BMC Genomics

BioMed Central

Research

The FEATURE framework for protein function annotation: modelling new functions, improving performance, and extending to novel applications

Inbal Halperin[†1], Dariya S Glazer[†1], Shirley Wu[†2] and Russ B B Altman[*2,3]

Address: [1]Department of Genetics, 318 Campus Drive, Clark Center S240, Stanford, CA 94305, USA, [2]Program in Biomedical Informatics, MSOB X-215, 251 Campus Drive, Stanford, CA 94305, USA and [3]Department of Bioengineering, 318 Campus Drive, Clark Center S170, Stanford, CA 94305, USA

Email: Inbal Halperin - inbal@helix.stanford.edu; Dariya S Glazer - dsglazer@stanford.edu; Shirley Wu - shwu19@stanford.edu; Russ B B Altman* - russ.altman@stanford.edu

* Corresponding author †Equal contributors

from IEEE 7th International Conference on Bioinformatics and Bioengineering at Harvard Medical School
Boston, MA, USA. 14–17 October 2007

Published: 9 September 2008

BMC Genomics 2008, **9**(Suppl 2):S2 doi:10.1186/1471-2164-9-S2-S2

This article is available from: http://www.biomedcentral.com/1471-2164/9/S2/S2

Abstract

Structural genomics efforts contribute new protein structures that often lack significant sequence and fold similarity to known proteins. Traditional sequence and structure-based methods may not be sufficient to annotate the molecular functions of these structures. Techniques that combine structural and functional modeling can be valuable for functional annotation. FEATURE is a flexible framework for modeling and recognition of functional sites in macromolecular structures. Here, we present an overview of the main components of the FEATURE framework, and describe the recent developments in its use. These include automating training sets selection to increase functional coverage, coupling FEATURE to structural diversity generating methods such as molecular dynamics simulations and loop modeling methods to improve performance, and using FEATURE in large-scale modeling and structure determination efforts.

Discussion

Introduction: importance and overview

A central goal of molecular biology is to understand the functions of proteins, including their catalytic properties, binding sites, cofactors, interaction partners, and subcellular localization. Traditional experimental methods for function characterization cannot cope with the rate at which genomics efforts are generating data. Computational methods for function recognition require far less time and expense and so can augment experimental methods. Computational tools make it possible to query many proteins for many different functions at varying levels of specificity, from general enzymatic activity to binding sites.

Usually, computational methods require either the sequence or structure of the molecule of interest. One effective approach in sequence-based function prediction methods is to compare the known sequence to a collection of sequences whose functions are known, whether on

a global or a local level. A high level of similarity found by such a comparison to an annotated sequence may allow the transfer of this annotation to the sequence of interest, based on presumed homology. BLAST [1] performs efficient sequence searches to facilitate such analyses. Searches within databases such as Pfam [2] and PROSITE [3], which contain models of short sequence motifs highly correlated with specific functions, may also allow function assignment based on sequence.

Inherently more important for the function of the molecule is its structure. The emergence of structural genomics (SG) has led to rapid advances in our knowledge of structure and structure determination. With the efficiency of structure determination methods now allowing high throughput experiments [4,5], the number of structures available in the Protein Data Bank (PDB) [6] is providing a wealth of insight into structure-function relationships. Furthermore, based on structures with known function, it should be possible to assign putative function to structures for which there exists no direct functional information. Annotation of molecular function by similarity is possible on the structural level as on the sequence level – by evaluating the similarity of global folds or local environments [7]. Structural similarity methods may employ chemical, physical, energetic or geometric criteria to recognize functional environments [8-10].

Many SG projects are targeting novel structures with low sequence identity to known proteins, in order to increase the ability to cover all fold families with at least one solved structure. Precise function can be reliably transferred only if sequence identity is at least 40% [11]; structure is significantly less conserved when sequence similarity is less than 50% [12]. As such, traditional sequence-based methods will not be enough to annotate a significant number of the novel protein structures being solved. Furthermore, with many of the proteins possessing novel folds, traditional global fold-based methods will be less effective. Consequently, there is a need for structure-based methods that do not depend on global fold similarity or exact conservation of residues or residue geometry.

Our group is actively interested in structure based function prediction, and has, to this end, developed a robust function recognition algorithm called FEATURE, which examines 3D environments of molecules in a way that is neither strictly sequence nor fold based. FEATURE represents the local environments of a macromolecule using descriptors that capture chemical, physical and geometric features. In this article we provide an overview of the FEATURE framework for predicting protein function. In particular, we present recent efforts in improving and enhancing FEATURE's functional coverage and efficiency, and applying FEATURE in novel ways.

An overview of the FEATURE system

The FEATURE system can be broken down into three major components. The first is the way in which sites, or local protein microenvironments, are represented; the second part concerns model building and supervised machine learning methods; and the third involves site scoring and model evaluation. FEATURE is flexible in the sense that each of these three components is adaptable to the specific needs of an application.

Microenvironment representation

One of the most important aspects of any structure-based protein function modeling system is how information about a protein is represented and calculated. Protein structure information can be especially complex, so simplified abstractions are used to capture relevant features in a way that is computationally tractable. Methods such as CASTp [13] employ geometric abstractions to describe the shape, area, and volume of surface pockets and internal cavities, which are often correlated with functional sites. Geometry is also used to determine the relative position of several amino acids to each other as in 3D templates [14]. Other representations involve calculating values for physicochemical properties associated with locations or elements in the structure, such as solvent accessibility, hydrophobicity, electrostatic potential, the presence of residues or secondary structure, conservation or the presence of chemical groups [15-24]. Jambon *et al*. use a representation that combines both geometry and property-based components [25].

FEATURE models a local protein microenvironment using a large number of physicochemical properties calculated at varying distances from the site (see Figure 1a for a simplified example). A site is defined as a 3D location in a protein structure, and its microenvironment is defined as a sphere centered on that location. In the typical use of FEATURE, 80 physicochemical properties (listed in Table 1) are computed in each of six 1.25 Å thick spherical shells – from 0 to 1.25, 1.25 to 2.5, 2.5 to 3.75, etc, up to 7.5 Å. A FEATURE vector represents the site as a list of 480 values (see Figure 1b for a simplified example). The FEATURE method has also been tested successfully on other segmentations of volume, such as a cubic lattice [26,27].

The concentric spherical shells representation has both advantages and disadvantages. One disadvantage is that information about orientation and the relative position of atoms is discarded. However, discrete shells are favorable because they allow statistics to be gathered over the relevant volumes and calculation is relatively efficient, which allows FEATURE to serve as an initial filter for more

Table 1: Physicochemical properties used by the FEATURE algorithm

Atom – based	Residue – based	Secondary structure – based
ATOM-TYPE-IS-C	RESIDUE_NAME_IS_ALA	SECONDARY_STRUCTURE1_IS_3HELIX
ATOM-TYPE-IS-CT	RESIDUE_NAME_IS_ARG	SECONDARY_STRUCTURE1_IS_4HELIX
ATOM-TYPE-IS-Ca	RESIDUE_NAME_IS_ASN	SECONDARY_STRUCTURE1_IS_5HELIX
ATOM-TYPE-IS-N	RESIDUE_NAME_IS_ASP	SECONDARY_STRUCTURE1_IS_BRIDGE
ATOM-TYPE-IS-N2	RESIDUE_NAME_IS_CYS	SECONDARY_STRUCTURE1_IS_STRAND
ATOM-TYPE-IS-N3	RESIDUE_NAME_IS_GLN	SECONDARY_STRUCTURE1_IS_TURN
ATOM-TYPE-IS-Na	RESIDUE_NAME_IS_GLU	SECONDARY_STRUCTURE1_IS_BEND
ATOM-TYPE-IS-O	RESIDUE_NAME_IS_GLY	SECONDARY_STRUCTURE1_IS_COIL
ATOM-TYPE-IS-O2	RESIDUE_NAME_IS_HIS	SECONDARY_STRUCTURE1_IS_HET
ATOM-TYPE-IS-OH	RESIDUE_NAME_IS_ILE	SECONDARY_STRUCTURE1_IS_UNKNOWN
ATOM-TYPE-IS-S	RESIDUE_NAME_IS_LEU	SECONDARY_STRUCTURE2_IS_HELIX
ATOM-TYPE-IS-SH	RESIDUE_NAME_IS_LYS	SECONDARY_STRUCTURE2_IS_BETA
ATOM-TYPE-IS-OTHER	RESIDUE_NAME_IS_MET	SECONDARY_STRUCTURE2_IS_COIL
ATOM-NAME-IS-ANY	RESIDUE_NAME_IS_PHE	SECONDARY_STRUCTURE2_IS_HET
ATOM-NAME-IS-C	RESIDUE_NAME_IS_PRO	SECONDARY_STRUCTURE2_IS_UNKNOWN
ATOM-NAME-IS-N	RESIDUE_NAME_IS_SER	
ATOM-NAME-IS-O	RESIDUE_NAME_IS_THR	
ATOM-NAME-IS-S	RESIDUE_NAME_IS_TRP	
ATOM-NAME-IS-OTHER	RESIDUE_NAME_IS_TYR	
HYDROXYL	RESIDUE_NAME_IS_VAL	
AMIDE	RESIDUE_NAME_IS_HOH	
AMINE	RESIDUE_NAME_IS_OTHER	
CARBONYL	CLASS1_IS_HYDROPHOBIC	
RING-SYSTEM	CLASS1_IS_CHARGED	
PEPTIDE	CLASS1_IS_POLAR	
	CLASS1_IS_UNKNOWN	
	CLASS2_IS_NONPOLAR	
	CLASS2_IS_POLAR	
	CLASS2_IS_BASIC	
	CLASS2_IS_ACIDIC	
	CLASS2_IS_UNKNOWN	
	PARTIAL-CHARGE	
	VDW-VOLUME	
	CHARGE	
	CHARGE-WITH-HIS	
	NEG-CHARGE	
	POS-CHARGE	
	HYDROPHOBICITY	
	MOBILITY	
	SOLVENT-ACCESSIBILITY	

In order to represent a local microenvironment, FEATURE determines the value of physicochemical properties in each of six concentric, spherical shells centered on the site of interest. Properties include those at the atom level, residue level, and secondary structure level.

expensive structure-based function prediction methods. Further advantages of this representation include unambiguous definition of a predicted site as a single point (i.e. Cartesian coordinates in the frame of the protein), accurate capture of properties of a cumulative nature such as partial charge, and the ability to change or add properties. The use of a single central point for each site means that models can be built with minimal prior knowledge of the geometry of the site – in other words, there is no need to establish other conserved points with which to define a non-spherical coordinate system. The use of spherical symmetry around this point also means that during search, each putative site center can be rapidly evaluated without the need to test alternative orientations around the point. Importantly, it allows identification of the physical and chemical features that are characteristic of functional sites, making the resulting models straightforward to interpret.

Model building by supervised machine learning

FEATURE uses supervised machine learning to combine significant properties into a model that can classify functional sites. In order to build a model, or description of a functional site, FEATURE requires two training sets. One consists of positive sites, which are 3D locations associated with positive examples of the function to be mod-

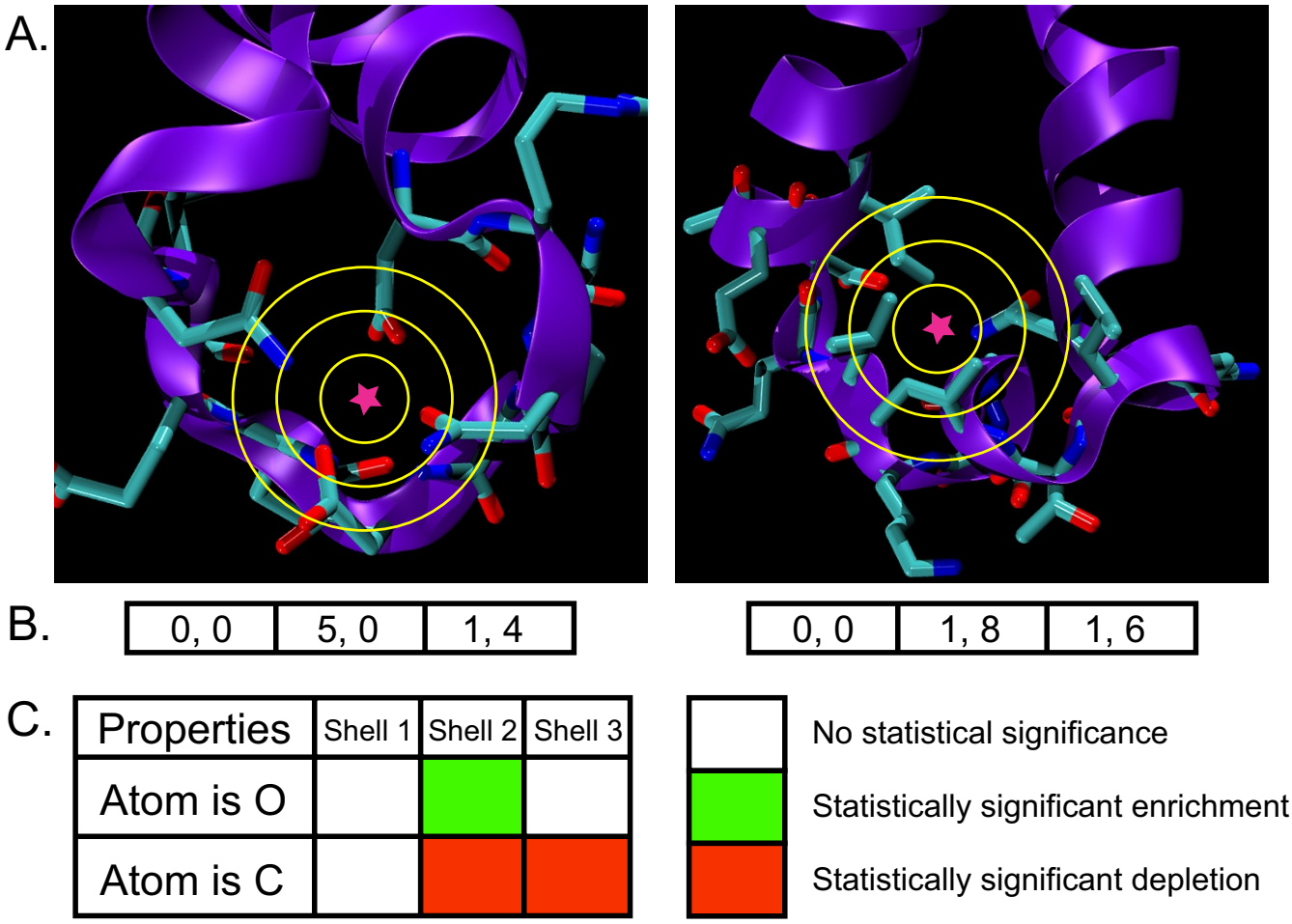

Figure 1
Simplified example for FEATURE model building. A. An example of a positive site (left) and negative site (right), and their respective microenvironments. Properties are calculated in concentric spherical shells centered on each site (star symbol). **B.** FEATURE vectors calculated from the images in **A**, with oxygen atom count being the first property, and carbon atom count the second. The vectors are divided by shell for clarity. **C.** An example of a visualized FEATURE model is shown, based on the FEATURE vectors in **B**, and images in **A**. In Shell 2, oxygen atoms are more abundant in the positive site (5 counts) than in the negative site (1 count) and so oxygen atom count is considered a significantly enriched property in Shell 2 of the model. In contrast, carbon atom count is less abundant in the positive site (0 counts) compared to the negative site (8 counts), so carbon atom count is considered a significantly depleted property in Shell 2 of the model. In Shell 3, both the positive and the negative sites have 1 oxygen atom, so the model contains no significant difference for oxygen atom count in Shell 3.

eled; the other consists of negative sites, which are 3D locations not known to be associated with the function (see Figure 1a). Negative sites can be chosen manually or automatically by randomly sampling 3D locations of structures in the PDB with a similar range of atom densities compared to the positive sites. FEATURE vectors are calculated for each site in the training set.

Given a set of FEATURE vectors, a distribution of values can then be collected for each property in each shell (see Figure 1b). We determine whether a property is significantly overrepresented, significantly underrepresented, or not significantly different in positive sites relative to negative sites in a given shell by comparing the positive and negative training set distributions for the property in that shell. The significance of a property for distinguishing sites from non-sites is calculated over all properties in all shells, and naïve Bayes [28] is used to weight the properties most informative for distinguishing the positive and negative sites. FEATURE models are visualized using "fingerprints", which are color-coded grids that depict the significance of each property in each shell (see Figure 1c). It is critical to stress that the choice of negative sites defines the background distribution for all features and thus

determines which features will be considered useful in identifying sites. Different models can result based on different strategies for defining the negative sites.

Site scoring and internal model evaluation

In order to determine performance statistics and score cutoffs for classification, the training sets are scored with the model, and sensitivity and specificity are estimated through k-fold cross-validation. Scores are calculated using a naïve Bayes scoring function, which operates on the assumption that the probability of a site belonging to a particular class is conditioned on the individual probabilities of observed, independent features. In the case of FEATURE, the features correspond to the physicochemical properties calculated in each shell, and their probabilities are derived from the training set distributions. A site's score is then the sum of the probabilities of obtaining an observed feature value given that the site is a positive site, taken over all significant features in the model. Score cutoffs are usually based on desired performance, and, as a default, are set to achieve 99% specificity on the training sets, as determined by cross-validation. In k-fold cross-validation, the training data is divided into k groups, and a model is trained on all but one of the groups and tested on the left out group.

Once a model is built and score cutoffs defined, potential sites can be scored using that model. FEATURE vectors are calculated for candidate sites in the same way as was done for training sites during model building, and scored using the same naïve Bayes scoring function. The resulting scores indicate the likelihood that the potential site is a positive site, depending on the score cutoff for that model. When available, the validity of every new model is assessed with an independent test set [18-20].

FEATURE in practice: workflow, training set selection, and manually-curated models

Creating a new model involves a typical workflow (see Figure 2) that begins by choosing a function of interest and defining a biologically reasonable definition of the Cartesian center point for that function (e.g. the central position in a binding site or the position of a key atom in an active site). Positive and negative training sets are then created and used to train the model. Cross-validation of the model on the training sets allows definition of score cutoffs based on desired performance, and whenever an independent test set is available, model performance can be further assessed. Once a model is built and a score cutoff has been defined, FEATURE can predict functional sites in structures of interest.

An especially important step in model training is the selection of sites for the positive training set, and, in order to tune performance, the negative training set. The first

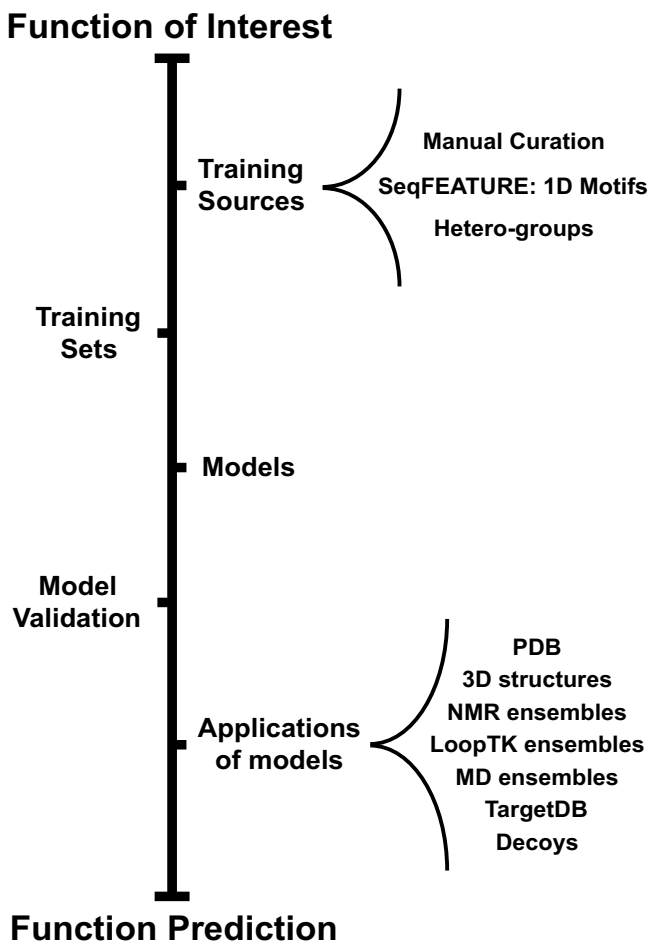

Figure 2

FEATURE framework overview. The outline of the steps necessary to predict a possible function for a protein is illustrated. In order to build a FEATURE model, one must first define the function of interest and create positive and negative training sets from the appropriate data sources. Then, the model is trained and evaluated on the training sets. The validated model can be used for function prediction. Certain steps in the outline, such as extracting training sets and model building are straightforward, as described in section "An overview of the FEATURE system". Other steps, such as determination of data sources for training sites and application of models, are more flexible. For example, training sites may be derived manually or automatically selected using annotated hetero-groups or sequence motifs. In addition, the resulting models can be applied towards static structures from the PDB or structure prediction decoys, or for dynamic function prediction over ensembles of structures generated using molecular dynamics simulation.

FEATURE models were manually curated in that the positive and negative training sets were built and verified by hand using published literature. These include calcium-binding [18] and ATP-binding [19] site models. The calcium-binding model has especially good performance,

and is currently being used in multiple ongoing projects to expand FEATURE's capabilities and applicability, described later in this overview. Our recently published zinc-binding model [29] is the best performing zinc binding predictor currently available. We have also applied FEATURE to function prediction in RNA structures with two magnesium binding models, one for diffuse binding and one for site-specific binding [30].

From its manually-curated beginnings, FEATURE has expanded to include automatic generation of training sets using sequence motifs, PDB annotations, and even a clustering of FEATURE vectors encompassing a non-redundant subset of the entire PDB. Functional coverage by the FEATURE system is enhanced when we employ multiple and diverse strategies for site selection. We describe our current work in the area of site selection in more detail below.

Increasing functional coverage

While having a highly specialized and performance-tuned model for recognizing a particular function is extremely valuable, it is becoming increasingly important to have wide coverage of protein function space. SG initiatives are causing a rapid expansion in the numbers of uncharacterized protein structures, many with very low sequence or even structural similarity to known proteins [31]. In order to expedite the annotation of structurally novel proteins, we need good and varied structure-based models of function. Structure-based models may also highlight heretofore unappreciated but interesting regions in partially characterized poly-functional proteins. Within the FEATURE framework, we have developed several strategies for expanding functional coverage.

SeqFEATURE – transforming 1D motifs into 3D models

Protein sequence data is extremely useful for deducing information about a protein's structure, interactions and function. Given its ubiquity, it comes as no surprise that there are numerous tools for recognizing function based on sequence. Pfam, Panther [32], PROSITE, and Superfamily [33] are just a few of the publicly available databases and methods for characterizing protein families or functions; many of them are conglomerated into single integrated tools like InterProScan [34] and ProFunc [8,10].

Most of the tools perform very well under most circumstances, but pattern matching tools such as PROSITE can be prone to false predictions and even the best tools, usually employing Hidden Markov Models, can be rendered less effective when sequence identity to known proteins is less than 30% [35]. 3D models have the potential to overcome this limitation, and can support a broader range of applications such as loop modeling and folding (see sections "Loop modeling" and "Decoy filtering").

In order to enhance both FEATURE's functional coverage and the performance of 1D motifs, we developed an extension to FEATURE, called SeqFEATURE, that transforms sequence-based models into structure-based ones [20,35]. Given a 1D motif, SeqFEATURE algorithm automatically extracts structures from the PDB that contain the motif to form a positive training set. One parameter that must be determined is the site center for each model. In the case of a 1D pattern, the center might be a functional atom on a functional residue contained in the pattern. SeqFEATURE finds all such 3D examples in a non-redundant subset of the PDB to be used as a positive training set. When a pattern contains more than one functional atom, multiple models are built centered on each one. The overlapping models can be used singly or in concert to predict the functional site.

Recently, we have applied SeqFEATURE to 44 regular expression patterns from the PROSITE database of functional motifs to produce a library of 136 automatically derived and trained models [35] (see section "Availability"). The models exhibit a wide range of performance, however, over three-quarters of them have an area under the curve (AUC) greater than 0.8 based on cross-validation. Further analysis using a test set derived from manually curated true positives, false positives, and false negatives for each PROSITE pattern showed that the models did not always detect all of the true positives, but they almost always made fewer false positive and false negative predictions than PROSITE.

In a comparison against some of the leading sequence and structure-based function prediction methods, the SeqFEATURE library performed competitively. When the sequence identity and structural similarity of the test set proteins to the training set proteins was reduced, however, the SeqFEATURE library demonstrated a marked robustness that was not matched by any of the other methods. FEATURE's independence from specific sequence and structure elements allows it to perform with greater sensitivity on novel or unique proteins than other methods that rely on conservation.

In principle, SeqFEATURE can be applied to build models for other sequence-oriented motif databases, such as Pfam or PRINTS [36], to generate many more functional site models quickly and automatically, greatly increasing FEATURE's coverage of protein function space. In addition, the enhanced performance at low sequence identity makes FEATURE a particularly relevant method for aiding the annotation of novel protein structures.

Hetero-groups-based functional site models

Many proteins and nucleic acid molecules require small molecular ligands or cofactors such as ATP or NAD in order to function properly. Ligands and cofactors, generally referred to as 'hetero-groups', are diverse. There are currently 7,642 types of hetero-groups in the PDB. These hetero-groups appear in as many as 76.6% of structures in this database. The prevalence of these hetero-groups among biological macromolecules makes them good candidates for automatic training of functional models using FEATURE.

The process of building a hetero-group-based model follows the guidelines described in section "An overview of the FEATURE system". A positive training set for a given hetero-group begins with collection of protein structures containing this hetero-group, namely holo structures. There are many databases of ligand-binding structures, including PDBSum [37], Relibase [38], Hic-Up [39], PLD [40], and PDB-Ligand [41]. The proteins that a given ligand binds are often homologous and present the same binding structure to the ligand. However, there are also many instances wherein a given ligand binds to the same or homologous protein in different binding environments. Therefore, representative structure selection among homologous proteins should be carefully executed. Some of the databases allow automatic superimposition of binding sites and sequence identity filtering which is necessary for representative selection. Once a non-redundant set of holo proteins is composed it may not have a sufficient number of structures. A minimum of five representative structures is required for a positive training set for FEATURE. Since larger datasets are more favorable, apo structures, determined without a hetero group, can supplement the datasets.

Automatic training of hetero-group based models presents us with many challenges. One major challenge is choosing the model center. An obvious strategy is to use the centroid as a center; however, this choice sometimes results in poor performance. Another option is to center on active atoms, but these need to be manually curated for the most part. The larger hetero-groups – containing as many as 390 atoms (e.g. RNA) – present another challenge, as they cannot be fully described within FEATURE's 'traditional' shell size of 7.5 Angstroms. The shell size can be enlarged only to some extent without altering the signal derived from accumulating properties of atoms within shells.

A better strategy is to build several 'sub-models' for different parts of the hetero-group and to combine them into a single model using a range of distances between model centers (see Figure 3). This approach increases the complexity of model building significantly because sub-models can be applied jointly in a combinatory fashion. Preliminary results for ATP-binding site prediction using a two-center approach suggest, however, that performance does improve with the addition of even one more center.

Clustering the PDB to discover and annotate new structural motifs

One limitation of the site selection strategies described in sections "SeqFEATURE – transforming 1D motifs into 3D models" and "Hetero-groups-based functional site models") is that they depend on existing annotation and cannot be used to discover new functions or potentially interesting structural motifs. To overcome this, we calculated FEATURE vectors for all residues in a non-redundant subset of the PDB – over 2 million vectors in all – and clustered them to reveal groups of residues sharing similar microenvironments [42]. In order to make calculation on this scale feasible, features were converted to binary values with minimal reduction in clustering accuracy. A number of the clusters corroborate with known PROSITE motifs, indicating that this strategy can reveal truly interesting groups of sites that may be used to construct new FEATURE models.

Although the capability to discover new motifs is important, its value is diminished unless there is a description of the possible biological or functional roles a new motif may have. One way to alleviate this problem is by generating descriptive text for each cluster automatically. Methods that address similar problems [43-45] rely, for the most part, on standard vocabularies such as the Gene Ontology [46], which are organized at a higher level of conceptual granularity than raw text. While the use of controlled terminologies can resolve many of the challenges surrounding text mining, processing the raw text itself in a way that expressly focuses on the distribution of words across documents and sets of documents may reveal less obvious connections. Such an approach could prove useful not only for characterizing clusters of similar protein microenvironments, but also for clusters or lists of any biological entities that have an associated literature, such as genes, drugs, or diseases.

Preliminary studies on test clusters of proteins derived from PROSITE motifs using a simple entropy-based scoring function demonstrate that this approach is able to detect the fundamental molecular function shared by the members of the cluster (i.e. the PROSITE motif) in addition to more detailed information, such as active site residues (see Table 2).

Improving FEATURE's performance

Extended functional coverage improves the FEATURE framework with respect to the functional space that can be explored. Additionally, it is possible to improve the ability of FEATURE to recognize functional sites, for example, by

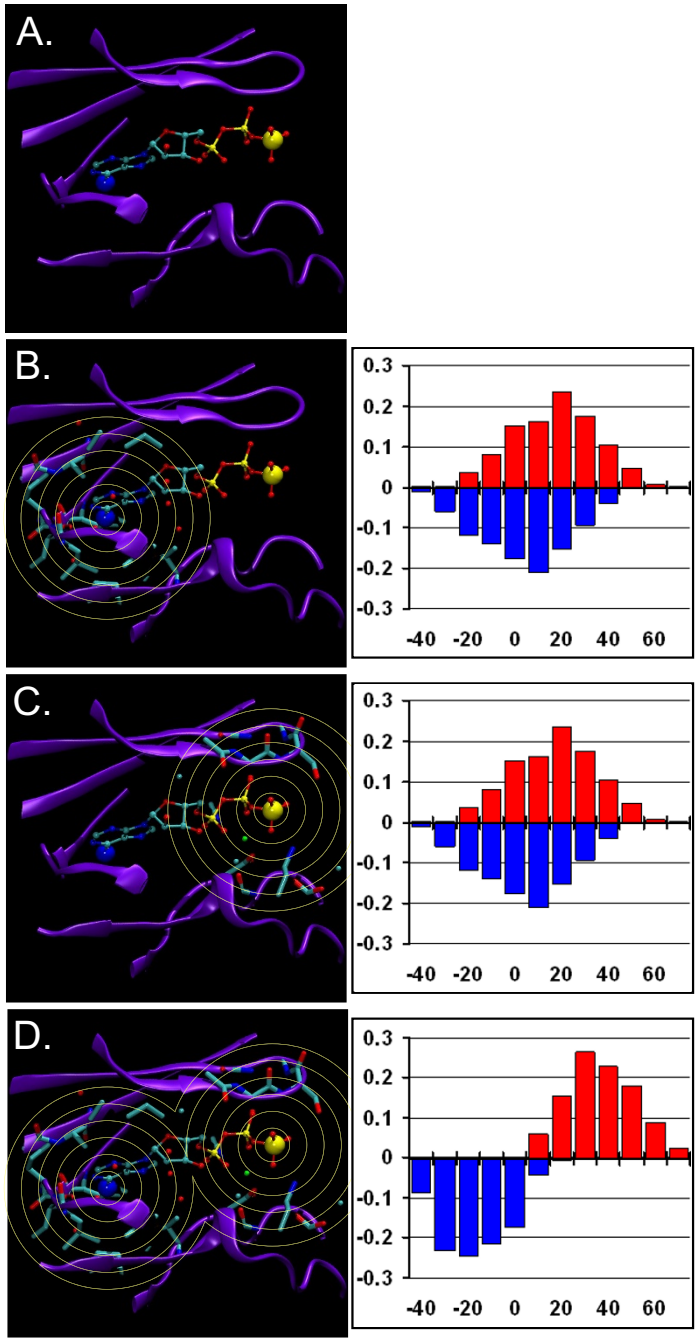

Figure 3
Illustration of the potential value of combining FEATURE models. **A.** An ATP binding pocket in PDB structure 1CSN. Enlarged are N6 (blue) and PG (yellow) atoms in ATP. **B.** Parts of the molecule considered by a putative FEATURE model based at N6 with shells out to 7.5 Å. Such a model might have poor ability to separate positive sites and negative sites, as shown in the histogram on the right with substantial overlap of (red) positive sites and (blue) negative sites. **C.** Parts of the molecule considered by a putative FEATURE model based at PG with shells out to 7.5 Å. Again, such a model might have poor discriminating ability, as shown in the score distributions on the right for (red) positive sites and (blue) negative sites. **D.** Parts of the molecule considered by an analysis which combines the two marginal models in **B** and **C**. With such an analysis it is possible to look for hits at both the N6 and PG and filter by appropriate distance separation between them, and thus achieve combined performance that is much better, as shown in the putative plot on the right.

Table 2: A text mining approach using an entropy-based scoring function rediscovers the molecular function of proteins sharing PROSITE motifs

Motif # of proteins # of documents	Terms
EF_HAND	ef-hand
36	calcium-bind
183	calcium
	ca 2+
	calcium-bind protein
	ca
	2+ bind
	2+
	ef-hand motif
	calmodulin
TRYSIN_SER	serin proteinas
11	proteinas
108	chymotrypsin
	serin
	serin proteas
	elastase
	ser-195
	his-57
	proteinas especially
	proteolyt
PROTEIN KINASE_ST	protein kinas
15	catalyt domain
107	phosphoryl
	substrat
	autophosphoryl
	phosphoryl site
	kinas
	threonin
	catalyt
	constitutively active

The method extracts text from the abstracts of references annotated in each protein's Swiss-Prot record, pre-processes the text (tokenization into terms, removal of non-content words, and basic stemming to normalize word forms), and scores terms based on their distribution across proteins and their relative significance in the entire corpus of Swiss-Prot referenced documents. With no additional normalization, concept and word redundancy may be observed. Although still very preliminary, the method is able to capture the molecular function for each cluster of proteins shown: "ef-hand" and "calcium binding" for EF_HAND; "serine proteinase", "proteolysis", and the active site residues "ser-195" and "his-57" for TRYPSIN_SER; and "protein kinase", "phosphorylation", "catalytic domain" and the substrate residue "threonine" for PROTEIN_KINASE_ST.

exploring the conformational space of the molecules in question. In order to perform their function, most proteins undergo dynamic changes within the active site. Methods that use static structures to predict function do not take structural dynamics into account. However, as the number of solved static structures increases in the PDB and the performance of static methods does not reach desirable levels, the importance of sampling the conformational space of the molecules becomes more apparent.

Dynamics improves efficiency of function annotation methods based on structure

The methods we have reviewed above generally rely on analysis of static structures solved by X-ray crystallography and Nuclear Magnetic Resonance (NMR) techniques. Both techniques, however, have characteristics that may preclude structure-based function prediction methods from performing at the highest levels of sensitivity. In X-ray crystallography, crystal packing may effectively rigidify proteins into compact conformations, which may not represent good averages of the conformational space of the molecules in solution. In order to overcome this limitation, time-resolved X-ray crystallography allows determination of many conformations at 1 picosecond intervals. Using this technique, Schotte *et al.* observed nuances of the inner workings of a myoglobin mutant as it progressed from a carboxy to a deoxy state [47]. However, time-resolved X-ray crystallography is not currently amenable to application in a high-throughput manner, since it requires molecules to be photosensitive, and data interpretation can be nontrivial [48]. These experiments illustrate that it is necessary to take into account the dynamic nature of molecules in order to understand its functional space better.

Although NMR structures do not generally achieve the resolution of structures solved by X-ray crystallography, they better represent the conformational space of the molecules because they typically produce an ensemble of structures. Since the molecules are all in solution during the NMR procedure, this ensemble of structures provides an opportunity to understand the dynamic behavior of molecules. Recent studies highlight the value of the structural diversity contained in the NMR ensembles. We examined several such ensembles (see Table 3) with a FEATURE Ca^{2+} binding model [19]. A subset of structures from most ensembles revealed Ca^{2+} binding sites (see Figure 4). The fact that all the structures did not exhibit Ca^{2+} binding behavior is noteworthy, because it demonstrates that the dynamics may influence our ability to recognize function.

Computational methods allow us to explore the dynamics of molecules on the scales that are not experimentally accessible while assessing their potential functions [49]. In particular, molecular dynamics (MD) simulations provide large ensembles of structures. Recent work demonstrates that MD simulations generate structural diversity useful for the assignment of function. Eyrisch *et al.* used MD simulations to improve efficiency of predicting functional surface pockets, which may be obscured in static PDB structures [50]. In the pharmaceutical industry, improvement in the prediction of functional pockets may assist in the development of more efficient drugs. Frembgen-Kesner *et al.* showed that cryptic drug binding sites, which appear only when the target has bound a ligand

Table 3: Results of NMR ensembles scanned with FEATURE Ca^{2+} binding site model

Protein Name	PDB ID	Number of Models	Number of Models Characterized as Calcium Binding
Lipoprotein receptor-related protein repeat 8	1CR8	20	20
Lipoprotein receptor-related protein repeat 3	1D2L	20	20
RALBP1-intercating protein	1IQ3	18	18
Rous Sarcoma virus receptor	1JRF	20	20
Tyrosine-protein kinase SRC	1KSW	20	20
Calerythrin	1NYA	20	20
Human Notch1	1PB5	16	16
Porcine pancreas phospholipase A2	1SFW	18	1
Rational design of a calcium-binding adhesion protein	1T6W	20	3
Human beta parvalbumin	1TTX	20	20
Cytochrome c peroxidase *	2B10	10	4
Matrilysin	2DDY	25	24
Calcium-binding protein p22	2E+30	20	17
Sodium/calcium exchanger 1 domain 1	2FWS	20	18
Sodium/calcium exchanger 1 domain 2	2FWU	20	18
Rat megalin	2IIP	20	19
Relaxin receptor 1	2JM4	24	22
Yeast frequenin	2JU0	15	15

Scanning of the 18 NMR ensembles with the Ca^{2+} binding model revealed structural heterogeneity among the structures in the ensembles. In several, most of the models exhibited Ca^{2+} binding conformations, while in others, only a few. The first and the second columns contain names and PDB IDs of the examined proteins, respectively. The third and fourth columns show the total number of models and how many of those were identified by FEATURE as Ca^{2+} binding in the NMR ensemble, respectively. * – Results of this scan are shown in Figure 4.

already, become more apparent over the course of MD simulations than in the original PDB structures [51].

We demonstrated the value of examining structural diversity generated by MD simulations with FEATURE to identify Ca^{2+} binding sites [52]. In the case of parvalbumin β, results of FEATURE coupled with dynamics recapitulated the behavior of the protein's Ca^{2+} binding sites with and without synthetic mutations (PDB IDs 1B8C and 1B9A, respectively). Further experiments are underway to establish the extent to which sampling conformational diversity with MD simulations improves efficiency of functional predictions made by FEATURE. Functions other than Ca^{2+} binding may be explored with various FEATURE models or alternative structure-based function prediction methods by evaluating MD ensembles of structures.

Loop modeling

Although SG initiatives are accelerating biological structure determination, it still lags behind the production of new genomic sequences. Roughly a third of all protein sequences can be modeled based on similarity to a known three-dimensional structure, but one of the major limiting factors is the ability to model structurally variable loop regions [53]. Loops participate in many active and binding sites in proteins. *A priori* knowledge of a loop's function can potentially be used to limit its conformational space, thereby assisting in achieving a more accurate ensemble. Such knowledge can result from sequence-based or structure-based predictions or from experiments.

In order to explore FEATURE's utility in loop modeling, loop conformations were generated by two methods: seed sampling and deformation sampling [54]. Both methods satisfy constraints on kinematic closure and clash avoidance. Seed sampling generates structurally diverse loops, whereas deformation sampling explores a more limited region close to the provided starting conformation. We examined the ability of these methods to generate 'functional' loops conformations that are similar to the native structure and could be recognized by FEATURE. Calcium binding loops of parvalbumin (1B8C, Ala51-Ile58) and grancalcin (1K94, Ala62-Asp69) were modeled with seed sampling and deformation sampling respectively. Both routines were able to build at least one functional loop, as evaluated by FEATURE, within a ~100,000 conformation ensemble.

Increasing the accuracy of loop conformation prediction using FEATURE as a filter for functionally plausible conformations can be applied not only to homology modeling but also to the task of modeling missing loops in experimentally-derived structures. Since loops tend to participate in ligand binding, dimer formation and enzymatic activity, they are an essential part of the structure and may hold clues to the elusive structure-function relationship. We are currently validating this method on a

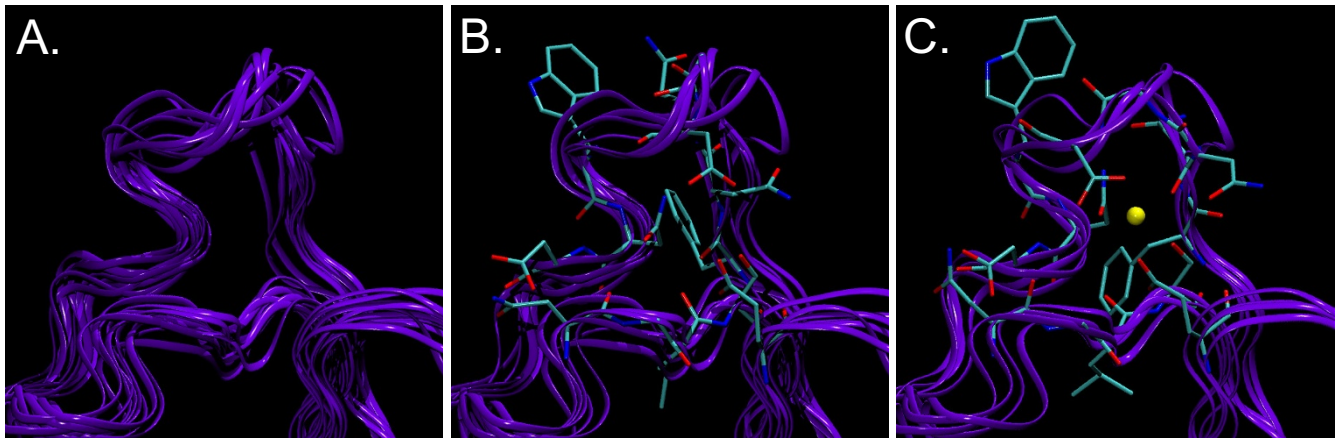

Figure 4
NMR ensemble scanning results for PDB structure 2B1O. 2B1O is a structure of a protein which is known to bind calcium (Ca^{2+}). The NMR ensemble for 2B1O contains different conformations of the structure, some of which show different proclivities for binding Ca^{2+}. **A** shows 10 NMR generated structures for one of the known Ca^{2+} binding loops, superimposed to minimize RMSD; **B** shows loops that FEATURE does not identify as Ca^{2+} binding, corresponding to NMR models 1, 3, 4, 5, 6, and 10; and **C** shows loops that FEATURE does identify as Ca^{2+} binding, corresponding to NMR models 2, 7, 8, and 9. In **B** and **C**, sidechains in the vicinity of the FEATURE hits are shown for the highest scoring NMR model (score ~39 for **B** and ~64 for **C**). In **C**, one of the hits that scored over the model threshold of 50 is shown as a yellow ball. Notice the differences in the conformations between side chains in **B** and **C**: the entire loop is wider in **C**, and coordinating oxygens form a ring around the hit, while in **B** they are more scattered. There is also a difference in the conformation of phenylalanine ring, which essentially blocks the Ca^{2+} binding spot in **B** but is rotated away from the site to allow possible Ca^{2+} binding in **C**.

dataset of existing loops in order to predict missing functional loops reliably.

Extending FEATURE to new applications

The flexibility of the FEATURE framework has proven to be extremely useful for increasing FEATURE's functional coverage and improving not only individual FEATURE models, but also the performance of methods solving slightly different problems, such as loop modeling. Here, we describe some novel applications of FEATURE that have broadened its utility.

Structural genomics and scanning for function in high-throughput

High-throughput projects such as structural genomics are producing greater numbers of uncharacterized and novel proteins than ever before. Often, these proteins bear little resemblance to known proteins in either sequence or structure, making annotation especially challenging. Previously, we showed that the sensitivity of the SeqFEATURE library of automatically derived functional site models (described in section "Increasing functional coverage") is more robust than that of some of the leading sequence and structure-based function prediction methods when sequence identity and structural similarity to known proteins are low [35]. As a result, the SeqFEATURE models should be valuable for suggesting potential functions for novel SG targets.

With this in mind, we scanned all of the SG targets in TargetDB [55] associated with unknown function through October 2007 using the SeqFEATURE library, filtered for the highest confidence predictions (based on model-dependent score cutoffs), and compared them to predictions made by a number of popular sequence and structure-based methods [35]. For a substantial fraction of these targets, the sequence-based methods made no significant predictions; for a smaller fraction, the structure-based methods had no or low confidence predictions as well. Those targets for which SeqFEATURE made a high confidence prediction but other methods did not are compelling candidates for further study (see section "Availability").

In keeping with the need for high throughput, we have also scanned the entire PDB (up to February 2006, about 35,000 proteins) with the entire SeqFEATURE library (see section "Availability"). The scan took about one day to complete on 13 parallel processors, suggesting that a

large-scale scan of many structures with many functional site models is actually quite efficient. With the structure determination pipeline improving and novel protein structures increasing every year, scanning for function in a high-throughput fashion will become a necessary enterprise.

Decoy filtering

One of the major goals of three-dimensional (3D) structure prediction methods, such as comparative modeling, threading and *ab initio* folding, is to elucidate function from a 3D structure. Determining the occurrence and location of active and binding sites within a structure helps achieve this goal. In 1999, Wei *et al.* predicted two calcium-binding sites in model structures, or decoys, of a vitamin D-dependent protein [56]. These decoys, generated by Park and coworkers [57,58], include near native structures. Root mean squared deviation (RMSD), which measures pairwise structural similarity, ranged from 0.95 Å to 9.39 Å between the decoys and the native structure.

Despite the existence of near native decoys, the quality of the calcium-binding microenvironments had only a very weak correlation with the overall RMSD. Moreover, the correlation between 'local RMSD' and FEATURE scores was also weak. Only when the quality of the local structural neighborhood around the calcium site is high does the modeling of the binding sites become reliable. Perturbation of atoms' positions within the native structure generated 100 decoys with a local RMSD of 0 to 1.7 Å [56]. The RMSD of these structures correlated with FEATURE's ability to recognize the functional site.

Recently we re-examined decoy selection with FEATURE. Current improved methodologies for *ab initio* folding are able to generate decoys similar in quality to the previously used perturbed structures. Some small proteins (under 100 amino acids) can be refined up to a near-atomic resolution level [59]. Using FEATURE, we scanned five hundred low scoring decoys for twelve calcium-binding proteins generated with Rosetta [60]. FEATURE scores were able to reduce the number of decoys while enriching for near-native conformations, sometimes with improvements of the average RMSD to known crystal structure moving from 9 to 5 Angstroms (Das Rhiju and Halperin Inbal, unpublished results).

These preliminary results support the potential value of incorporating FEATURE into the *ab initio* folding scheme. Much of the calculation time in *ab initio* folding is spent on the side chain packing of the different main chain conformations generated in the main chain optimization stage. The ability to reduce the number of main chain conformations after this stage while keeping most of the correct conformations would be highly valuable for lowering the computational cost.

Availability

FEATURE models, data, and source code are available online for public use. The WebFEATURE website [61,62] allows functional scans of PDB structures using any of the manually curated models or the models in the SeqFEATURE library, as well as the option to scan using the entire SeqFEATURE library. The improved zinc binding model is also available for scanning [63]. Single SeqFEATURE model scans require only a few seconds to run, scanning with the entire SeqFEATURE library may take about a minute, and manually curated models may take varying lengths of time depending on the size of the input structure. Job status notification can occur either interactively on the website or through email notifications and results can be interactively viewed in a web browser.

Data from the PDB scan and high-confidence predictions for TargetDB structures can be downloaded from the "Data" section of the WebFEATURE site [64]. Source code for FEATURE is accessible from SimTK [65], a repository for biological structure software maintained by the SimBIOS Center for Biomedical Computation [66,67]. FEATURE has been downloaded about 150 times since being made available on SimTK. In addition, WebFEATURE is currently seeing almost 2,500 unique visitors a month.

Conclusion

FEATURE is a powerful function recognition framework that has been adapted to new paradigms in function annotation and structure modeling. Importantly for the annotation of structural genomics targets, FEATURE robustly models molecular functions without relying on significant sequence or fold similarity. Creating training sets automatically from many different sources and discovering new functions through unsupervised clustering of microenvironments improves functional coverage. Function annotation approaches that recognize and treat dynamic nature of molecules as essential are proving to be more successful than their static counterparts, and FEATURE can be easily coupled to simulations to enhance function recognition. Structure determination and loop modeling efforts also benefit from the addition of FEATURE as a filter. As structural genomics and structure determination efforts advance and evolve, structure-based modeling will become more important. FEATURE is uniquely poised to take advantage of and assist in these efforts.

Abbreviations

SG = Structural Genomics. PDB = Protein Data Bank. AUC = Area Under the Curve. NMR = Nuclear Magnetic Reso-

nance. PSSM = Position Specific Scoring Matrix. RMSD = Root Mean Squared Deviation

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

IH, DSG, and SW wrote the manuscript and carried out research described therein. RBA conceived and edited the manuscript.

Acknowledgements

FEATURE is supported by NIH grant LM05652 (method development) and GM072970 to RBA. SW has been supported by LM07033. DSG has been supported by Stanford Genome Training Grant NIH 5 T32 HG00044. We thank Jessica S. Ebert for contributions and comments on the manuscript.

This article has been published as part of *BMC Genomics* Volume 9 Supplement 2, 2008: IEEE 7th International Conference on Bioinformatics and Bioengineering at Harvard Medical School. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/9?issue=S2

References

1. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
2. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A: **Pfam: clans, web tools, and services.** *Nucleic Acids Res* 2006, **34(Database issue):**D247-D251.
3. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ: **The PROSITE database.** *Nucleic Acids Res* 2006, **34(Database issue):**D227-D230.
4. Marsden RL, Lewis TA, Orengo CA: **Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint.** *BMC Bioinformatics* 2007, **8(86):**.
5. Chandonia J-M, Brenner SE: **The impact of structural genomics: expectations and outcomes.** *Science* 2006, **311:**347-351.
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.
7. Novotny M, Madsen D, Kleywegt GJ: **Evaluation of protein fold comparison servers.** *Proteins* 2004, **54:**260-270.
8. Suzuki A, Ando T, Yamato I, Miyazaki S: **FCANAL: structure based protein function prediction method. Application to enzymes and binding proteins.** *Chem-Bio Informatics Journal* 2002, **2(1):**60-72.
9. Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM: **Towards fully automated structure-based function prediction in structural genomics: a case study.** *J Mol Biol* 2007:1511-1522.
10. Laskowski RA, Watson JD, Thornton JM: **ProFunc: a server for predicting protein function from 3D structure.** *Nucleic Acids Res* 2005, **33(Web Server issue):**W89-W93.
11. Wilson C, Kreychman J, Gerstein M: **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.** *J Mol Biol* 2000, **297:**233.
12. Chothia C, Lesk A: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5(4):**823.
13. Binkowski TA, Naghibzadeg S, Liang J: **CASTp: computed atlas of surface topography of proteins.** *Nucleic Acids Res* 2003, **31:**3352-3355.
14. Watson JD, Laskowski RA, Thornton JM: **Predicting protein function from sequence and structural data.** *Current Opinion In Structural Biology* 2005, **15:**275-284.
15. Fetrow J, Skolnick J: **Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases.** *J Mol Biol* 1998, **281(5):**949-968.
16. Wallace AC, Borkakoti N, Thornton JM: **TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites.** *Protein Sci* 1997, **6:**2308-2323.
17. Pal D, Eisenberg D: **Inference of protein function from protein structure.** *Structure* 2005, **13(1):**121-130.
18. Wei L, Altman RB: **Recognizing protein binding sites using statistical descriptions of their 3D environments.** *Pac Symp Biocomp* 1998:497-508.
19. Wei L, Altman RB: **Recognizing complex, asymmetric functional sites in protein structures using a Bayesian scoring function.** *J Bioinform Comput Biol* 2003, **1(1):**119-138.
20. Liang MP, Brutlag DL, Altman RB: **Automated construction of structural motifs for predicting functional sites on protein structures.** *Pac Symp Biocomp* 2003, **8:**204-215.
21. Kufareva I, Budagyan L, Raush E, Totrov M, Abagyan R: **PIER: protein interface recognition for structural proteomics.** *Proteins* 2007, **67(2):**400-417.
22. Pettit FK, Bare E, Tsai A, Bowie JU: **HotPatch: a statistical approach to finding biologically relevant features on protein surfaces.** *J Mol Biol* 2007, **369:**863-879.
23. Youn E, Peters B, Radivojac P, Mooney SD: **Evaluation of features for catalytic residue prediction in novel folds.** *Protein Sci* 2007, **16:**216-226.
24. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N: **ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures.** *Nucleic Acids Res* 2005:W299-302.
25. Jambon M, Imberty A, Deleage G, Geourjon C: **A new bioinformatic approach to detect common 3D sites in protein structures.** *Proteins* 2003, **52:**137-145.
26. Bagley SC, Altman RB: **Conserved features in the active site of nonhomologous serine proteases.** *Fold Des* 1996, **1(5):**371-379.
27. Bagley SC, Wei L, Cheng C, Altman R: **Characterizing oriented protein structural sites using biochemical properties.** *Proc Int Conf Intell Syst Mol Biol* 1995:12-20.
28. Domingos P, Pazzani M: **On the optimality of the simply Bayesian classifier under zero-one loss.** *J Mach Learn Res* 1997, **29:**103-137.
29. Ebert JC, Altman RB: **Robust recognition of zinc binding sites in proteins.** *Protein Sci* 2008, **17(1):**54-65.
30. Banatao DR, Altman RB, Klein TE: **Microenvironment analysis and identification of magnesium binding sites in RNA.** *Nucleic Acids Res* 2003, **31(15):**4450-4460.
31. Levitt M: **Growth of novel protein structural data.** *Proc Natl Acad Sci USA* 2007, **104(9):**3183-3188.
32. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function.** *Genome Res* 2003, **13:**2129-2141.
33. Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J: **The SUPERFAMILY database in 2004: additions and improvements.** *Nucleic Acids Res* 2004, **32(Database issue):**D235-D239.
34. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier.** *Nucleic Acids Res* 2005, **33(Web Server issue):**W116-W120.
35. Wu S, Liang MP, Altman RB: **The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation.** *Genome Biol* 2008, **9(1):**R8.
36. Attwood T, Blythe M, Flower D, Gaulton A, Mabey J, Maudling N, McGregor L, Mitchell A, Moulton G, Paine K, Scordis P: **PRINTS and PRINTS-S shed light on protein ancestry.** *Nucleic Acids Res* 2002, **30(1):**239-241.
37. Laskowski RA: **PDBsum: summaries and analysis of PDB structures.** *Nucleic Acids Res* 2001, **29:**221-222.
38. Hendlich M, Bergner A, Gunther J, Klebe G: **Relibase – design and development of a database for comprehensive analysis of protein-ligand interactions.** *J Mol Biol* 2003, **326:**607-620.
39. Kleywegt G, Jones T: **Databases in protein crystallography.** *Acta Crystallogr D Biol Crystallogr* 1998, **54:**1119-1131.

40. Puvanendrampillai D, Mitchell J: **Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein-ligand complexes.** *Bioinformatics* 2003, **19:**1856-1857.
41. Jae-Min S, Doo-Ho C: **PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures.** *Nucleic Acids Res* 2005, **33:**D238-D241.
42. Yoon S, Ebert JC, Chung EY, De Micheli G, Altman RB: **Clustering protein environments for function prediction: finding PROSITE motifs in 3D.** *BMC Bioinformatics* 2007, **8(Suppl 4):**S10.
43. Daraselia N, Yuryev A, Egorov S, Mazo I, Ispolatov I: **Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks.** *BMC Bioinformatics* 2007, **8(243):**.
44. Raychaudhuri S, Chang JT, Sutphin PD, Altman RB: **Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature.** *Genome Res* 2002, **12:**203-214.
45. Zheng B, McLean DC, Lu X: **Identifying biological concepts from a protein-related corpus with a probabilistic topic model.** *BMC Bioinformatics* 2006, **7:**58.
46. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G: **Gene ontology: tool for the unification of biology.** *Nat Genet* 2000, **25:**25-29.
47. Schotte F, Lim C, Jackson TA, Smirnov AV, Soman J, Olson JS, Phillips GNJ, Wulff M, Anfinrud PA: **Watching a protein as it cunctions with 150-ps time-resolved X-ray crystallography.** *Science* 2003, **300:**1944-1947.
48. Bourgeois D, Schotte F, Brunori M, Vallone B: **Time-resolved methods in biophysics. 6. Time-resolved Laue crystallography as a tool to investigate photo-activated protein dynamics.** *Photochem Photobiol Sci* 2007, **6:**1047-1056.
49. Henzler-Wildman K, Kern D: **Dynamic personalities of proteins.** *Nature* 2007, **450:**964-972.
50. Eyrisch S, Helms V: **Transient pockets on protein surfaces involved in protein – protein interaction.** *J Med Chem* 2007, **50:**3457-3464.
51. Frembgen-Kesner T, Elcock AH: **Computational sampling of a cryptic drug binding site in a protein receptor: explicit solvent molecular dynamics and inhibitor docking to p38 MAP kinase.** *J Mol Biol* 2006, **359:**202-214.
52. Glazer DS, Radmer RJ, Altman RB: **Combining molecular dynamics and machine learning to improve protein function prediction.** *Pac Symp Biocomput* 2008:332-343.
53. Olson M, Feig M, Brooks Cr: **Prediction of protein loop conformations using multiscale modeling methods with physical energy scoring functions.** *J Comput Chem* 2007, **29(5):**820-831.
54. Yao P, Dhanik A, Marz N, Propper R, Kou C, Liu G, Bedemy H van den, Latombe J, Halperin I, Altman R: **Efficient algorithms to explore conformational spaces of flexible protein loops.** *IEEE/ACM Trans Comput Biol Bioinform* in press.
55. Chen L, Oughtred R, Berman HM, Westbrook J: **TargetDB: a target registration database for structural genomics projects.** *Bioinformatics* 2004, **20(16):**2860-2862.
56. Wei L, Huang E, Altman RB: **Are predicted structures good enough to preserve functional sites?** *Structure* 1999, **7:**643-650.
57. Park B, Huang E, Levitt M: **Factors affecting the ability of energy functions to discriminate correct from incorrect folds.** *J Mol Biol* 1997, **266:**831-846.
58. Park B, Levitt M: **Energy functions that discriminate X-ray and near native folds from well-constructed decoys.** *J Mol Biol* 1996, **258:**367-392.
59. Bradley P, Malmström L, Qian B, Schonbrun J, Chivian D, Kim D, Meiler J, Misura K, Baker D: **Free modeling with Rosetta in CASP6.** *Proteins* 2005, **61(Suppl 7):**128-134.
60. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka M, Bhat D, Chivian D, Kim D, Sheffler W, Malmström L, Wollacott A, Wang C, Andre I, Baker D: **Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home.** *Proteins* 2007, **69(Suppl 8):**118-128.
61. Liang MP, Banatao DR, Klein TE, Brutlag DL, Altman RB: **WebFEATURE: an interactive web tool for identifying and visualizing functional sites on macromolecular structures.** *Nucleic Acids Res* 2003, **31(13):**3324-3328.
62. **WebFEATURE** [http://feature.stanford.edu/webfeature]
63. **FEATURE metal scanning data** [http://feature.stanford.edu/metals]
64. **WebFEATURE data** [http://feature.stanford.edu/webfeature/data]
65. **SimTK** [http://simtk.org/]
66. **SIMBIOS** [http://simbios.stanford.edu/]
67. **SIMBIOS.** *NIH GM072970* .