Appeared in Proc. Pacific Symposium on Biocomputing, pp. 334–345, 2005

ALGORITHMS FOR STRUCTURAL COMPARISON AND STATISTICAL ANALYSIS OF 3D PROTEIN MOTIFS

BRIAN Y. CHEN[1], VIACHESLAV Y. FOFANOV[2], DAVID M. KRISTENSEN[3], MAREK KIMMEL[2], OLIVIER LICHTARGE[3,4], LYDIA E. KAVRAKI[1,3]

Rice University Departments of Computer Science[1] and Statistics[2], Houston, TX 77005, USA

Baylor College of Medicine Struct. and Comp. Biol. and Mol. Biophys. Prog.[3] and Department of Molecular and Human Genetics[4], Houston, TX 77030, USA

The comparison of structural subsites in proteins is increasingly relevant to the prediction of their biological function. To address this problem, we present the Match Augmentation algorithm (MA). Given a structural *motif* of interest, such as a functional site, MA searches a *target* protein structure for a *match*: the set of atoms with the greatest geometric and chemical similarity. MA is extremely efficient because it exploits the fact that the amino acids in a structural motif are not equally important to function. Using motif residues ranked on functional significance via the Evolutionary Trace (ET), MA prioritizes its search by initially forming matches with functionally significant residues, then, guided by ET, it augments this partial match stepwise until the whole motif is found. With this hierarchical strategy, MA runs considerably faster than other methods, and almost always identifies matches in homologs known to have cognate functional sites. Second, in order to interpret matches, we further introduce a statistical method using nonparametric density estimation of the frequency distribution of structural matches. Our results show that the hierarchy of functional importance within structural motifs speeds up the search within targets, and points to a new method to score their statistical significance.

1. Introduction

Determining the function of proteins remains a primary goal of biology[1]. Tools such as PSI-BLAST[2], EMATRIX[3], and PROSITE[4] help predict function using sequence similarity. But with the increasing availability[5] of protein structures, other techniques have been developed to predict function via geometric comparison of functional subsites, such as JESS[6], PINTS[7], webFEATURE[8], and Geometric Hashing[9]. Starting from the basic observation that biological motifs are hierarchical in nature, this paper contributes a hybrid technique combining evolutionary information[10,11] with protein geometry and chemical labels to efficiently identify proteins with local structural similarity to a motif of interest.

Contributions and Outline We wish to annotate protein structures with structural motifs that are functionally relevant. This problem naturally decomposes into the problem of motif design, and the problem of motif search. The Evolutionary Trace[12,11,13] (ET) was developed to identify functionally relevant motifs, and the topic of this paper is two novel algorithms for mo-

tif search and statistical interpretation of matches. Match Augmentation (MA) takes ET-based motifs, composed of subsets of three-dimensional (3D) protein structures ranked by evolutionary significance via ET and labeled by amino acid identity. MA hierarchically searches target structures of whole proteins, labeled with amino acids, for a match. Section 2.4 demonstrates that heuristic prioritization based on evolutionary rankings allows MA to identify cognate active sites in homologous proteins, with a 60 fold speed-up.

Much like sequence comparisons, MA may identify similar structures by chance alone. Hence, we ask if matches of cognate active sites have significantly greater structural similarity than what is expected by chance. In the second part of this paper, we apply Nonparametric Density Estimation (NDE) to attach statistical significance to a match, much like the BLAST p-value[2]. In Section 3.2, while current motifs are not yet optimal, and the test set is not complex, our results nevertheless verify that statistically significant matches are correlated with identifying cognate active sites.

2. Match Augmentation (MA)

Several methods exist for comparing protein structure, such as SSAP[14], DALI[15], tools for graph theoretical comparison[16], and Geometric Hashing[9], which was adapted to alignment by atom position[17], by backbone C-alpha[18], multiple structural alignment[19], and alignment of hinge-bending and flexible protein models[18]. More specifically, the problem of searching for motifs within protein structures, has also been approached using Geometric Hashing to find catalytic triads[20], using JESS[6], PINTS[7], and webFEATURE[8]. All structural comparison techniques share a dependence on heuristics, because the complexity of the structural pattern matching problem is at least NP-hard[21]. Heuristics are essential because biological input is too large for exhaustive approaches.

Our algorithm, MA, is unique because it uses the evolutionary significance of an amino acid in combination with structural and chemical data, stored as amino acid labels, to produce added performance with a novel combination of hashing and backtracking.

2.1. *Definition of Hierarchical Motifs*

Structure comparison is fundamentally hard[21], but heuristics using evolutionary data may improve performance. One source of such data is the Evolutionary Trace, which identifies functionally significant residues via Multiple Sequence Alignment (MSA) of homologous proteins[22,11]. For each residue, ET produces evolutionary significance *ranks*, which quantify the

relative importance of individual residues to the function of the protein, and a list of functionally-compatible amino acid *alternates*, where mutation of a residue to its alternates is tolerated during evolution. MA can also accept rank information from other sources.

Ranks facilitate a prioritized attitude towards geometric and chemical comparison. Between high ranked motif points and corresponding target points, geometric and chemical similarity intuitively suggests greater functional similarity than correspondences involving low ranked motif points. Therefore, we seek correspondences for high ranking motif points before low ranking motif points, in a prioritized manner. Prioritization is the center point of our algorithmic design.

Our motifs $S = \{s_1, \ldots, s_m\}$ are sets of m points in space whose coordinates are taken from backbone and sidechain atoms of high ranking residues around a ligand binding site, or other functional structure. Each point s_i in the motif, or *motif point*, has an associated rank $p(s_i)$ and a set of alternate amino acid *labels* $l(s_i) = \{a_1, a_2, ...\}$ taken respectively from the significance rank and alternate sidechains generated by ET. Typically, our motifs are between 4 and 9 motif points.

2.2. *The Problem*

We seek a correspondence between S and the target T, often hundreds of atoms encoded as n *target points*: $T = \{t_1, \ldots t_n\}$, where each t_i is taken from atom coordinates, and labeled $l(t_i)$ for the amino acid t_i belongs to.

The correspondence is a *match* M, a bijection between $\{s_{M_i} \ldots s_{M_i}\} \in S$ and $\{t_{M_i} \ldots t_{M_i}\} \in T$ of the form $M = \{(s_{M_1}, t_{M_1}) \ldots (s_{M_m}, t_{M_m})\}$, with Euclidean distance between points a and b defined as $||a - b||$ and:

> **Criterion 1** $\forall i$, s_{M_i} and t_{M_i} are biologically *compatible*: $l(t_{M_i}) \in l(s_{M_i})$.
> **Criterion 2** LRMSD alignment, via rigid transformation A of S, causes $\forall i, ||A(s_{M_i}) - t_{M_i}|| < \epsilon$, our threshold for geometric similarity.

MA identifies the match with smallest LRMSD among all matches that have paired all s_{M_i} to distinct t_{M_i}. Matches of subsets of S to T are rejected.

2.3. *Description of the Algorithm*

Following our prioritized data, we designed MA in a prioritized fashion, where correspondences with higher ranked points are identified first. MA is composed of two parts: *Seed Matching* and *Augmentation*. The purpose of Seed Matching is to identify a match for the *seed* $S' = \{s_1, s_2, s_3\}$, the three highest ranked motif points. The k lowest LRMSD *seed matches* are passed to Augmentation to be iteratively expanded into matches for the remaining motif points, in descending rank order. Augmentation outputs

the match with smallest LRMSD. We use $k = 30$ and $\epsilon = 3.0$Å.

Seed Matching We must find k sets of 3 target points $T' = \{t_A, t_B, t_C\}$ which are compatible with $S' = \{s_1, s_2, s_3\}$, respectively, with similar inter-point distances as S'. Interpret T as a geometric graph, where target points are vertices. Suppose t_i,t_j are compatible with s_1,s_2. Then if $-2\epsilon \leq ||t_i - t_j|| - ||s_1 - s_2|| \leq 2\epsilon$, we define a green edge between t_i and t_j. Similarly, red and blue edges are defined between target points compatible with s_1,s_3 and s_2,s_3 respectively, where again inter-point distances are within 2ϵ.

Edges are found by range search on a geometric data structure. When we find an edge, if it forms a triangle of all three colors, we have a seed match with compatible labels and similar inter-point distances. When we identify a triangle, LRMSD with S' is calculated and if all points are aligned within ϵ, the new seed match is stored. The k lowest LRMSD seed matches are passed to Augmentation. Targets of size n have at most $\binom{n}{3} = O(n^3)$ matching triangles, but this worst case would be a geometrically regular set of identical triangles, which never occurs in natural proteins. Performance on biological data is commonly $O(n^2)$.

Augmentation Augmentation is an application of depth first search. Given a seed match, we must find correspondences for unmatched motif points within the target. Considering the LRMSD alignment of the seed matches, we plot the position of the highest ranked unmatched motif point s_i as if it were rigidly aligned with the rest of the seed. In the spherical vicinity V of this position, we identify all $t_i \in T$ compatible with s_i. For each t_i, we calculate the LRMSD alignment A of the seed with (s_i, t_i). If $||A(s_i) - t_i|| < \epsilon$, the seed match, with (s_i, t_i), becomes a *partial match*.

V often contains several t_i compatible with s_i. We test all t_i, storing accepted partial matches on a stack. After all t_i are tested, we pop the first partial match off the stack, and begin testing with the next unmatched motif point s_{i+1}. This is essentially depth first search (DFS), implemented with a stack. When no unmatched s_i remain, or no compatible t_i within V can be aligned to satisfy $||A(s_i) - t_i|| < \epsilon$, LRMSD is calculated for the entire match, and the match is stored. Final output is the match of all s_i to distinct t_i with lowest LRMSD.

Performance is dependent on the number of motif points m, and c_r, the number of compatible t_i found in V, giving runtime $O(m^2(c_r^{m-3}))$. c_r is bounded because repulsive Van der Waals forces limit the number of atoms found in V. The quadratic factor is the aggregate cost of LRMSD calculations, and the exponential is the cost of DFS with c_r possibilities per iteration. With m usually 4-9 points, MA is extremely efficient.

2.4. *Experimental Results*

To demonstrate the accuracy of MA, we searched for motifs within structures of evolutionarily related proteins. We use targets identified by sequence similarity because each residue in the motif has a cognate residue in the target: we know what match to expect beforehand. Using functional analogs may seem more relevant for functional annotation, but successfully matching analogs would only demonstrate how well our motifs represent function. Our focus is on methodology, and because analogs lack easily identifiable cognate residues, their use would sacrifice precise verifiability.

Data Set Our primary data (Figure 1) is 12 families of enzymes with known active sites. Each family is composed of a set of homologous sequences identified by BLAST, some of which have known structures in the Protein Data Bank[23] (PDB). Of the structures found, each family is assigned a *major* structure; the rest are *minor*. ET is applied on each family of sequences, and the significance ranks and labels generated are mapped onto the major structure for each family. Between 4 and 9 of the most functionally significant residues surrounding the active site on the major protein are selected, and their alpha carbons become the points in the motif. Specifics on amino acid selection and functional sites used for each motif can be found at `http://www.cs.rice.edu/~brianyc/papers/PSB2005/`.

Alpha carbons (C_α) were used in our motifs as preliminary data. Rather than debate the adequacy of C_α atoms to represent function, we seek only to document the correctness of our techniques. Future publications will comparatively document issues of motif design on a larger scale.

{**16pk**, 1vpe, 1php} {**1bqk**, 8paz, 1aaj, 1aan, 1ag6, 1b3i, 1baw, 1bxa, 1bxv, 1paz, 1pza, 1pzb, 1pzc, 1zia, 1zib, 2plt, 2rac, 3paz, 1aac} {**1amk**, 1tpe} {**1aky**, 5ukd, 1qf9, 1uke, 1zin, 1zio, 1zip, 2ak2, 2ukd, 3ukd, 4ukd, 1ak2} {**1a6m**, 1ymc, 1dwr, 1dws, 1dwt, 1m6c, 1mbs, 1mno, 1mwd, 1myg, 1pmb, 1wla, 1ymb, 1azi} {**1a3k**, 1slt, 1sla, 1slc, 1qmj} {**1finA**, 1hcl, 1hck, 1b38} {**1ukrA**, 1xyn, 1xnb, 1yna} {**3lzt**, 2ihl, 2lz2, 1jhlA, 1ghlA, 1fbiX, 1lz3, 1hhl, 1jug, 2eql, 1gd6A, 1f6rA, 1hfx} {**7a3hA**, 1g01A,1egzA} {**1juk**, 1j5tA, 1i4nA} {**1f8eA**, 1nn2, 1nsbA}

Figure 1. Families (bracketed) used in experimentation. Bolded proteins are major.

Experimental Protocol We search for each motif in the minor structures of the same family. These are homologous proteins (HPs). ET uses MSAs, so a functional residue in one sequence correlates with cognate residues of related function, at the same position, in all sequences of the family. Thus we can verify MA: if we find a *cognate match* where the target points are

cognate to the motif points, we have a correct match, residue by residue. For comparison, we also searched for each motif in the minor proteins of the other families. These proteins are not homologous (NHPs).

Results In 69 out of the 73 motif-HP pairs (95.4%), MA matches 100% of the source motif with cognate residues in the target. Of the remaining four cases, two of the target structures (1m6c and 1mno) were experimental structures that had a point mutation which changed the label of residue 68 (in both cases) from a valine to an asparagine in order to over-stabilize oxygen binding in myoglobin (1a6m). As a result, the labels of the points corresponding to residue 68 in both 1m6c and 1mno were incompatible, and, correctly, the points were not matched. While this was not intended, it demonstrates the ability of our algorithm to eliminate potential matches with incorrect labels. In the other two cases, a match existed with lower LRMSD than the cognate match. These occurred between major protein 1amk with target 1tpe, and 1f8eA with 1nsbA. In each case the cognate match had a higher LRMSD (approx. .5Å) than the match MA identified. This is no fault of MA. Instead, it suggests that 1amk and 1f8eA are sub-optimal motifs, which bear accidental similarity to functionally unrelated structures: Ideally, motifs should have structural similarity only with proteins with functional similarity. True failures of MA would be the opposite: We would return a match with LRMSD higher than the cognate match, showing that the cognate match was overlooked. This never occurs. From our experiments, we found that MA is accurate and efficient on biological data, identifying cognate residue correspondences, except when the motif bears incidental structural similarity to unrelated residues.

Matches between motifs and HPs tended to have lower LRMSDs than between the same motif and NHPs. This is apparent in Figure 2, which plots LRMSD for all matches found. 9 out of 12 motifs considered had matches of HPs (Blue, Fig. 2) with LRMSD lower than most matches of NHPs (Red, Fig. 2). Two of the motifs breaking this trend were 1amk and 1f8eA, motifs which had incidental similarity with functionally unrelated residues, suggesting again that these motifs are not specific representatives of function. The remaining motif, 1finA, was defined on a flexible active site, so cognate active sites, flexible themselves, had less geometric similarity.

Performance We compared performance to our implementation of Geometric Hashing (GH), as described by Rosen[24], because the source code is not available. All published heuristics compatible with our data were implemented. GH has been applied many times[17,18,20,19], but cannot be prioritized as is the case with MA. GH identified identical HP matches and

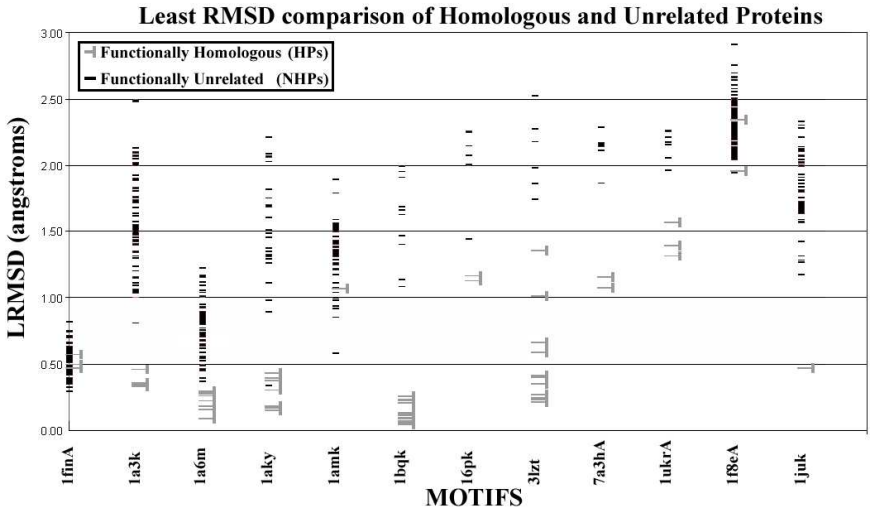

Figure 2. Experimental Results: 12 motifs and 73 targets plotted by LRMSD

similar NHP matches, but on our motifs of 4 to 9 motif points, and targets with 123 to 398 target points, MA was about 60 times faster. Average execution time was 6.195 seconds for GH, and only 0.103 seconds for MA using identical thresholds. Without loss of accuracy, Seed Matching narrows the search to matches of the highest ranking motif points, whereas GH considers all points equally. Evolutionary prioritization seems to strongly improve performance. Experiments were run on Athlon 1900+ CPUs. GH and MA memory footprints varied between 5 and 20 megabytes, depending on input.

3. Statistical Analysis of Geometric Similarity

Structural similarity is important to functional annotation only if a strong correlation exists between identifiably significant structural similarity and functional similarity. However, as seen in Figure 2, algorithms like MA and GH can identify matches in NHPs with unrelated functions, so the existence of a match alone does not guarantee functional similarity. LRMSD can be a differentiating factor. If matches of HPs represent statistically significant structural similarity over what is expected by random chance, we could differentiate on LRMSD, as long as we can evaluate the statistical significance of the LRMSD of a match.

BLAST[2] first calculated the statistical significance of sequence matches with a combinatorial model of the space of similar sequences. Determining

the statistical significance of structural matches has also been attempted. Modeling was applied for the PINTS database[7] to estimate the probability of a structural match given a particular LRMSD. An artificial distribution was parameterized by motif size and amino acid composition in order to fit a given data set, and the p-value is calculated relative to that distribution. Another approach was taken in the algorithm JESS[6], using comparative analysis to generate a significance score relative to a specific population of known motifs. Both methods have some disadvantages. The artificial models of PINTS are not parameterized by the geometry of motifs, and, all else equal, produce identical distributions for motifs of different geometry. JESS, on the other hand, is dependent on a set of known motifs; should this set change, all significance scores would have to be revised.

3.1. *A Method for Characterizing Geometric Similarity*

We begin by defining the statistical significance of geometric similarity. A match of motif S to target T with LRMSD r is statistically significant if the *p-value*, the probability of finding a target T' within a space of proteins Z, where the best match identified has LRMSD $r' < r$, is very low.

Determining the p-value is difficult because it requires the frequency distribution D_S of match LRMSD between a given motif S and all possible targets: all protein structures. But we have little knowledge of the space of protein structures; many proteins defy current techniques for structure determination. Rather than hypothesizing about unknown proteins, we use a set of known structures Z, and accept that our p-value reflects only this concrete set. Specifically, any Z can have biases in the structural similarity of its members to some motif. This doesn't make the selection of Z poor, because the primary purpose of our technique is to reflect that bias in the p-value calculated. We use the Protein Data Bank[23] as Z: the set $\{PDB\}$.

D_S is essentially a histogram of how many proteins match S at any LRMSD. Once D_S is determined it can be interpreted as a distribution density function, which can be integrated to find the probability $P(r)$ of finding a match within $\{PDB\}$ with LRMSD less than a given r:

$$P(r) = \int_0^r D_S \tag{1}$$

One basic assumption of PINTS[7] and JESS[6] was that explicit calculation of D_S is computationally infeasible; that running algorithms like MA with a given motif and every target would take too long. We tried this approach first, finding a match between S and each member of $\{PDB\}$. This brute force approach generated D_S in 3 hours for some motifs, or up

to 631 for others. While this is acceptable for some applications, others, such as motif design, require frequent updates, for which this is too long. Scanning is embarrassingly parallel, but we provide a simpler solution first.

We used random sampling of the $\{PDB\}$ to avoid considering every target. D_S was estimated using Nonparametric Density Estimation[25] (NDE). The distribution D_S estimated from the sample needs to be smoothed, to neutralize spikes caused by the practice of submitting numerous similar structures to the PDB, and to interpolate between our sample points. Kernel Density Smoothing[25] was applied with a gaussian kernel to smooth the data. To avoid undersmoothing or oversmoothing, optimal bin-width determined by S-J estimation[26] was deemed best[27].

3.2. *Experimental Results*

Nonparametric Density Estimation We begin by demonstrating the effectiveness of sampling. We use a snapshot of the PDB from 8.17.2003. PDB files with multiple chains were divided into individual files, generating 55,305 structures. A handful of unparseable files were removed, and certain degeneracies were fixed, such as negatively indexed residues.

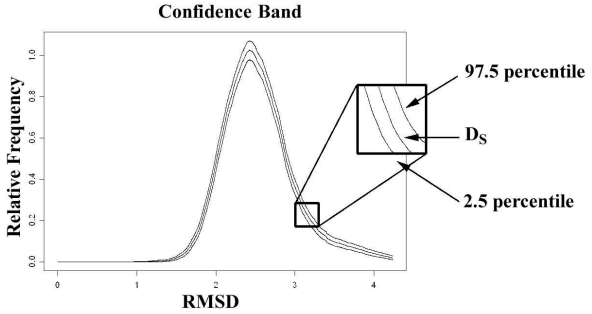

Figure 3. Quality of sampling fit

$\{PDB\}$ was scanned using MA, with each motif S_i from Section 2.4. Brute force generated a reference distribution D_{S_i}. To verify sampling stability, each D_{S_i} was sampled at 5%, 5,000 times. For all S_i, 95% of sampled curves fell within confidence bands tight around D_{S_i}. The confidence band, Figure 3, graphing frequency to LRMSD for D_{3lzt}, is typical of how tightly D_{S_i} is approximated. Kolmogorov-Smirnoff[28] tests confirmed a lack of statistically significant differences between sampled distributions and D_{S_i}. Non-redundant PDB subsets produced no significant differences from $\{PDB\}$.

Random sampling directly improves performance. Brute force computation time was 12:48 (hrs:mins) on average, while sampling took 0:38 on average. The best case fell from 2:40 to 0:08 and the worst case from 631:41 to 31:30. Sampling cuts runtime by almost exactly 95%. Sampling does efficiently estimate D_{S_i} without statistically significant loss of accuracy.

Revisiting Earlier Results After generating D_S for all motifs from the previous section, in Figure 4, we calculated p-values for each LRMSD from Figure 2. The majority of p-values generated for HPs were between 1% and 0.01%. In contrast, most p-values generated for NHPs are above 10%. Notable exceptions are the p-values for matches of motifs 1amk and 1f8eA, which had accidental similarity to functionally unrelated structures. These had p-values above 10%. This verifies on a PDB-scale that 1amk and 1f8eA poorly represent functional sites: they have geometric and chemical similarity to 10% of all PDB proteins. The motif defined on 1finA, which had a flexible active site, also lacks statistical significance in its matches, because the geometry of functional residues may change relative to the motif. Matches of HPs represent identifiably significant structural similarity, except where the motif itself poorly represents protein function.

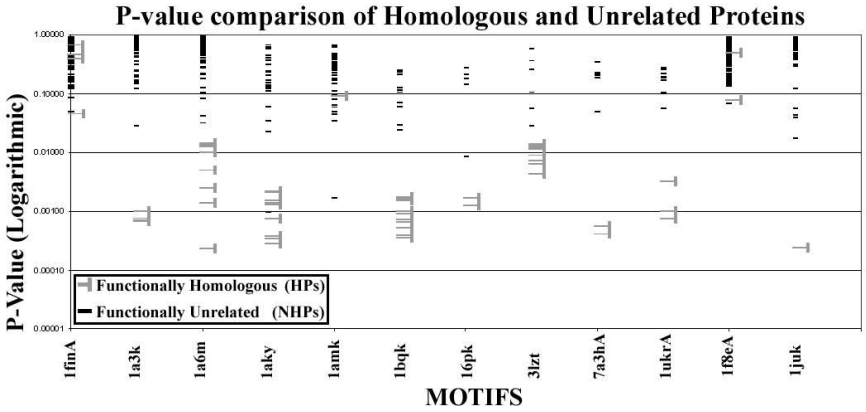

Figure 4. p-values of LRMSDs from Figure 2 (log scale)

Discussion NDE avoids inflexibility, characteristic of parametric approaches[7], because it is not limited to a parametric model. While generally considered less powerful in other applications, NDE is appropriate here given breadth and complexity of protein structure space. Sampling, combined with NDE, greatly accelerates the process of calculating p-values.

On our data set of evolutionarily related proteins, our results show a

correlation of statistically significant structural similarity to evolutionary relatedness between proteins, as long as the motifs properly represent function. This correlation indicates that statistically significant geometric and chemical similarity can be markers of cognate active sites.

4. Summary and Future Work

MA efficiently identifies homologs via rigid structural comparison. On our data set, 95.4% of active sites cognate to a given motif were correctly identified, and the remainder were not found because of the difficulty of representing some active sites with motifs. By optimizing on evolutionary data, MA is about 60 times faster than standard GH.

NDE via sampling calculates the statistical significance of matches identified. Testing indicates that we can drastically cut the number of calculations necessary to estimate D_S without significant loss of accuracy. Furthermore, our results mesh with previous observations: matches between motifs and HPs were statistically significant, except for motifs which poorly represent protein function. Statistically significant LRMSD is correlated with the detection of cognate active sites.

This paper presents a fast method for identifying matches which permits an efficient statistical analysis of our data. Our future studies will develop methods for motif design, and test the sensitivity and specificity of these methods for functional annotation.

Acknowledgements This work is supported by the National Science Foundation NSF DBI-0318415. Additional support is gratefully acknowledged from training fellowships the Gulf Coast Consortia (NLM Grant No. 5T15LM07093) to B.C. and D.K.; from March of Dimes Grant FY03-93 to D.K.; from a Whitaker Biomedical Engineering Grant and a Sloan Fellowship to L.K; and from a VIGRE Training in Bioinformatics Grant from NSF DMS 0240058 to V.F. Experiments were run on equipment funded by AMD and EIA-0216467.

References

1. Jones S. et. al. Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.*, 8(1):3–7, 2004.
2. Altschul S.F. et. al. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids. Res.*, 25(17):3389–3402, Sept 1997.
3. Wu T.D, Nevill-Manning C.G, and Brutlag D.L. Fast probabilistic analysis of sequence function using scoring matrices. *Bioinf.*, 16(3):233–44, 2000.
4. Hulo N. et. al. Recent improvements to the PROSITE database. *Nucl. Acids. Res.*, 32:D134–D137, 2004.
5. Blundell T.L., Jhoti H., and Abell C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Disc.*, 1:45–54, 2002.

6. Barker J.A. and Thornton J.M. An algorithm for constraint-based structural template matching. *Bioinf.*, 19(13):1644–1649, 2003.
7. Stark A., Sunyaev S., and Russell R.B. A model for statistical significance of local similarities in structure. *J. Mol. Biol.*, 326:1307–1316, 2003.
8. Laing M.P. et.al. Webfeature: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucl. Acids Res.*, 31(13):3324–7, 2003.
9. Wolfson H.J. and Rigoutsos I. Geometric hashing: An overview. *IEEE Comp. Sci. Eng.*, 4(4):10–21, Oct 1997.
10. Lichtarge O. and Sowa M.E. Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.*, 12(1):21–27, 2002.
11. Mihalek et. al. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, 336:1265–82, 2004.
12. Lichtarge O., Bourne H.R., and Cohen F.E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257(2): 342–358, 1996.
13. Yao H. et. al. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, 326:255–261, 2003.
14. Taylor W,Orengo A. Protein structure alignment. *J.Mol.Biol*, 208:1–22, 1989.
15. Holm L. and Sander C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233:123–138, 1990.
16. Artymuik P.J. et. al. A graph-theoretic approach to the identification of 3D patterns of amino acid side chains. *J.Mol.Biol*, 243:327–44, 1994.
17. Bachar O. et. al. A computer vision based technique for 3D sequence independent structural comparison of proteins. *Prot. Eng.*, 6(3):279–288, 1993.
18. Verbitsky G., Nussinov R., and Wolfson H. Structural comparison allowing hinge bending. *Prot: Struct. Funct. Genet.*, 34(2):232–254, 1999.
19. Liebowitz N et al. Automated multiple structure alignment and detection of a common substructural motif. *Prot: Struct. Func. Genet*, 43:235–245, 2001.
20. Wallace A.C., Laskowski R.A., and Thornton J.M. Derivation of 3D coordinate templates for searching structural databases. *Prot.Sci*, 5:1001–13, 1996.
21. Akutsu T. On determining the congruity of point sets in higher dimensions. In *Proc. ISAAC: 5th Symp. Alg. Comp.*, 1994.
22. Sowa M.E. et. al. Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat. Struct. Biol.*, 8:234–237, 2001.
23. H.M. Berman et al. The protein data bank. *Nucl.Acids Res*, 28:235–42, 2000.
24. Rosen M. et. al. Molecular shape comparisons in searches for active sites and functional similarity. *Prot. Eng.*, 11(4):263–277, 1998.
25. Silverman B.W. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall: London, 1986.
26. Jones M.C., Marron J.S., and Sheather S.J. A brief survey of bandwidth selection for density estimation. *J. Amer. Stat. Assoc.*, 91:401–407, Mar 1996.
27. Sheather S.J. and Jones M.C. A reliable data-based bandwidth selections method for kernel density estimation. *J. Roy. Stat. Soc.*, 53(3):683–690, 1991.
28. Birnbaum Z.W. and Tingey F.H. One-sided confidence contours for probability distribution functions. *Ann. Math. Stat.*, 22(4):592–596, Dec 2003.