Critical Review of Functional Site Prediction

Biochem218 (Spring 2009)
Grace Tang

Introduction

In the late 1900s, advancements in X-ray crystallography, Nuclear Magnetic Resonance (NMR), gene cloning, and protein expression enabled the creation of structural genomics initiatives worldwide. Centers in Japan, Canada, Europe, and United States collectively aimed to increase the structural knowledge of proteins and to achieve a more through coverage of protein structure space.[1] At that time, many protein families lacked structural representatives in the Protein Data Bank (PDB), a repository for three-dimensional protein structures. In 2005, an assessment of structural knowledge across sequence space showed only thirty-six percent of PFAM families (2736 of 7677) to possess at least one member with a structure. [2,3] Structural genomics target selection is the process of selecting a structure to characterize. It has been focused on proteins with low homology to existing structures to improve the coverage of protein structure space.[4] However, this has come at the cost of increasing the number of structures with unknown function. Over one-third of the structures from the structural genomics initiatives have no additional experimental data to infer functions from and many remain annotated as hypothetical proteins.[5]

Traditionally, sequence similarity is used to infer function for novel proteins. A given protein sequence is queried using BLAST against the Swiss-Prot database to search for experimentally annotated sequences of high significance.[6] This is minimally applicable for proteins from structural genomics since target selection often involves a 30% identity cutoff to protein families with a known structure.[7] Sequences with this low level of identity are in the twilight-zone, where function and structure cannot be reliably inferred from sequence alone.[7] One approach is to search for local one-dimensional sequence motifs or patterns, under the assumption that some homology remains. An alternative and more powerful approach is to search the structure for known three-dimensional motifs. These two classes of motifs can provide clues to function by locating local areas of conversation, potentially active sites and/or binding sites. While this does not fully explain function, it provides a basis upon which future experimental studies can be designed. Understanding the biological roles of these novel structures is fundamental to realizing their potential.

One-Dimensional (1D) Motifs

1D-motifs are local conserved sequences associated to the functional or structural elements of a protein. They commonly represent enzyme catalytic sites, prosthetic group attachment sites, metal binding sites, disulfide bond formation sites, small molecule binding sites, and macromolecule binding sites. Motifs, which are also referred to as patterns, signatures, and fingerprints, are derived from multiple sequence alignments of homologous proteins. As divergent sequences are introduced into an alignment, gaps and insertions are inevitably introduced. Sectors of high conservation begin to emerge, which are reduced down to form the motifs (Figure 1). Experimental evidence has repeatedly shown these regions to be intimately linked to the functionality of proteins, thereby explaining their increased conservation during evolution. The high information content of these segments renders them less sensitive to low pairwise identity compared to other functional inference techniques like pairwise alignment. For these reasons, 1D motif scanning is a popular approach for gaining functional insight to the diverse proteins from structural genomics initiatives. Here, I will discuss three of the most popular 1D-motif databases as well as a meta-server helpful for integrating results from different sources.

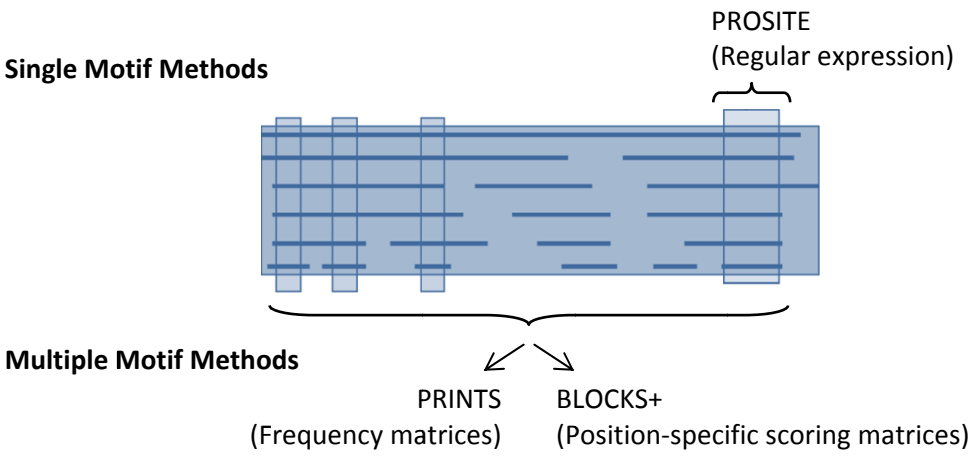


Figure 1. Overview of 1D motif creation: single motif and multiple motif methods. In parentheses are the representation methods for the conserved sequences.

PROSITE[8]

PROSITE is one of the earliest annotated databases of motifs aimed for protein family and domain recognition. Conversed segments from multiple sequence alignments are simplified into a consensus sequence. On average, patterns vary between ten to twenty amino acids in length and are represented as a regular expression. These motifs can come in a variety of forms. Motifs can be strictly conserved with each position of the pattern recognizing a specified amino acid. Others are more flexible and allow or exclude subsets of physiochemically similar residues. Positions where all twenty amino acids are permitted are considered spacers or wildcards. These spacers can vary in length to permit variable gap and linker lengths between conserved residues.

PROSITE patterns are qualitative, meaning sequences either match or do not match. Small deviations from the pattern are fully rejected. Since no acceptance threshold is employed, there is no sense of how well or poor a given sequence matches a pattern. Unlike many other motif finders, there is no significance score for a match. However, the motifs do have a precision and recall score assessed using the Swiss-Prot database.[9] The Swiss-Prot database is manually curated by biologists, providing high quality and reliable information for protein sequences. A scan of Swiss-Prot identifies all the hits for a motif and cross-referencing to the experimental information allows for quick categorization into true positives, true negatives, false positives, and false negatives. In general, PROSITE aims to achieve high precision at the cost of recall. By only including the most conserved amino acids into a motif, the number of false positives is reduced, and the number of false negatives is increased.

PROSITE patterns are easy to understand and quick to use. Pattern recognition is not computationally intensive, so large numbers of protein sequences can be scanned in a reasonable time frame. Patterns are also focused on the most conserved residues within a domain, residues which are often tightly associated with function. These positions are most relevant for further research and of most interest to scientists. However, the specific implementation method of PROSITE has numerous weaknesses. The high false negative rate of the patterns is of concern, especially when working with novel proteins. Proteins from the structural genomics initiatives are known to have low sequence homology to other proteins, and as such, may have slightly divergent motifs. PROSITE patterns being qualitative would preclude identification of these motifs. When suspected patterns are not detected, it is unclear how divergent the sequence was. Without a significance score, it is also hard to assess how likely a given sequence is a false negative. Often times, the most interesting predictions are those bordering on the edge of significance, but with the architecture of PROSITE, these cannot be found. Additionally, PROSITE aims to represent a functional site as a single motif. However, this is not always possible. Linear motifs such as leucine zippers are suitable for single motif representation, but serine protease active sites are not.[10] A three-dimensional active site is often defined by regions distant in

linear sequence that cannot be captured by a single motif. Therefore, PROSITE, while fast and straightforward, is limited in its ability to detect distant homologs and represent complex active sites.

PRINTS[11]

PRINTS is a more recent annotated compendium of protein fingerprints. Rather than express a protein family or domain or functional site as a single motif, a series of generally non-overlapping ungapped motifs are used. This provides a more complete representation of a set of homologous proteins while restricting recognition to only regions of high conservation. These sets of motifs, which are separated in primary sequence, are often contiguous in three-dimensional space, further justifying their usage. Creation of the PRINTS database also began with building multiple sequence alignments of homologous proteins. Rather than using the most conserved sector of an alignment (PROSITE), all well conserved sectors ten to twenty amino acids in length are merged to form a protein fingerprint. PRINTS uses frequency (identity) matrices to capture the information within the multiple sequence alignments. The use of similarity matrices is specifically avoided because they are inherently noisy.

In the PRINTS database, fingerprints with two motifs have similar performance to single motifs. Single motifs are by nature more prone to false positives since high conservation across a single motif is less stringent than high conservation across multiple motifs. Fingerprints composed of more patterns generally have higher specificity. Simultaneously, as the number of motifs in a fingerprint increases, the model becomes increasingly tolerant of mismatches within a motif and within a fingerprint. This enables detection of more distant sequences. Matches are determined on the basis of how many signatures are matched, how well the signatures are matched, and if the ordering and placement of the signatures is correct.

A key benefit of PRINTS over PROSITE is the significance scores for the matches. It gives a sense of how well a profile matches a fingerprint, allowing for better user interpretation. The significance cutoff can also be manually adjusted by the user to get PRINTS to return conservative or non-conservative predictions. Representation of patterns as frequency matrices rather than regular expressions also reduces information loss. For example, a regular expression that accepts both a leucine and valine may favor leucine 90% of the time and valine 10% of the time. This bias is not visible in a regular expression but can be captured by a frequency matrix. Additionally, the design of PRINTS, which represents families of proteins as a set of patterns, is more intuitive. Complex three-dimensional functional sites are often mapped to multiple positions of a protein sequence. Multiple pattern representation is capable of capturing all relevant regions while single pattern representation is not. The presence or lack of other patterns in the fingerprint can provide insight into subfamily, family, and superfamily classification and shed light on functional specificity. The architecture of PRINTS enables it to detect motifs in distant sequences and can offer additional information about functional divergence.

BLOCKS and BLOCKS+[12]

The BLOCKS database represents conserved regions of proteins as blocks of ungapped multiple sequence alignments. Like PRINTS, multiple blocks are used to represent homologous proteins if multiple regions of high conservation are found. Matches are returned with a significance score that assesses how likely the match occurs from chance. The significance cutoff can be freely adjusted by the user to return results of varying significance. This allows for identification of highly similar sequences as well as more divergent sequences. Unlike PRINTS, BLOCKS uses a position-specific scoring matrix (PSSM) to represent motifs. Included in the PSSM are pseudo-counts to account for the fact that the alignments are incomplete.[13] This enables the motifs to tolerate amino acids that are not observed in the multiple sequence alignment. When looking for homology in remote sequences, BLOCKS can have improved performance over PRINTS, which uses the raw frequency count to represent a motif. Likewise, it is also

more flexible than PROSITE's regular expressions. However, the more lenient PSSM can also lead to increased false positives and lower the significance for true positives.

BLOCKS is unique in that it is generated using a fully automated algorithm, rather than by careful manual curation. It is therefore capable of building upon other curated databases to increase its coverage. The original BLOCKS was built upon the protein families in PROSITE. As PROSITE does not have complete coverage of all protein families, BLOCKS was expanded to BLOCKS+. BLOCKS+ contains not only the families represented in PROSITE, but also those from PRINTS, PFAM-A, ProDom,[14] and Domo.[15] By only building upon annotated databases, BLOCKS+ retains a relatively high caliber of information. To ensure the blocks were non-redundant, blocks-versus-blocks alignments were performed and only those with no significant matches were added to the BLOCKS+ database. BLOCKS+ has more complete coverage of protein sequence space compared to both of the databases discussed earlier. However, though BLOCKS+ is built off of PROSITE and PRINTS, this does not imply the results will be identical. Differences in representation and scoring functions will lead to variable results given a single input.

InterPro[16]

InterPro was developed to integrate the various protein motif predictors into a single interface. InterPro includes Gene3D, PANTHER, PFAM, PIRSF, ProDom, SMART, SUPERFAMILY, TIGRFAMs, PROSITE, and PRINTS. By nature of how each database represents motifs, whether the motifs are single or multiple, and how matches are scored, will change the detection capabilities of each database for a given sequence. No single tool is better than another for all sequences. These databases are complementary to each other rather than competitive. Compiling results from multiple independent servers therefore adds confidence to predictions when the predictions are compatible. BLOCKS+ is not included because is not curated and not independently built. Only PROSITE, PRINTS, and BLOCKS+ are covered in this review because they are the main databases for local one-dimensional motifs.

Three-Dimensional (3D) Motifs

One-dimensional motifs can be identified using a variety of tools and databases. Regions of high conservation can be depicted using various representations and in multiple patterns. These patterns are relatively easy to build because of the wealth of primary sequence information. The number of available sequences also permits robust statistical significance calculations. However, one-dimensional motifs are inherently limited in that they are built solely on sequence conservation. Sequence conservation cannot capture the complex chemistry and orientation required for a functional site. Additionally, as sequences diverge, the signature will become increasingly weak, as only residues absolutely necessary for function will be conserved. A prime example is serine proteases. Representation in primary sequence is difficult, as the crux of the signal arises from the catalytic triad: the serine, histidine, and aspartic acid. Pinpointing three residues in a full length protein sequence is extremely difficult and subject to high false positive rates. Conserved functional sites therefore cannot always be well identified and represented by primary sequence alone.

It is also well understood that structure is more conserved than sequence and that structure is intimately linked to function.[17] When sequence identity enters the twilight zone, sequence conservation patterns rapidly drop off while structural conservation is maintained. These functional sites can be compactly represented as a single site in three-dimensional structure space, but often as multiple segments in one-dimensional sequence space, as is the case for serine proteases. Motif building from a structural approach will be able to take advantage of a strong, focused, local signal. Here, I will discuss a variety of 3D-motif building approaches as well as a meta-server helpful for integrating results from different sources.

Fuzzy Functional Forms[18, 19]

Fuzzy functional forms (FFFs) are three-dimensional descriptors for a functional site based on its geometry and conformation. These can be used to rapidly screen protein structures derived from X-ray crystallography or NMR for a site of interest. Building FFFs begins with identification of functionally important residues through multiple sequence alignments, literature search, and structural examination. Structures containing the sites are superimposed via these residues. Distances and angles between the functionally important residues are then calculated. Performance of the resulting FFFs is assessed against experimentally validated test set structures. The model building procedure must be iterated to find the most informative set of functional residues to include in the model. Information about the secondary structure can also be included if experimentally shown to be important. The end set of distances and angles form the geometrical and conformational constraints for the functional site. This information can then be encoded into computer algorithms to screen new protein structures.

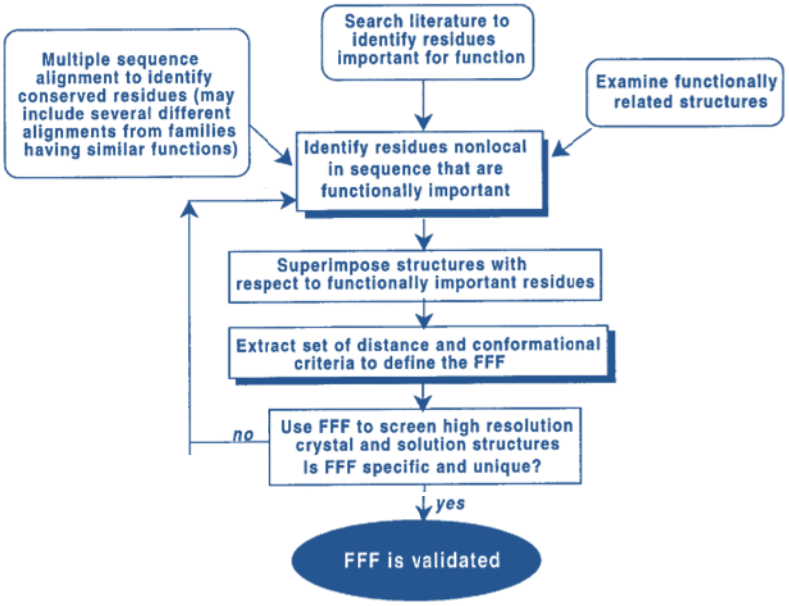


Figure 2. Outline of the protocol for producing a FFF.

FFFs have been shown to have improved performance over the main local sequence motif databases (PROSITE, PRINTS, and BLOCKS+). The models have increased sensitivity in finding functional sites in proteins of unknown function. Few residues are built into the model, meaning there is more tolerance for the surrounding residues compared to 1D motifs. Information regarding the relative positions of the functional residues also helps reduce the false positive rate compared to sequence based methods. Proteins that by chance have similar residues conserved are unlikely to have them in the proper orientation. In depth analysis of residue positions sometimes even provides subfamily classification information. The added layer of three-dimensional structural information onto sequence conservation significantly enhances functional site recognition and interpretation.

Specifically, FFFs have good predictive ability for enzyme active sites, like the catalytic triad of serine proteases. However, a significant drawback is that it requires absolute conservation of the residues built into the model. A model built on a functional site containing an aspartic acid will not recognize a homologous site containing a glutamic acid. These two residues are similar and are often interchanged in many homologs. An additional FFF would need to be built in order to detect both versions of the functional site. FFFs are also time consuming to build, as the functional residues must be manually selected. Including spurious or non-functional residues will significantly reduce the performance of the models. Automation of functional residue selection would greatly increase the usability of FFFs for large scale 3D motif creation.

PROCAT[20, 21]

The PROCAT system uses Template Search and Superposition (TESS), a geometric hashing algorithm to construct 3D structural motifs. Compared to FFFs, it requires significantly less user input for motif building. TESS begins with a single residue of interest, specified by the user. Protein structures containing the functional site are aligned according to this residue. A grid of one angstrom separation is then placed about the residue. For atoms that fall within the grid, their atom type and grid position are noted. Atoms that are significantly enriched within a grid are retained as signatures of the 3D motif. Matches to the motifs are those with an atom corresponding to each atom within the template. To address the scenario where an atom falls along the edge of a box, a potential hit can have the corresponding atom located in an adjacent box. The final score of a match to the motif is based on the root mean square deviation between the corresponding atoms. TESS requires many distance calculations during motif building and database screening. As a result, hash tables are used to store atom position information to reduce computation time.

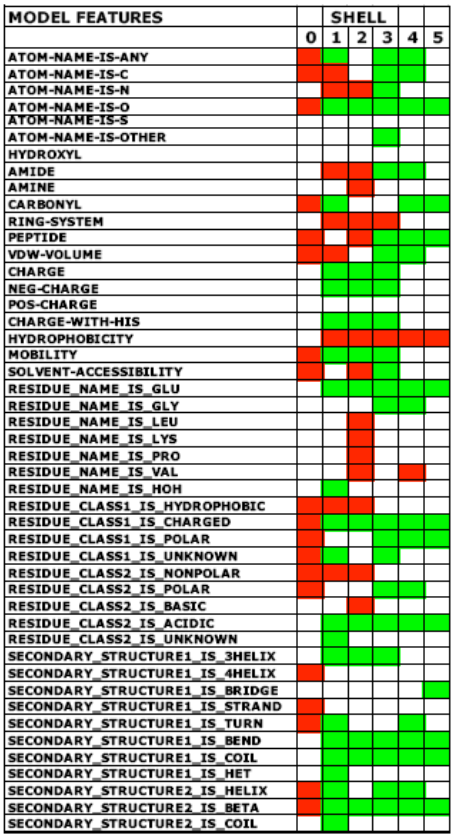
PROCAT, like FFFs, has been shown to work well for enzyme active sites. The incorporation of structural information significantly bolsters performance in terms of specificity and sensitivity. By automating the selection of functional residues, the method is feasible for large scale motif development and screens. However, PROCAT and FFFs only use a small portion of the available structural information. The relative positioning of residues and atoms is the key information that is being used. Information regarding the solvent accessibility and flexibility of the residues is ignored. Secondary structure information is only peripherally included. In the end, both methods grossly oversimplify the physical sites. This has limited their application to enzyme active sites, where positioning of resides and atoms are the key signatures. For functional sites not as precisely oriented as enzyme active sites, incorporation of more structural characteristics would be required for high performance.

FEATURE[22, 23] and SeqFEATURE[24-26]

FEATURE is a machine-learning algorithm that models the local physiochemical microenvironment surrounding a functional site. Building a model requires a set of positive training sites and a set of negative training sites (non-sites). A point in space (either on atom or off atom) around which the microenvironment will be centered must also be specified. A series of radial shells is drawn about this point, and the features within each shell are evaluated (Figure 3). FEATURE uses a broad set of features to define a microenvironment, including atom type, atom name, residue type, residue name, chemical groups, secondary structure, and solvent accessibility. Each feature in a shell is determined to be significantly overrepresented, significantly underrepresented, or not significantly different between sites and non-sites. These probabilities collectively form a functional site model and can be visualized as a “fingerprint” (Figure 3). A query structure is evaluated through a similar process of defining a microenvironment center and assessing the features within the shells. Assuming independence of the features, a naïve-Bayes scoring function can quantitatively assess the likelihood of the query site containing the functional site. Comparison of the query site score to the specificity cutoffs for the model determines whether or not it is a site. Model specificity cutoffs and performance statistics are calculated using k-fold cross-validation.

FEATURE’s strength lies in its fairly comprehensive coverage of the features of a functional site. Sequence conservation, structural characteristics, and location of features are all taken into account. Computational complexity is low by treating the microenvironment as a series of radial volumes. This removes the need to orient structures into a particular reference frame prior to analysis. However, this design loses orientation information associated with the features. A potential improvement to FEATURE is to allow for a more expensive secondary scans where the shells are further divided into quadrants. This would require a pre-processing step to adjust the query structure into the same reference frame as

the model. A secondary scan using multi-quadrant shells can potentially reduce the number of false positive sites, which are unlikely to have a functional orientation of overrepresented atoms.



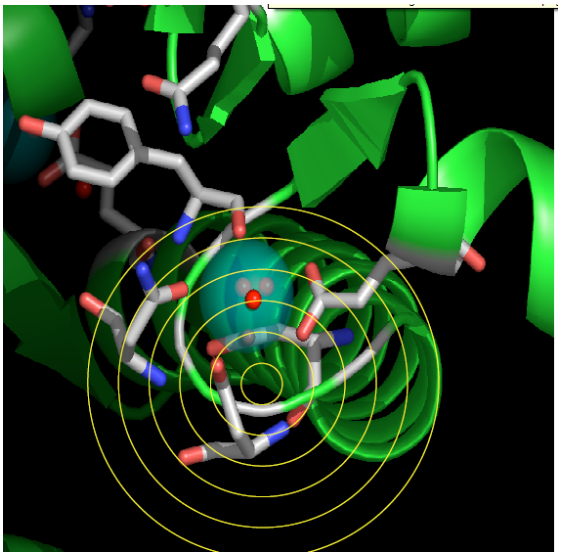


Figure 3.
(LEFT) Model summary represented as a "fingerprint." Green denotes overrepresented and red denotes underrepresented. (TOP) FEATURE microenvironment divided into radial shells.

The dominant drawback of FEATURE is the manual creation of the positive and negative training sets. The performance of the models heavily relies on the quality of the training data. Contamination of either training set with the other will blur the distinction between a site and non-site. However, unlike FFFs and PROCAT, no prior knowledge of the conserved features is required. Any site possessing the function of interest can be used as a positive training structure. FEATURE will automatically discover the important features when building the model. To circumvent the time-consuming process of manually building test sets, and to increase its utility, SeqFEATURE was developed. Given a 1D sequence motif representing a functional site, SeqFEATURE automatically creates a 3D structural motif. It creates a positive training set by finding all non-redundant PDB structures with the motif. A model is then built for each functional atom of the 1D motif. High confidence predictions can be made by collectively scanning with the overlapping models.

SeqFEATURE's performance is competitive against 1D motif databases and other structure-based function predictors. A set of libraries that were automatically built using SeqFEATURE and PROSITE motifs showed lower false positive and false negative rates compared to PROSITE. With test sets composed of proteins with low percent identity to the training set, all 1D motif databases had decreased performance while SeqFEATURE remained consistent. Likewise, for test sets composed of proteins with low structural similarity to the training set, SeqFEATURE outperformed other structure-based methods. SeqFEATURE therefore shows incredible potential for providing functional insight to very divergent proteins.

ProFunc[27]

ProFunc is a server for predicting protein function using both sequence and structure-based methods. It acknowledges that each tool has its strengths and weaknesses and should therefore be used with others to enhance their results. It requires as input a PDB structure and scans this using methods that look at the protein's sequence, methods that look at the protein's structure, and methods that match the structure to other PDB structures using 3D templates. The sequence analysis is fairly exhaustive, as it will call upon InterPro, PDB, UniProt, and SUPERFAMILY.[28] Structural-based analysis begins with secondary structure matching (SSM), which uses a graph-based representation of secondary structure to find fold relatives to the query structure.[29] The structure is also submitted to SURFNET, which will scan for surface clefts.[30] Scans against helix-turn-helix DNA binding motifs and anion and cation binding motifs are also performed. The last set of searches is against templates for enzyme active sites, ligand-binding sites, DNA-binding sites, and 'reverse' templates built from the target structure. The templates define 3D conformations using two to five amino acid residues and are scanned against the query structure to look for matches. ProFunc has yet to incorporate other structure-based function prediction methods like FEATURE and SeqFEATURE. Nonetheless, ProFunc is a key step towards a comprehensive function prediction tool.

Conclusion

Function prediction using structure-based methods offer many advantages over sequence based methods. Primarily, function and structure are intimately linked, making structural conservation significantly more prominent that sequence conservation. It is therefore ideal to build 3D motifs to take advantage of the strong structural signatures for functional site identification. The most evident limitation, when it comes to a comparison of 1D and 3D motifs, is that the available training data for 3D motifs is rather small. Building 3D motifs can also be time consuming both for training set creation and computational evaluation. Nonetheless, there has been significant success with the different three-dimensional models despite these limitations.

As the individual algorithms and approaches continue to be refined and expanded, the lack of conformational diversity for a structure begins to become an issue. Structures deposited in the PDB are a single snapshot of the protein in the case of X-ray crystallography or multiple snapshots of the protein in the case of NMR. It is well understood that proteins do not exist as a single structure in nature, but rather as an ensemble of structures.[31] Side chains and flexible loops are constantly sampling conformational space. Not only are single snapshots incomplete representations but often times inaccurate representations. Crystallization conditions used to solve the structures can introduce artifacts.[32] Functional sites are often not in a functional conformation, and therefore cannot be detected by any structure-based function prediction methods. Generating an ensemble of structures experimentally is not feasible considering cost and experimental complexity. The alternative approach is to generate collections of structures using molecular dynamics (MD). MD simulations are physics based models where the forces on atoms from bond and non-bonded interactions are computed. This translates into accelerations, which when applied to the atoms, form a trajectory over time. MD simulation is computational intensive but recent advancements for GPU-accelerated (Graphical Processing Unit) MD have made it computationally feasible to generate a library of conformational diversity.[33] Preliminary studies have already shown that MD simulation can improve the detection of calcium-binding sites. Follow-ups on the benefits of MD simulation on other models are also being looked into.[34]

All structure-based function prediction methods stand to benefit from conformational libraries. The various methods have constantly been improving their sensitivity and specificity. Eventually, the sensitivity may be restricted by the quality of the data (structures) themselves. It is therefore important

to begin the development of a conformational library in order to realize the full potential of three-dimensional protein structures.

References

1. Thornton, J. Structural genomics takes off. *Trends Biochem Sci* **26**, 88-9 (2001).
2. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res* **32**, D138-41 (2004).
3. Chandonia, J.M. & Brenner, S.E. The impact of structural genomics: expectations and outcomes. *Science* **311**, 347-51 (2006).
4. Marsden, R.L. & Orengo, C.A. Target selection for structural genomics: an overview. *Methods Mol Biol* **426**, 3-25 (2008).
5. Watson, J.D., Laskowski, R.A. & Thornton, J.M. Predicting protein function from sequence and structural data. *Curr Opin Struct Biol* **15**, 275-84 (2005).
6. McGinnis, S. & Madden, T.L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* **32**, W20-5 (2004).
7. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85-94 (1999).
8. Sigrist, C.J. et al. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* **3**, 265-74 (2002).
9. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. *Methods Mol Biol* **406**, 89-112 (2007).
10. Krem, M.M. & Di Cera, E. Molecular markers of serine protease evolution. *EMBO J* **20**, 3036-45 (2001).
11. Attwood, T.K. The PRINTS database: a resource for identification of protein families. *Brief Bioinform* **3**, 252-63 (2002).
12. Henikoff, J.G., Pietrokovski, S., McCallum, C.M. & Henikoff, S. Blocks-based methods for detecting protein homology. *Electrophoresis* **21**, 1700-6 (2000).
13. Henikoff, J.G. & Henikoff, S. Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci* **12**, 135-43 (1996).
14. Servant, F. et al. ProDom: automated clustering of homologous domains. *Brief Bioinform* **3**, 246-51 (2002).
15. Gracy, J. & Argos, P. DOMO: a new database of aligned protein domains. *Trends Biochem Sci* **23**, 495-7 (1998).
16. Hunter, S. et al. InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**, D211-5 (2009).
17. Chothia, C. & Lesk, A.M. The evolution of protein structures. *Cold Spring Harb Symp Quant Biol* **52**, 399-405 (1987).
18. Di Gennaro, J.A. et al. Enhanced functional annotation of protein sequences via the use of structural descriptors. *Journal of Structural Biology* **134**, 232-245 (2001).
19. Cammer, S.A. et al. Structure-based active site profiles for genome analysis and functional family subclassification. *J Mol Biol* **334**, 387-401 (2003).
20. Wallace, A.C., Laskowski, R.A. & Thornton, J.M. 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**, 1001-13 (1996).
21. Wallace, A.C., Borkakoti, N. & Thornton, J.M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci* **6**, 2308-23 (1997).
22. Wei, L. & Altman, R.B. Recognizing protein binding sites using statistical descriptions of their 3D environments. *Pac Symp Biocomput*, 497-508 (1998).

23. Bagley, S.C., Wei, L., Cheng, C. & Altman, R.B. Characterizing oriented protein structural sites using biochemical properties. *Proc Int Conf Intell Syst Mol Biol* **3**, 12-20 (1995).
24. Liang, M.P., Brutlag, D.L. & Altman, R.B. Automated construction of structural motifs for predicting functional sites on protein structures. *Pac Symp Biocomput*, 204-15 (2003).
25. Halperin, I., Glazer, D.S., Wu, S. & Altman, R.B. The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. *BMC Genomics* **9 Suppl 2**, S2 (2008).
26. Wu, S., Liang, M.P. & Altman, R.B. The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation. *Genome Biol* **9**, R8 (2008).
27. Laskowski, R.A., Watson, J.D. & Thornton, J.M. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* **33**, W89-93 (2005).
28. Madera, M., Vogel, C., Kummerfeld, S.K., Chothia, C. & Gough, J. The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res* **32**, D235-9 (2004).
29. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* **60**, 2256-68 (2004).
30. Laskowski, R.A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* **13**, 323-30, 307-8 (1995).
31. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964-72 (2007).
32. Chayen, N.E. & Saridakis, E. Protein crystallization: from purified protein to diffraction-quality crystal. *Nat Methods* **5**, 147-53 (2008).
33. Buck, I. et al. GPUs: stream computing on graphics hardware. *ACM Transactions on Graphics* **23**, 777-786 (2004).
34. Glazer, D.S., Radmer, R.J. & Altman, R.B. Combining molecular dynamics and machine learning to improve protein function recognition. *Pac Symp Biocomput*, 332-43 (2008).