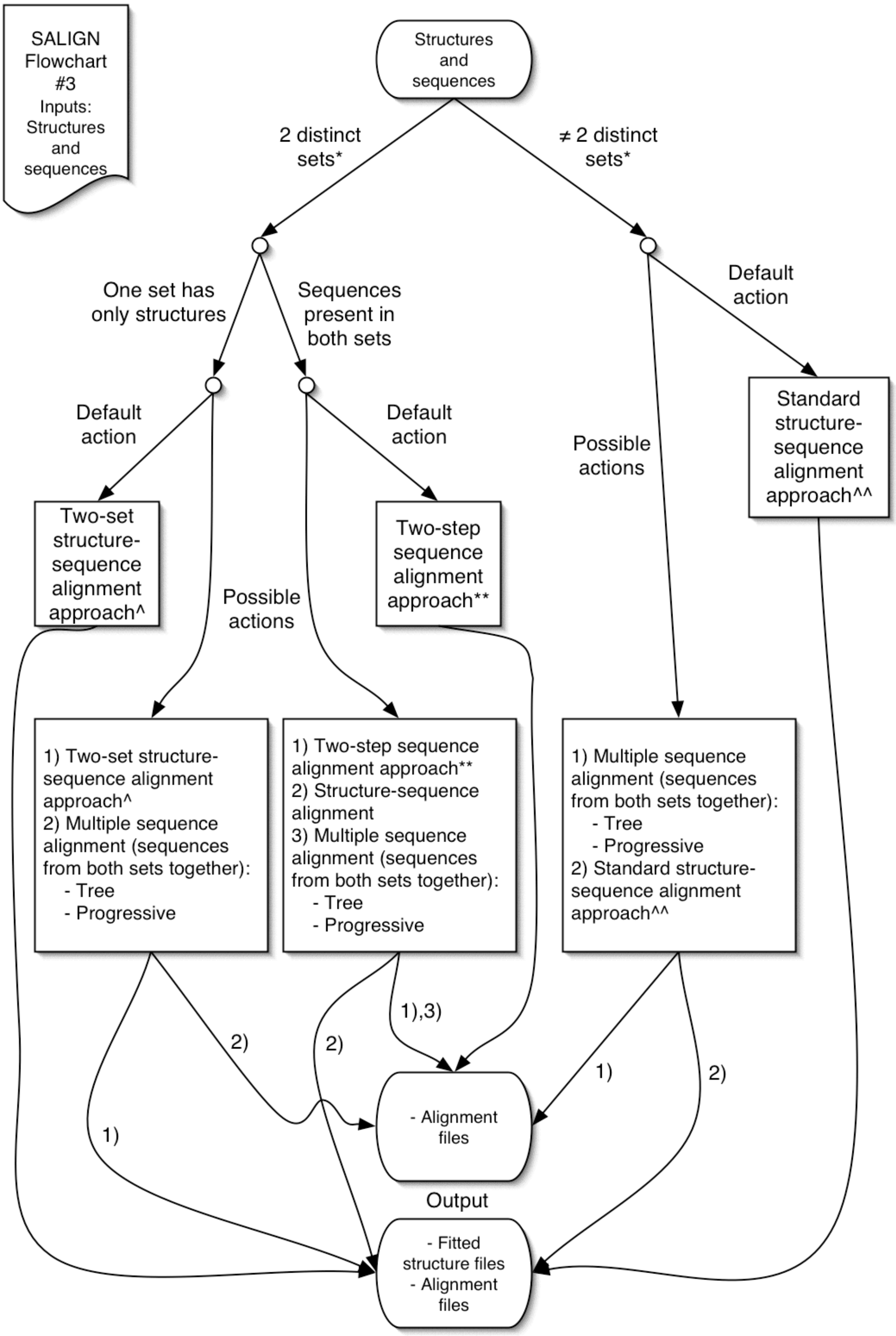
Xu, L. Z., Sanchez, R., Sali, A., Heintz, N. (1996) Ligand specificity of brain lipid-binding protein. *J Biol Chem*, **271**, 24711-24719.

Zhu, Z. Y., Sali, A., Blundell, T. L. (1992) A variable gap penalty function and feature weights for protein 3-D structure comparisons. *Protein Eng,* **5**, 43–51.

Appendix


SALIGN Flowchart #1 Inputs: Only structures
Structures
≤ 50
> 50
Default action
Default action
Tree multiple structure alignment
Progressive multiple structure alignment
Possible actions
Output
Output
- Fitted structure files
- Alignment files
1) Multiple structure alignment:
- Tree
- Progressive
2) Multiple sequence alignment:
- Tree
- Progressive
- Fitted structure files
- Alignment files
2)
1)
Output
Output
- Alignment files
- Fitted structure files
- Alignment files


SALIGN Flowchart #2 Inputs: Only sequences
Sequences
2 distinct sets*
≠ 2 distinct sets*
Default action
Possible actions
≤ 500 sequences
> 500 sequences
Default action
Possible actions
Default action
Two-step sequence alignment approach**
1) Multiple sequence alignment (sequences from both sets together):
- Tree
- Progressive
2) Two-step sequence alignment approach**
Tree multiple sequence alignment
Multiple sequence alignment:
- Tree
- Progressive
Progressive multiple sequence alignment
Output
- Alignment files


SALIGN Flowchart #3
Inputs: Structures and sequences
Structures and sequences
2 distinct sets*
≠ 2 distinct sets*
One set has only structures
Sequences present in both sets
Default action
Default action
Default action
Possible actions
Possible actions
Possible actions
Two-set structure-sequence alignment approach^
Two-step sequence alignment approach**
Standard structure-sequence alignment approach^^
1) Two-set structure-sequence alignment approach^
2) Multiple sequence alignment (sequences from both sets together):
- Tree
- Progressive
1) Two-step sequence alignment approach**
2) Structure-sequence alignment
3) Multiple sequence alignment (sequences from both sets together):
- Tree
- Progressive
1) Multiple sequence alignment (sequences from both sets together):
- Tree
- Progressive
2) Standard structure-sequence alignment approach^^
2)
1)
2)
1),3)
1)
2)
- Alignment files
Output
- Fitted structure files
- Alignment files

SALIGN
Flowcharts

This page contains descriptions of terms introduced in the flowcharts, as well as information about additional output files.

*
Set of sequences/structures:

Any number of sequences/structures that have been uploaded "together". Each uploaded alignment file is considered one set, and all pasted sequences are grouped together as one set.

**
Two-step sequence alignment approach:

Step 1: The two sets are multiply aligned (sequence-sequence) independently. Sets consisting of more than 500 entries are not aligned in step 1 and should thus be prealigned.

Step 2: The resulting alignments from step 1 are aligned to each other by matching their profiles.

Even when applied to structures, this approach does only utilize sequence information.

^
Two-set structure-sequence alignment approach:

Step 1: The structure set is aligned using the structure-structure feature. The mixed set is aligned using the sequence-sequence feature.

Step 2: The resulting alignments from step 1 are aligned to each other by a structure-sequence alignment if neither set contains > 100 entries. For larger sets a profile-profile alignment is performed.

^^
Standard structure-sequence alignment approach:

Step 1: Independent multiple alignments of structures (structure-structure) and sequences (sequence-sequence) are performed, regardless of the distributions in the uploaded files.

Step 2: The resulting alignments from step 1 are aligned to each other by a structure-sequence alignment if neither set contains > 50 entries. For larger sets a profile-profile alignment is performed.

In addition to the output files in the flowcharts, all output packages include MODELLER log files, which give details pertaining to the alignment process, and MODELLER input files, which can be used with any stand-alone version of MODELLER, version 8 and higher. If a tree alignment has been performed, a dendrogram file is also provided.