Molsoft LLC
SBIR Phase II - Final Scientific/Technical Report
Project Title: "GAP - Genomics Annotation Platform"
(Grant No. DE-FG03-01ER83282)

Introduction

The Genomics Annotation Platform (GAP) was designed to develop new tools for high throughput functional annotation and characterization of protein sequences and structures resulting from genomics and structural proteomics, benchmarking and application of those tools. Furthermore, this platform integrated the genomic scale sequence and structural analysis and prediction tools with the advanced structure prediction and bioinformatics environment of ICM. The development of GAP was primarily oriented towards the annotation of new biomolecular structures using both structural and sequence data. Even though the amount of protein X-ray crystal data is growing exponentially, the volume of sequence data is growing even more rapidly. This trend was exploited by leveraging the wealth of sequence data to provide functional annotation for protein structures. The additional information provided by GAP is expected to assist the majority of the commercial users of ICM, who are involved in drug discovery, in identifying promising drug targets as well in devising strategies for the rational design of therapeutics directed at the protein of interest. The GAP also provided valuable tools for biochemistry education, and structural genomics centers. In addition, GAP incorporates many novel prediction and analysis methods not available in other molecular modeling packages.

This development led to signing the first Molsoft agreement in the structural genomics annotation area with the University of Oxford Structural Genomics Center. This commercial agreement validated the Molsoft efforts under the GAP project and provided the basis for further development of the large scale functional annotation platform.

Project Objectives

The following objectives have been achieved under this grant:

- Developing a new paradigm for annotation, visualization and distribution of functional and structural information with a compact single file containing both data and interactive graphics environment. These files can replace the current text files and give biologists and chemists a much more powerful way to look at proteins. This platform is already in use by the University of Oxford Structural Genomics Center.

- Implementing a genome browser with a focus on genes/proteins with the following capabilities: rapidly reading and displaying genomic data from the National Center for Biotechnology Information (NCBI), reading single nucleotide polymorphism (SNP) data as a table, selecting organisms from a hierarchical taxonomic tree, and allowing text and graphical annotation of genomes. All of these features are accessible through a graphical user interface (GUI). The unique feature of this Molsoft proteome browser

is that the whole human genome can be read into a local "native" browser. Also several genomes can be loaded and displayed side-by-side.

- Developing tools and procedures to consider biological symmetry in the context of structural proteomics.
- Developing the Molsoft XPDB database to include structural information on the biologically relevant complex present in the Protein Data Bank BIOMT records as well as including structural annotation of ligand binding sites, SNPs, and functional domains.
- Developing and implementing a robust and sensitive method to calculate the evolutionary conservation of residues which can be applied to large scale functional annotation. Functionally important residues, such as those involved in ligand binding, catalysis, and protein-protein interactions, are usually more conserved, so that evolutionary conservation is useful for predicting these sites (published).
- Implementing a novel method for identifying variable size surface regions with a high solvation energy and displaying them on a graphical representation of the protein surface to aid in identifying protein-protein interaction sites (published and available as a module in commercial version of the ICM program).
- Developing an efficient program for Support Vector Machines that is accessible through both the ICM scripting language and the GUI (implemented as a module in the commercial version of the program).
- Developing and implementing an accurate method, utilizing Support Vector Machines and evolutionary conservation, to predict protein-protein interaction interfaces (published).

Commercialization and Product Development

The GAP project led to four types of outcomes. First, new tools and interfaces were added to the established molecular visualization and modeling program, ICM. Some of them are full fledged separately licensed modules (e.g. SVM) and others have the new functionality implemented as extensions to the existing modules. Second, new data were generated for existing protein structures in the PDB and the annotated files were generated in the form of the XPDB database available from Molsoft. Third, a web server, ODA, was created to serve the protein interface prediction via the Internet. This server uses a computer cluster at Molsoft. Fourth, the technology (single file animated protein structures) is licensed to structural genomics centers so that they can prepare those files and distribute them.

ICM is an integrated molecular modeling program with capabilities including protein structure and sequence analysis, flexible ligand docking, homology modeling, and cheminformatics tools [1]. These functions are accessible either through both the ICM scripting language or a graphical user interface. There are currently two commercial versions of ICM and one free version, ICM-Browser, for reading structure files either in PDB or proprietary binary formats and displaying them. The commercial versions include the full feature one,

ICM-Pro, and an extended browser, ICM-Browser-Pro, that allows the user to also modify and save structures. The genome browser, REVCOM evolutionary conservation, Support Vector Machines, and protein-protein interface prediction functionality will eventually be included in ICM-Pro. There is also currently the ability to market portions of the program as separate modules for ICM-Pro. The Support Vector Machines and other machine learning algorithms are currently available as a separate module of the commercial versions of ICM. The other newly developed functions may be transferred to such a module in the future. The new functions developed under GAP are also included in future plans for ICM-Bio, a version targeted specifically at research biologists and which has all structural and sequence analysis capabilities but not virtual ligand screening functions.

Multiple Genome Browser and Editor

The rapidly growing quantity of genomic data and its increasing importance to biological and pharmaceutical research makes having a browser capable of visualizing and annotating such data an important addition to the ICM modeling environment. The selection of an organism for study is facilitated by a GUI window with the hierarchical taxonomic tree. The user simply navigates the tree and clicks on the organism of interest to load the genome data. Initially only basic information such as gene names and their location on the chromosome are read and displayed for fast loading and visualization. Individual genes may then be annotated either with different border or fill colors or with text through the GUI or using the scripting language. The newly annotated genome may then be saved as a file for later use.

Data for individual genes may also be accessed through the browser. The SNP data for a particular gene, which describes variations in the gene sequence among different individuals, may be read as a table for analysis. This table also has links to the NCBI dbSNP database, which contains detailed information such as the frequencies of occurrence and known relations with disease. When a structure of the protein product is available, this SNP information may be visualized on the structure to give insight into its effect on protein function. The DNA sequence of the gene or amino acid sequence of its products can also be read through the browser and used for further analysis.

Tree Viewer, Clustering and Microarray Analysis

An interactive tree viewer has also been added to ICM. It allows a tree data structure to be visualized in a separate frame. The taxons may be colored and subtrees may be collapsed or expanded using the mouse. This can be used for viewing trees generated from many applications including multiple sequence alignments, taxonomic analysis, gene ontology (GO) hierarchy, and microarray clustering.

Distance based clustering methods, including the weighted UPGMA method and neighbor-joining method, have also been implemented. These can be applied to a multitude of similarity or distance data including chemical descriptors for cheminformatics, sequence similarity, and microarray data. Since these clustering methods are hierarchical, the results may be naturally visualized using the tree viewer described above.

Microarray data analysis using ICM brings together the genome browser, clustering and tree viewer functionality. The microarray data is first normalized and clustered. Next gene annotation data may be linked to the microarray data and the clustering tree visualized. The clustering level is changed interactively within the clustering tree display by using the mouse. Individual genes in the clusters, which often are involved in common pathways or have related functions, may then easily be examined using the table viewer. An example of this type of analysis is shown in Figure 1.

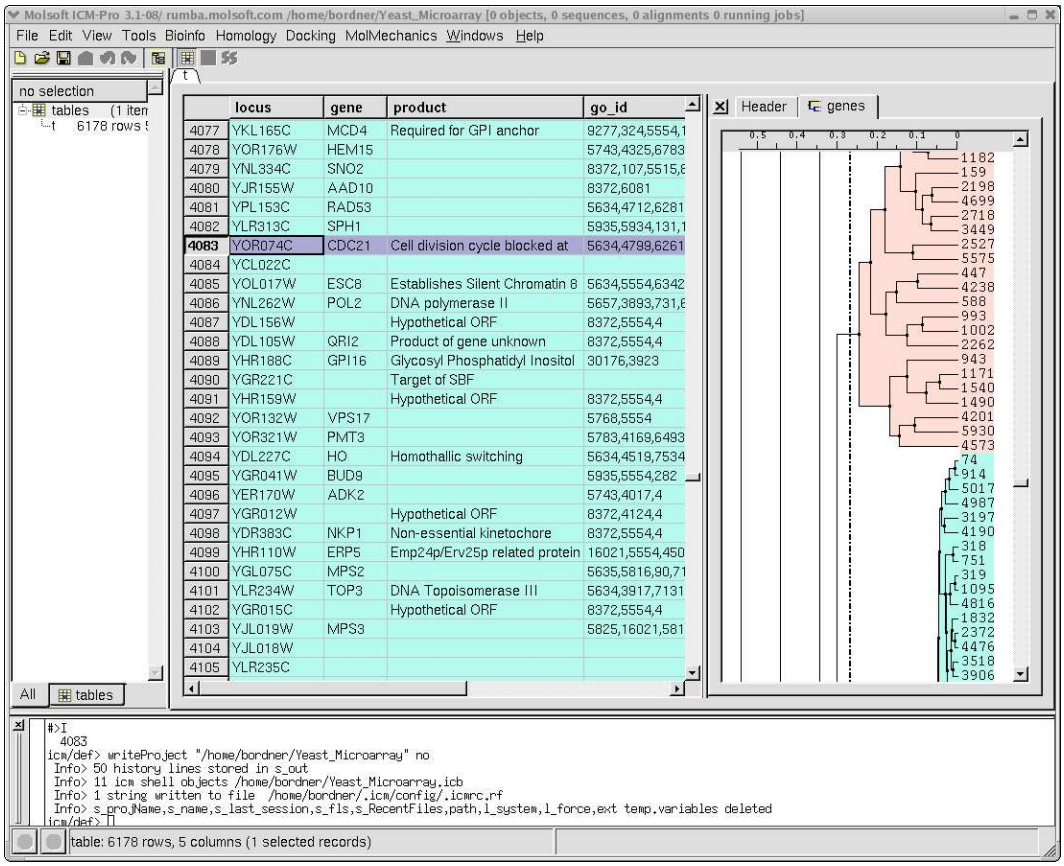


Figure 1: *Cluster analysis of genome-wide mRNA transcript level time series data for the yeast cell cycle collected using microarrays (data from Cho et al. [2]). Genes with a similar pattern of expression during the cell cycle are clustered to reveal functional associations. For example, the cluster highlighted in blue contains many genes known to be involved in DNA replication.*

Structural Database with Biological and Functional Annotation

The XPDB database is a Molsoft product that contains protein structures, with annotation derived from sequence databases, stored in a compact binary format that may be read and

displayed by ICM-Browser, ICM-Browser-Pro, or ICM-Pro. This product has been substantially expanded to include information on the biologically relevant complex as well as expanded annotations from the UniProt and NCBI genomic databases. A new function to generate the structure for the biologically relevant complex uses BioMT information present in the Protein Data Bank (PDB) file. Although this information was not generally present in PDB format files before 1999, it is included in the corresponding mmCIF format files. The mmCIF files for older PDB entries were therefore parsed and a table of the BioMT information generated for them. The X-ray crystallographic structure may contain multiple copies of the biological unit or the constituent proteins may be arranged differently so that additional processing is necessary to generate the structure of the complex. Only the symmetry information necessary to generate the structure of the complex is stored in the XPDB files in order to make them compact and to insure fast browsing of structures with little computational overhead. A user then simply clicks a button on the GUI with the mouse to dynamically generate the full structure of the complex. This functionality is important because the complete quaternary structure is often essential in generating protein models for virtual ligand screening as the natural substrate may bind to the interface between two protein subunits. This increases the value of the XPDB product and its utility for drug target selection and characterization. Additional sequence information has also been added to the XPDB structural annotation. This includes information on ligand binding sites, SNPs, and functional sites retrieved from a recently released UniProt database and NCBI genomic data files. In addition, graphical links to the NCBI dbSNP database have been added. When the user clicks on a SNP displayed on the structure, a web browser automatically loads the corresponding NCBI webpage with detailed information such as experimental methods, references, frequencies in populations, and phenotypes. The annotation of SNPs on the structure of a drug target is invaluable because individual genetic variations may affect a patient's response to the drug. It may also help identify variations that contribute to drug resistance. An example is shown in Figure 2.

REVCOM Evolutionary Conservation Method

The degree of evolutionary conservation varies among amino acids in a protein because of functional constraints [3, 4]. Residues that are important for a protein's tertiary structure and folding, enzymatic function, ligand binding or interaction with other proteins are generally more conserved. Therefore the identity of these conserved residues gives important clues to a protein's structure as well as predicting binding sites to target with virtual ligand screening in the search for a new drug. Conservation calculations begin with a multiple sequence alignment containing the protein sequence of interest. The simplest methods then calculate the conservation using the residue frequencies in each alignment column. One of the most widely used measures of this type is the Shannon entropy. However, there are problems with these methods including their dependence on the choice of sequences and their lack of an evolutionary model. This may be corrected by including phylogenetic relationships between the sequences via a phylogenetic tree calculated from the multiple sequence alignment, as our new method does. Distantly related sequences in the alignment, while they have the potential to provide strong evidence of conservation, also generally introduce errors in the

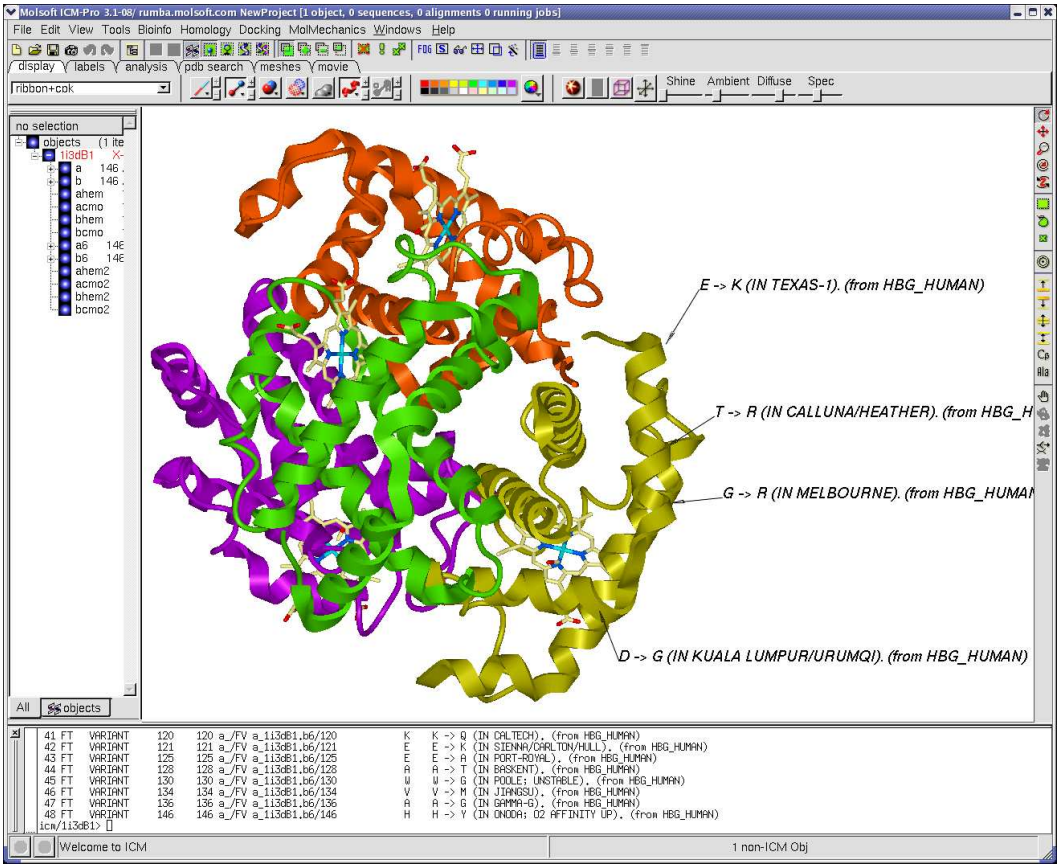
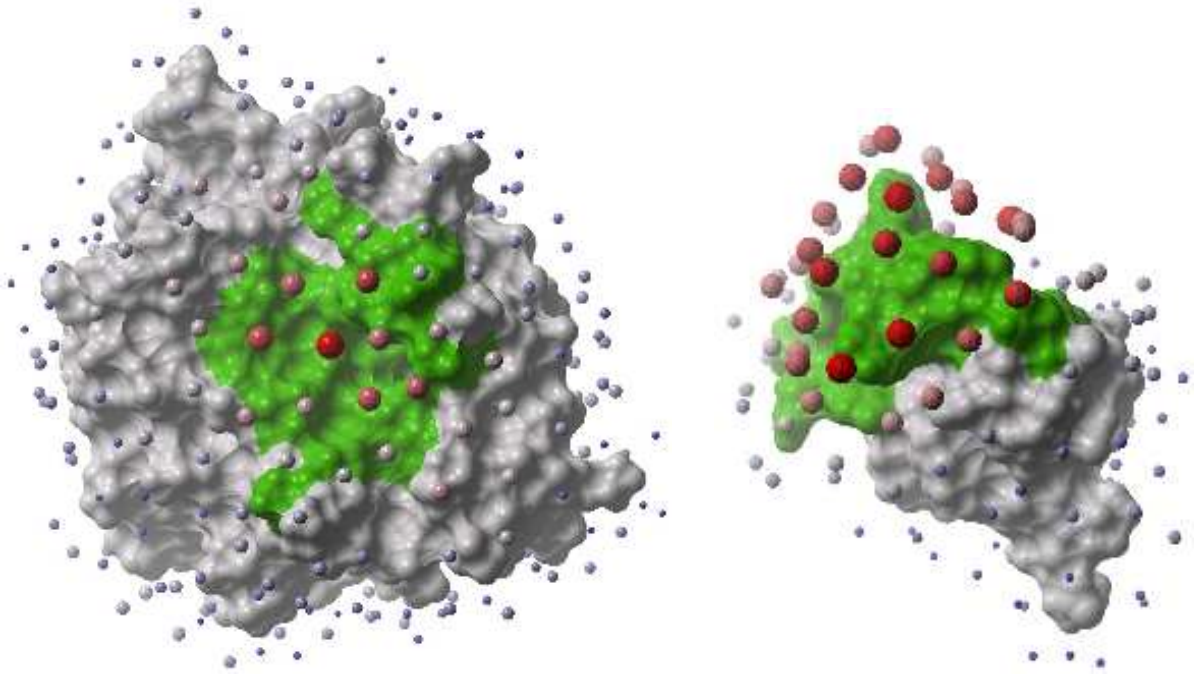


Figure 2: *An example of the updated XPDB database, which incorporates information on the biological relevant complex. The structure of Hemoglobin (Hb) Bart's (PDB entry 1I3D), an abnormal tetrameric γ_4 hemoglobin associated with α-thalassemias, is shown here with SNP annotation from UniProt. The subunits are shown in different colors. The coordinates of the complex were not contained in the original file but were generated by applying crystal symmetry transformations.*

alignment and subsequent conservation calculation. These errors include local alignment errors in divergent sequence segments, alignment to a nonhomologous sequence mistakenly included in a database search, and uncertainties in the substitution matrices at large evolutionary distances. The Robust Evolutionary Conservation Measure (REVCOM) method uses phylogenetic trees to reduce dependence on the choice of sequences, Bayesian estimation of evolutionary rates to avoid overfitting, and consistently incorporates alignment reliability into the conservation calculation [5]. A large non-redundant data set of 1494 protein-protein interface structures from the PDB was compiled to test the REVCOM method as well as the protein-protein interface prediction method described below. Upon comparison with the entropy conservation measure, the REVCOM method was found to detect the higher conservation of interface residues for a larger number of interfaces in the data set. It was also found to give conservation values that were more stable to the introduction of distantly related sequences than the entropy measure.

Optimal Desolvation Area as a Predictor of Protein-Protein Interaction Sites

Hydrophobic interactions are known to be an important contribution to protein-protein binding affinity [6, 7], particularly for obligatory complexes [8, 9]. It was also found that rigid body docking results cluster around known protein-protein interaction sites except with different orientations [10], possibly due to favorable desolvation energies. Thus the identification and visualization of protein surface patches with low desolvation potentials can aid in predicting protein-protein interaction sites. The novel Optimal Desolvation Area (ODA) method calculates the desolvation energy for a series of concentric spheres, centered at regularly spaced points above the protein surface [11]. The sphere yielding the lowest desolvation potential for each center is then chosen. The resulting desolvation potential for each center is then visualized by spheres at each center of sizes proportional to the value of the potential. Examples of ODA calculation results for two proteins, chymotrypsin and HPTI, are shown in Figure 3.



(a) (b)

Figure 3: *Surface points around unbound proteins colored according to the optimal desolvation energy values (red are the lowest energy values). Surface points with the lowest energies are seen to cluster around the known interface residues that are colored in green. (a) shows chymotrysin's interface with APPI (PDB entry 1CA0) and (b) shows HPTI's interface with chymotrypsinogen (PDB entry 1CGI).*

Support Vector Machines

Support Vector Machines (SVMs) are powerful kernel based methods for supervised learning. They can efficiently learn nonlinear patterns from a set of examples and generalize to make predictions. They have been successfully applied to a wide range of problems including optical pattern recognition, text catagorization, time-series prediction, gene expression profile analysis, and DNA and protein sequence analysis [12]. The data is first mapped to feature space using a nonlinear function, and the SVM learning procedure then minimizes a weighted sum of the empirical risk and the complexity to find the optimal separating hyperplane in feature space.

SVMs for classification and regression have been implemented in ICM. An efficient chunking optimization method for learning that solves a series of two-dimension quadratic programming problems has been used for speed. All of the standard kernel functions (linear, polynomial, Gaussian, and sigmoidal) have been implemented. Functions are also included for cross validation, which provides the most reliable assessment of model accuracy. All SVM functions are accessible from the scripting language as well as a GUI interface, and reasonable default values are provided for parameters as an aid to novice users. Because SVMs have a wide range of applications in computational biology, it is expected that they will be utilized for a variety of future predictive methods in ICM.

SVM Prediction of Protein-Protein Interfaces

Protein-protein interactions are important for many biochemical and biological functions in an organism. Although high-throughput experimental methods, such as yeast two-hybrid screens and mass spectrometry of coimmunoprecipitated complexes, can be used to find interacting proteins in a proteome, there are many false positives and negatives [13]. More importantly, detailed structural information of the intermolecular interactions is necessary both for understanding the function of the interaction and to design drugs to modulate the interaction. However such information is only available from the expensive and time consuming methods of obtaining X-ray crystallography or nuclear magnetic resonance (NMR) structures of the complexes, which can be applied to only a small fraction of an organism's proteome. Thus a fast computational method to predict protein-protein interaction interfaces, given the structure of one partner, yields important information that may either be confirmed by a targeted set of alanine-scanning mutagenesis experiments or used to guide computational docking of the component proteins. It may also aid a drug discovery project by providing valuable information to guide selection of small molecule binding pockets to target for virtual ligand screening.

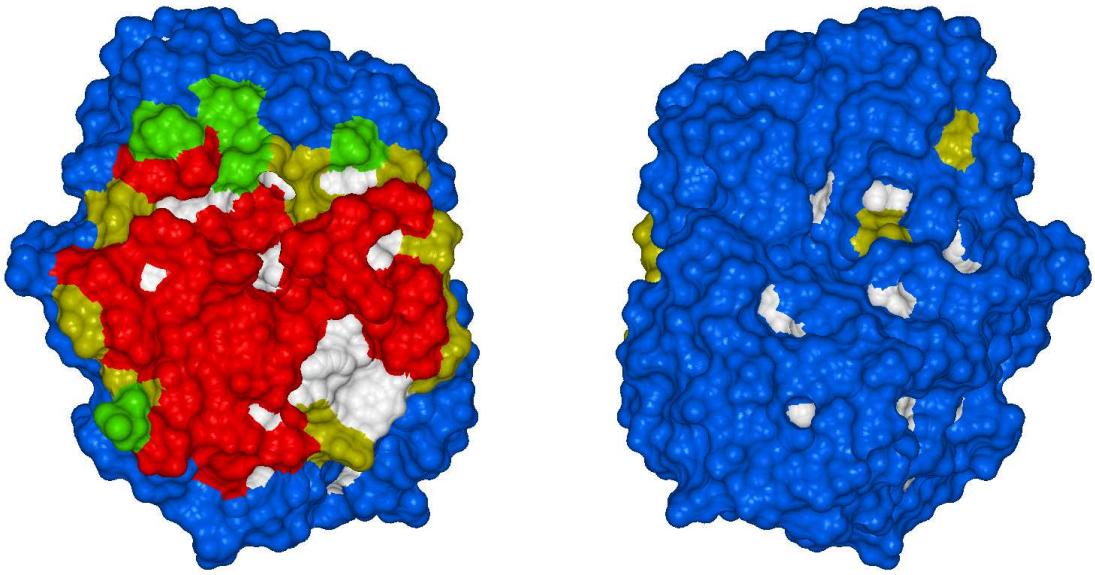
While evolutionary conservation can be used to predict protein-protein interaction interfaces, other features also distinguish them from the rest of the protein surface. These additional properties include hydrophobicity, solvation energy, and residue type propensity. Protein-protein interfaces tend to be hydrophobic, have a high solvation energy, and contain a higher proportion of large hydrophobic and polar residues and a lower proportion of charged residues than the remaining protein surface [8]. In addition, protein interfaces are comprised of localized clusters of surface residues. These other properties as well as

residue conservation should all be accounted for in an accurate and sensitive protein-protein interface prediction method.

An SVM trained on Z-scores for residue frequencies in columns from a multiple alignment of close homologs and evolutionary rates calculated using the REVCOM method was used to predict protein-protein interfaces [14]. Transforming the data using Z-scores is important because SVMs work best with homogeneous input data. These values were calculated for surface patches containing 15 residues and used to train the SVM. Because hydrophobicity and solvation energy were found to be highly correlated and did not contribute to the prediction accuracy, probably because they are also correlated with the residue type frequencies, they were not included in the SVM input data. A subset of the 1494 non-redundant protein-protein interfaces described above, containing only dimers, was used for SVM training and validation. Five fold cross validation on the set of 632 dimer interfaces was used to assess prediction accuracy. A high fraction of the predicted interfaces, 97%, overlapped the actual interfaces even though only 22% of the surface residues were included in the average patch. This indicates that the prediction is not only accurate but also generally applicable because the predictions were performed only for data not used to train the model. The prediction result for *E. coli* cytidine deamine (PDB entry 1ALN) is shown in Figure 4 as an example. This figure illustrates the general feature of the prediction results: central interface residues are more accurately predicted than those near the interface boundary. This was attributed to higher evolutionary conservation for central residues, a higher fraction of interface residues in the surface patches, and arbitrariness in the definition of the boundary of the actual protein-protein interface.

Reliability estimates of the interface residue predictions can be used to identify strongly predicted residues. Then, for example, experimental verification of the prediction using alanine-scanning mutagenesis experiments could be first performed only for those residues with high prediction reliabilities. The reliability estimate was implemented using a histogram technique for the SVM that relies on the fact that data points far from the separating hyperplane in feature space are more likely to belong to the corresponding class (interface or non-interface residue) than points near it. Furthermore, central interface residues were found to have a higher reliability, on average, than peripheral interface residues. This was attributed to the higher evolutionary conservation of the central residues and the higher proportion of non-interface residues in peripheral patches.

The SVM protein-protein interface prediction was also able to identify interfaces not present in the original X-ray crystal structure used for the prediction. This was particularly evident for transient heterodimers involved in intracellular signalling because of multiple binding partners, some of which are absent from the structure. For example, there are structures of complexes of the GTP-binding nuclear protein Ran with four different binding partners in the PDB: RCC1, NTF2, importin β, and the Ran binding domain of RanBP2. A prediction for Ran based on the RCC1 complex gave four separate predicted interface patches, only one of which corresponded to the RCC1 binding site. However all four predicted interfaces corresponded with the actual interface of at least one binding partner. This indicates that the new prediction method may be used to identify the protein-protein binding interfaces of alternative binding partners. The prediction results also indicated probable misannotation of oligomeric states for some proteins. For example, a prediction based on a structure of *T. maritima* glycl-tRNA synthetase (PDB entry 1J5W) reveals a large predicted



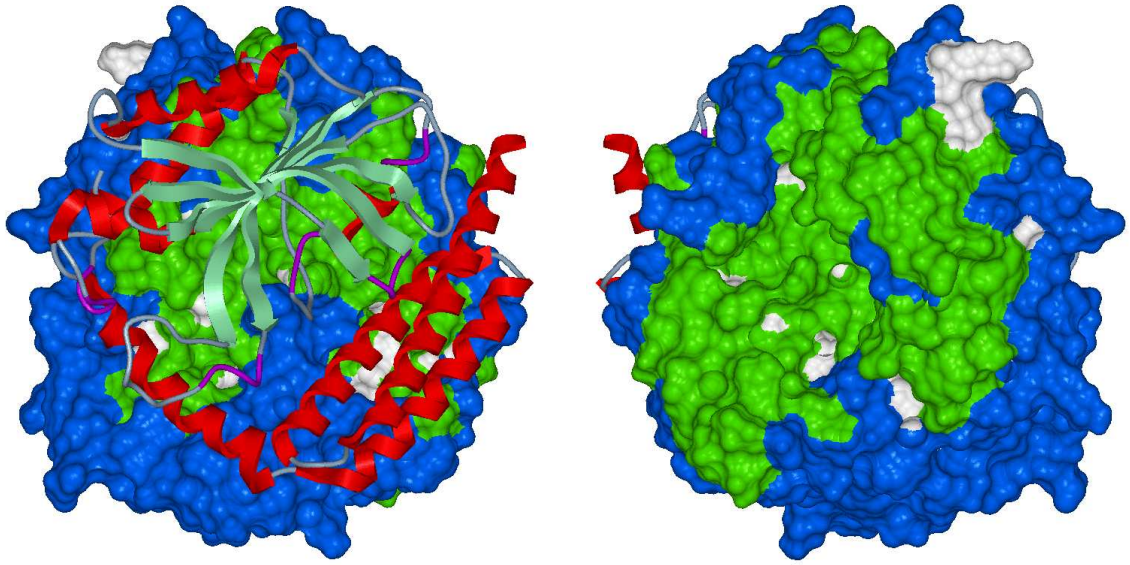
(a) front (b) back

Figure 4: *Prediction results for E. coli cytidine deaminase (PDB entry 1ALN). The molecular surface is colored according to the prediction results as follows: red - correctly predicted interface residue, blue - correctly predicted non-interface residue, yellow - non- interface residue predicted as interface one, and green - interface residue predicted as non-interface one. Arbitrariness in the definition of the interface boundary and weaker prediction results there contribute to the fact that the all central interface residues are correctly predicted. Almost all non-interface residues on the opposite side are also correctly predicted.*

interface opposite the correctly predicted known homodimeric interface (see Figure 5). This indicates that the actual complex is likely an $\alpha_2\beta_2$ tetramer, like most other eubacterial glycl-tRNA synthetases, and not a homodimer as the PDB file indicates. This discrepency in the annotation is probably due to the lack of the β subunit in the X-ray crystal structure as well as the fact that most archael and eukaryotic orthologs are homodimers.

Conclusion

The Genomics Annotation Platform project led to the development of new annotation tools and features that complement the existing functionality of the ICM molecular modelling and drug discovery program. First genome information can be rapidly read and manipulated through an easy to use GUI interface. This information has also been used to add annotation, such as SNPs, to structural models in the XPDB database. Also the XPDB database has been expanded to include information on the biologically relevant complex obtained from the PDB. Next, a new robust and sensitive method to calculate the evolutionary conservation of residues has been added. This is expected to be useful for identifying functionally important residues. Two methods for predicting protein-protein interfaces, which



(a) front (b) back

Figure 5: *Protein-protein interface prediction results for T. maritima glycl-tRNA synthetase (PDB entry 1J5W). The predicted interface surface is colored in green and the remaining surface in blue. The binding partner in the homodimer structure is shown in ribbon format. Although one predicted interface patch coincides with the actual interface in the structure of the complex another predicted interface is evident on the opposite face. This suggests that this enzyme forms an $\alpha_2\beta_2$ tetramer, as do most eubacterial orthologs, rather than a homodimer, as the indicated in the PDB file.*

may be drug targets, have been implemented in ICM. One method, the Optimal Desolvation Area method, identifies variable size surface regions with high solvation energy, and the other method uses SVMs to predict protein binding interfaces. The efficient implementation of SVMs in ICM are also expected to be useful for other prediction tasks encountered in structural modelling, cheminformatics. Overall the additional annotation of protein structural models provided by GAP will assist users in identifying promising protein regions to target for virtual ligand screening.

References

[1] Molsoft, LLC. ICM software manual. Version 3.0. 2004.

[2] Cho R.J., Campbell M.J., Winzeler E.A., Steinmetz L., Conway A., Wodicka L., Wolfsberg T.G., Gabrielian A.E., Landsman D., Lockhart D.J., Davis R.W.. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 1998;2:65–73.

[3] Fitch W.M., Margoliash E.. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem Genet 1967;1:65–71.

[4] Uzzell T., Corbin K.W.. Fitting discrete probability distributions to evolutionary events. Science 1971;172:1089–1096.

[5] Bordner A.J., Abagyan R.A.. REVCOM: A robust Bayesian method for evolutionary rate estimation. *submitted for publication.*

[6] Vakser I.A., Aflalo C.. Hydrophobic docking: a proposed enhancement to molecular recognition techniques. Proteins 1994;20:320–329.

[7] Young L., Jernigan R.L., Covell D.G.. A role for surface hydrophobicity in protein-protein recognition. Protein Sci 1994;3:717–729.

[8] Jones S., Thornton J.M.. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.

[9] Tsai C.J., Lin S.L., Wolfson H.J., Nussinov R.. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci 1997;6:53–64.

[10] Fernandez-Recio J., Totrov M., Abagyan R.. Identification of protein-protein interaction sites from docking energy landscapes. J Mol Biol 2004;335:843–865.

[11] Fernandez-Recio J., Totrov M., Skorodumov C., Abagyan R.. ODA (optimal desolvation area): New predictor for protein-protein interaction sites. *to appear in Proteins.*

[12] Müller K., Mika S., Rätch G., Tsuda K., Schölkopf B.. An introduction to kernel-based learning algorithms. IEEE Trans Neural Net 2001;12:181–202.

[13] Legrain P., Wojcik J., Gauthier J.M.. Protein–protein interaction maps: a lead towards cellular functions. Trends Genet 2001;17:346–352.

[14] Bordner A.J., Abagyan R.A.. Statistical analysis and prediction of protein-protein interfaces. *to appear in Proteins.*