Automated Construction of Structural Motifs for Predicting Functional Sites on Protein Structures

M.P. Liang, D.L. Brutlag, R.B. Altman

Pacific Symposium on Biocomputing 8:204-215(2003)

AUTOMATED CONSTRUCTION OF STRUCTURAL MOTIFS FOR PREDICTING FUNCTIONAL SITES ON PROTEIN STRUCTURES

M.P. LIANG, D.L. BRUTLAG, R.B. ALTMAN
Stanford University, 251 Campus Drive X-215,
Stanford, CA 94305-5479, USA
E-mail: {mliang, brutlag, altman}@smi.stanford.edu

Structural genomics initiatives are beginning to rapidly generate vast numbers of protein structures. For many of the structures, functions are not yet determined and high-throughput methods for determining function are necessary. Although there has been extensive work in function prediction at the sequence level, predicting function at the structure level may provide better sensitivity and predictive value. We describe a method to predict functional sites by automatically creating three dimensional structural motifs from amino acid sequence motifs. These structural motifs perform comparably well with manually generated structural motifs and perform better than sequence motifs. Automatically generated structural motifs can be used for structural-genomic scale function prediction on protein structures.

1 Introduction

1.1 Structural Genomics

With the sequencing of the human genome and many other organisms, rapid determination of gene and protein function is becoming increasingly important. Structural genomics initiatives are helping elucidate the function of these gene products by developing high-throughput methods for determining structures for all unique protein folds. These new structure targets are specifically selected not to have sequence similarity to existing proteins[1-4]. With the anticipated explosion of available structures, it is imperative to develop computational methods for high-throughput function prediction on protein structures.

1.2 Sequence Motifs

There has been extensive work in identifying conserved residues in protein sequences with similar function. Amino acid sequence patterns that represent these conserved residue positions can be created from multiple alignments of sequences with similar function. These patterns, or sequence motifs, can be used to assign function to sequences that contain the pattern.

Numerous sequence motif databases have been established with different methods for creating sequence motifs. Some of the databases are manually curated by experts, while others are automatically derived. The BLOCKS+ database provides an integration of many sequence motif databases by generating blocks of ungapped alignment of sequences from families in the databases[5, 6].

The eMotif database uses sequence alignments from the BLOCKS+ database to create sequence motifs (eMotifs) at various specificities for a family of protein sequences. By providing motifs at different specificities for a block of sequences, eMotifs can represent families with high specificity and yet provide sensitivity by combining multiple motifs[7, 8].

Although sequence motifs can provide insight into protein function, when new proteins do not have significant sequence similarity with known proteins, using only sequence information fails. Proteins that do not have high sequence similarity may still have similar function because of conservation of physicochemical properties at the structural level[9, 10].

1.3 Structural Motifs

Analogous to sequence motifs, structural motifs provide a description of conserved properties in the three dimensional structure of proteins sharing molecular function. Investigators have devised different techniques to construct and define structural motifs; each technique emphasizes different conserved properties.

Wallace et al have developed a system, PROCAT, for identifying catalytic sites by geometric orientation of residues with known functional importance. By using previous knowledge of the critical residues involved in the catalytic activity, a structural motif representing the conserved relative positions of those residues is constructed. This motif can be used to scan a new protein structure for occurrence of the catalytic site using a geometric hashing algorithm[11, 12].

Fetrow and Skolnick have developed Fuzzy Functional Forms (FFF) for representing distances between interesting residues. The critical residues involved in a functional site are identified by careful examination of the literature. Examples of known structures containing these residues are used to find mean distance and variance between the residues. The structural motif representing the conserved distance and variance of the residues is used to identify functional sites on protein structures[13, 14].

Wei and Altman have developed the FEATURE system that describes the physicochemical environment around functional sites. The environment, characterized by observing the frequency of physicochemical properties in radial shells around the site of interest, represents a structural motif used for predicting the functional sites[15, 16].

Unlike PROCAT and FFF, FEATURE does not require the conserved properties to be known in advance, but can discover them automatically given a training set. Once a set of examples of the functional site and a background control set is provided, FEATURE can automatically build the structural motif without having to manually identify which residues are considered important.

We introduce a new method, SeqFEATURE, that automatically creates structural motifs around functional sites by characterizing the structural environment around sequence motifs using FEATURE. Since the three dimensional structures of protein are more conserved than the sequence of amino acids[10], focusing on the protein structure should provide better sensitivity in identifying protein function.

2 Methods

2.1 SeqFEATURE overview

SeqFEATURE is a method for automatically building a structural motif from sequence motifs (Figure 1). The sequence motifs can be obtained from any of the sequence motif databases; here we use the eMotif database. Protein structures that contain the sequence motif are identified and a training set for the FEATURE system is automatically generated. FEATURE then uses the training set to construct a structural motif describing the physicochemical environment around the sequence motif.

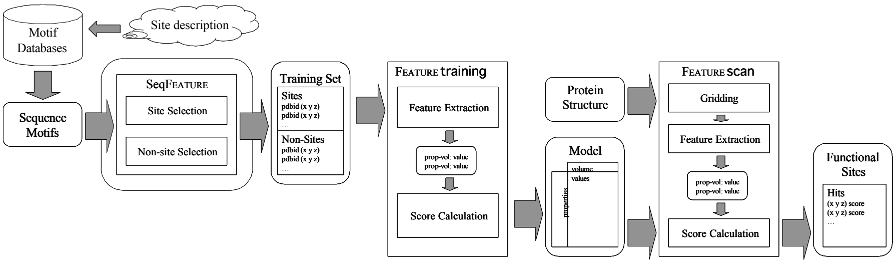

Figure 1: Data flow diagram for SeqFEATURE. Sequence motifs are selected by querying a sequence motif database. These motifs are fed into SeqFEATURE for automatic selection of sites and non-sites to create a training set. The training set is used by FEATURE for creating a model of the microenvironment around the functional site. The model is used to predict functional sites on new protein structures.

2.2 FEATURE system

The FEATURE system is detailed and evaluated elsewhere[16]; here we describe it briefly. The FEATURE system is used to build structural motifs and predict functional sites. FEATURE builds a physicochemical model of the environment around functional sites. The model represents the statistical distribution of physicochemical properties at radial distances from the site of interest. The physicochemical properties range from atom names and atom properties to residue names and secondary structure (Table 1). By using properties at the atomic level and not just at the residue level, FEATURE provides a closer view of the chemistry necessary for function.

Table 1: List of FEATURE's physicochemical properties

AtomName	C	N	O	S
	ANY	OTHER		
ChemicalGroup	Hydroxyl	Amide	Amine	Carbonyl
	RingSystem	Peptide		
AtomProperties	VDWVolume	Charge	NegCharge	PosCharge
	ChargeWithHis	Hydrophobicity	Mobility	SolventAccessibility
ResidueName	ALA	ARG	ASN	ASP
	CYS	GLN	GLU	GLY
	HIS	ILE	LEU	LYS
	MET	PHE	PRO	SER
	THR	TRP	TYR	VAL
	HOH	OTHER		
ResidueProperties	Hydrophobic	Charged	Polar	NonPolar
	Basic	Acidic		
SecondaryStructure	3Helix	4Helix	5Helix	Bridge
	Strand	Turn	Bend	Coil
	Het	Unknown		

FEATURE requires a training set composed of positive examples of the functional site and negative background control examples. Each example is a specific 3D coordinate in a protein structure locating the site of interest. The environment around each site of interest is divided into radial shell volumes. FEATURE then creates a feature vector v for each site by counting the occurrence of physicochemical properties within each volume around the site.

The structural motif of the functional site is constructed by comparing the distribution of feature vectors from the positive examples (sites) with those from the negative examples (non-sites). Properties at each volume are evaluated as statistically more present or absent in the functional site by using the Wilcoxon rank sum test.

Finally, using a Bayesian inference method, FEATURE predicts functional sites at specific locations on protein structures. It places a 3D grid over the new protein structure and calculates the feature vector described above from the environment around each grid point. The grid point is given a score proportional to the likelihood that it is a functional site given its feature vector v.

$$\text{Score} = \log \frac{p(\text{Site} \mid v)}{p(\text{Site})} \tag{1}$$

By assuming that the components of the feature vector are independent, i.e. the property-volume values v_i are independent, the score can be broken down to summing the probability that the grid point is a site given each vector component v_i.

$$\text{Score} = \sum_i \log \frac{p(\text{Site} \mid v_i)}{p(\text{Site})} \tag{2}$$

2.3 eMotif database

Here we use the eMotif database[8] as the source of sequence motifs for automatic creation of structural motifs, although any of the other sequence motif databases would be acceptable. The eMotif database provides a broad collection of sequence motifs, each representing functional signatures or domains built from BLOCKS+.

eMotif takes advantage of predefined conserved residue substitution groups based on shared chemical or physical properties of the amino acids, such as size or hydrophobicity. The resulting motifs represent variation in motif positions that are biologically meaningful.

In addition, multiple eMotifs are built for a block of sequences with different levels of specificity and sensitivity. These properties make eMotif particularly amenable to predicting functional sites at a proteomic level.

2.4 Training set selection

We select sequence motifs related to the functional site from the eMotif database using a keyword search. A training set of positive example sites and negative control non-sites is automatically generated from the sequence motif.

Structures from the Protein Data Bank[17] (PDB) whose sequence contains the sequence motifs are selected. Bennett et al. have created a database of all structures containing eMotifs (3MOTIF) along with accompanying tools for visualizing the eMotifs on the 3D structures[18, 19].

To pick a positive site example, the residues matching the sequence motif are extracted from the structure and the geometric center of their alpha carbons is selected.

To pick a control non-site example, the atom density around the selected positive sites is calculated and a random point on the same protein structure with similar atom density is selected. Points within a particular distance from the positive site are excluded.

2.5 Evaluation metrics

The performance of a structural motif or sequence motif is measured by its sensitivity and positive predictive value.

Sensitivity is the true positive rate, representing how many of the true sites are detected by the motif. It is calculated by taking the ratio of the number of sites that are predicted correctly ($\mathrm{TP}_{\mathrm{sites}}$) to the total number of sites ($\mathrm{TP}_{\mathrm{sites}}$ + $\mathrm{FN}_{\mathrm{sites}}$).

$$\text{Sensitivity} = \frac{\text{Number of Sites Predicted Correctly}}{\text{Total Number of Sites}} = \frac{\mathrm{TP}_{\mathrm{sites}}}{\mathrm{TP}_{\mathrm{sites}} + \mathrm{FN}_{\mathrm{sites}}} \qquad (3)$$

Positive predictive value (PPV) measures how often the positive predictions are correct. Positive predictions (hits) are predictions that a functional site occurs at a particular location. Positive predictive value is calculated by taking the ratio of the number of hits that are near true sites (TP_{hits}) to the total number of hits (TP_{hits} + FP_{hits}).

$$\mathrm{PPV} = \frac{\text{Number of Hits near True Sites}}{\text{Total Number of Hits}} = \frac{TP_{hits}}{TP_{hits} + FP_{hits}} \qquad (4)$$

It is important to point out that the units for sensitivity and PPV are not interchangeable. Sensitivity is a ratio of sites, whereas PPV is a ratio of hits. Because hits detect the area around a single point representing the site, there could be more than one predicted location per actual site.

3 Results

3.1 EF-Hand calcium binding sequence motif

Calcium ions have a spectrum of essential biological roles, affecting a myriad of regulatory processes, enzymatic activity, and protein stability. The EF-Hand is the most common motif for binding calcium in proteins[20, 21].

Sequence motifs for the EF-Hand family of calcium binding sites were selected from eMotif by keyword search. Thirteen motifs were found with specificity ranging from 10^{-3} to 10^{-9} (Table 2). The motifs were 12 to 13 residues in length.

Table 2: EF-Hand family eMotifs at varying specificities

eMotif	*specificity*
d.[dn]........[de]	10^{-3}
[dn].[dn]....[ilmv]...[de][filvy]	10^{-4}
d.......[filmv].e[fwy]	10^{-4}
d.[dn]..g.[ilmv]...[de]	10^{-5}
[dn]...d....[fly].e[fwy]	10^{-5}
d.[dn]..g.[ilmv]...e[fly]	10^{-6}
d.[dn].d..[iv]...[de][fly]	10^{-6}
d.[dn].[dn]....[filmv].e[fwy]	10^{-6}
d.[dn].dg.[ilv]...[de][fly]	10^{-7}
d.[dn]..g.[ilmv].[fly].e[fwy]	10^{-7}
d.[dn].d..[ilv].[filmv].e[fwy]	10^{-7}
d.[dn].dg.[ilv].[filmvy].e[fwy]	10^{-8}
d.[dn].dg.[iv]..[de]ef	10^{-9}

3.2 Automated construction of structural motif

The 13 EF-Hand motifs had a total of 1374 hits on 62 structures. These were automatically selected and fed into the training set generator. The positive and negative examples were selected as described in the methods generating a total of 220 sites and 119 non-sites. Figure 2 shows the automatically constructed structural motif.

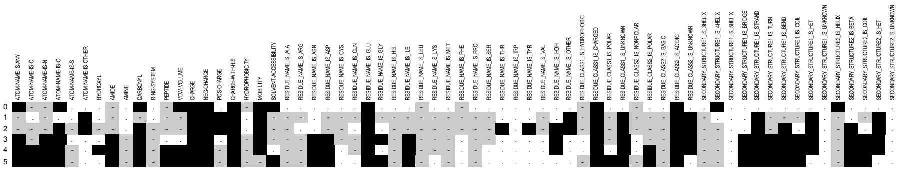

Figure 2: SeqFEATURE model of calcium binding automatically generated from EF-Hand sequence motifs. This model describes the 3D environment around the sequence motif. Each row represents the distribution of properties in increasingly larger spherical shell around the site. Properties significantly present are shaded darker and properties significantly absent are shaded lighter.

3.3 Manual construction of general calcium binding motif

A general calcium binding structural motif was created by manual generation of a training set for FEATURE. Forty protein structures from the PDB with unique folds for binding calcium were manually selected. The calcium ions in the protein structures were used as sites for the training set. In addition, random backbone atoms in 14 structures known not to have calcium binding activity were used as non-sites.

3.4 Performance on general calcium binding proteins

In the following performance figures, the presence of calcium ion in the crystal structure is used as the gold standard for identifying true sites.

The general calcium binding site test set had 54 structures containing 91 true sites. The number of hits from FEATURE is varied by adjusting the cutoff threshold for the score. The eMotif sequence motifs predicted 81 of 98 hits near sites, detecting 14 unique sites (15% sensitivity, 83% PPV). At similar sensitivity, SeqFEATURE with score cutoff at 115 had 23 of 24 hits near sites, detecting 14 unique sites (15% sensitivity, 96% PPV). The manual FEATURE with score cutoff at 65 had 17 of 18 hits near sites, detecting 16 unique sites (18% sensitivity, 94% PPV). Figure 3 shows the performance over varying sensitivity.

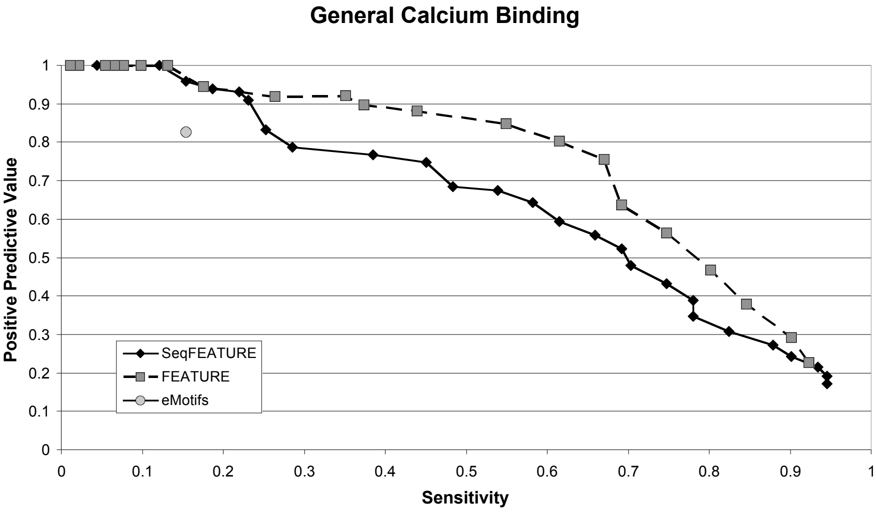

Figure 3: Performance of SeqFEATURE, FEATURE, and eMotifs on general calcium binding sites.

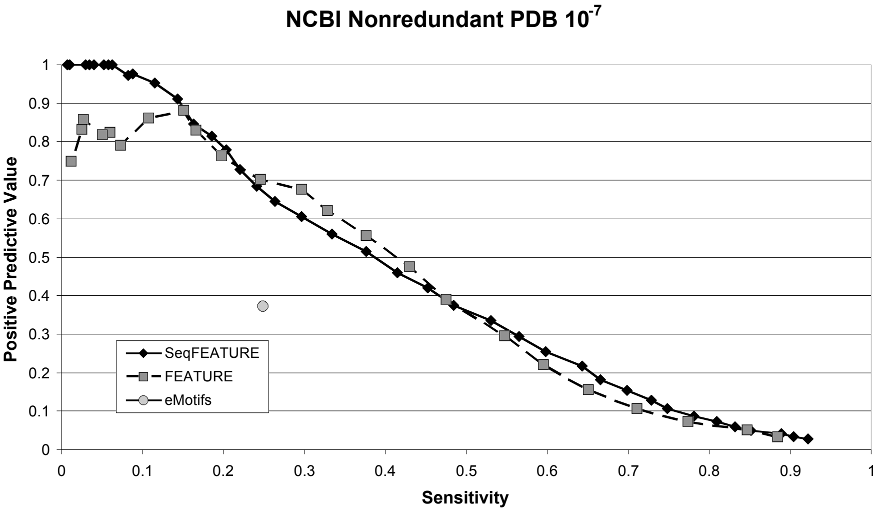

Figure 4: Performance of SeqFEATURE, FEATURE, and eMotifs on NCBI's non-redundant PDB.

3.5 Performance on non redundant database

The non-redundant PDB set from NCBI with Blast e-value of 10e-7 was used as another test set. There were a total of 2122 structures in the non-redundant PDB. 131 structures contained calcium ions, making 398 true sites.

The eMotif sequence motifs predicted 1376 hits of which 514 were near sites, detecting 99 unique sites (25% sensitivity, 37% PPV). At similar sensitivity, SeqFEATURE with score cutoff at 110 had 284 of 415 hits detect 96 unique sites (24% sensitivity, 68% PPV). Manual FEATURE with score cutoff at 60 had 137 of 195 hits detect 98 unique sites (25% sensitivity, 70% PPV). Figure 4 shows the performance over varying sensitivity.

4 Discussion

In both the general calcium binding sites and the non-redundant PDB test sets, the automatically generated EF-Hand structural motif from SeqFEATURE had comparable performance to manual construction of a calcium binding site from FEATURE. The FEATURE calcium binding site performed better in the general calcium binding test set, with higher positive predictive value across the sensitivity range. This is expected since the SeqFEATURE model was created from only examples of EF-Hand calcium binding whereas the manual FEATURE model was trained on a variety of different calcium binding structures.

The SeqFEATURE model of the EF-Hand was able to have better PPV and sensitivity than the EF-Hand sequence eMotifs in both test cases. This demonstrates how observing the three dimensional environment around sequence motifs can provide more information in predicting function than just using the sequence information.

SeqFEATURE and FEATURE use a discriminative method for learning the model for a functional site. By taking positive examples and negative control examples, these methods identify what shared attributes in the positive examples make them different from the control examples and filter out attributes and noise common to both sets. The eMotifs use a generative method for representing the conservation of residues. They only use a set of positive examples, identifying the properties conserved in the set and filtering out those that have too much variance. If the eMotifs are created using discriminative methods, they may provide more power in predicting function than the current generative methods.

In the evaluation of the structural and sequence motifs for calcium binding, we used the presence of calcium ion in the crystal structure as a gold standard. This standard may be subject to errors as some calcium ions may have been missed in the structure determination due to different experimental conditions. By taking a large set of examples, we hope these errors will be averaged out.

5 Conclusion

The three dimensional structure of the proteins provide more information for predicting functional sites than does sequence homology. A structural motif for calcium binding was automatically created from sequence motifs of the EF-Hand calcium binding family. By using structural information, the structural motif had better performance in detecting calcium binding than the sequence motif. This method for automatic construction of structural motifs from sequence motifs can be used to build a library of models for performing genomic-scale prediction of functional sites. Structural motifs for other types of functional sites, such as catalytic sites and small molecule binding sites, will be tested in the future.

Acknowledgments

This work was supported by the National Institute of Health (NIH) grants LM-05652, LM-07033, GM-63495, and HG-02235. We thank Rey Banatao, Liping Wei and Soumya Raychaudhuri for helpful discussions.

References

1. Brenner, S.E., "A tour of structural genomics". *Nat Rev Genet.* **2**, 10 (2001) pp. 801-9.
2. Burley, S.K., et al., "Structural genomics: beyond the human genome project". *Nat Genet.* **23**, 2 (1999) pp. 151-7.
3. Baker, D. and A. Sali, "Protein structure prediction and structural genomics". *Science.* **294**, 5540 (2001) pp. 93-6.
4. Skolnick, J., J.S. Fetrow, and A. Kolinski, "Structural genomics and its importance for gene function analysis". *Nat Biotechnol.* **18**, 3 (2000) pp. 283-7.
5. Henikoff, J.G., et al., "Increased coverage of protein families with the blocks database servers". *Nucleic Acids Res.* **28**, 1 (2000) pp. 228-30.
6. Henikoff, S., J.G. Henikoff, and S. Pietrokovski, "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations". *Bioinformatics.* **15**, 6 (1999) pp. 471-9.
7. Nevill-Manning, C.G., T.D. Wu, and D.L. Brutlag, "Highly specific protein sequence motifs for genome analysis". *Proc Natl Acad Sci U S A.* **95**, 11 (1998) pp. 5865-71.
8. Huang, J.Y. and D.L. Brutlag, "The EMOTIF database". *Nucleic Acids Res.* **29**, 1 (2001) pp. 202-4.
9. Chothia, C. and A.M. Lesk, "The relation between the divergence of sequence and structure in proteins". *Embo J.* **5**, 4 (1986) pp. 823-6.
10. Chothia, C. and A.M. Lesk, "The evolution of protein structures". *Cold Spring Harb Symp Quant Biol.* **52**, (1987) pp. 399-405
11. Wallace, A.C., N. Borkakoti, and J.M. Thornton, "TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites". *Protein Sci.* **6**, 11 (1997) pp. 2308-23.
12. Wallace, A.C., R.A. Laskowski, and J.M. Thornton, "Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases". *Protein Sci.* **5**, 6 (1996) pp. 1001-13.
13. Fetrow, J.S. and J. Skolnick, "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.* **281**, 5 (1998) pp. 949-68.
14. Fetrow, J.S., A. Godzik, and J. Skolnick, "Functional analysis of the Escherichia coli genome using the sequence- to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity". *J Mol Biol.* **282**, 4 (1998) pp. 703-11.

15. Bagley, S.C., et al., "Characterizing oriented protein structural sites using biochemical properties". *Proc Int Conf Intell Syst Mol Biol*. **3**, (1995) pp. 12-20
16. Wei, L. and R.B. Altman, "Recognizing protein binding sites using statistical descriptions of their 3D environments". *Pac Symp Biocomput*., (1998) pp. 497-508.
17. Berman, H.M., et al., "The Protein Data Bank". *Nucleic Acids Res*. **28**, 1 (2000) pp. 235-42.
18. Bennett, S.P., C.G. Nevill-Manning, and D.L. Brutlag, "3MOTIF: Visualizing Conserved Protein Sequence Motifs in the Protein Structure Database". Submitted for Publication (2002)
19. Bennett, S.P. and D.L. Brutlag, "Protein Sequence Motifs in the Protein Structure Database". Personal Communication (2002)
20. Donato, R., "Functional roles of S100 proteins, calcium-binding proteins of the EF- hand type". *Biochim Biophys Acta*. **1450**, 3 (1999) pp. 191-231.
21. Lewit-Bentley, A. and S. Rety, "EF-hand calcium-binding proteins". *Curr Opin Struct Biol*. **10**, 6 (2000) pp. 637-43.