A Statistical Model to Correct Systematic Bias Introduced by Algorithmic Thresholds in Protein Structural Comparison Algorithms

Viacheslav Y. Fofanov[1,§], Brian Y. Chen[2,§,*], Drew H. Bryant[2], Mark Moll[2],
Olivier Lichtarge[3,4], Lydia Kavraki[2,3,4], and Marek Kimmel[1]

[1]*Dept. of Statistics,* [2]*Dept. of Computer Science, Rice University,* [3]*Structural and Computational Biology and Molecular Biophysics,* [4]*Dept. of Molecular and Human Genetics, Baylor College of Medicine*

vfofanov@rice.edu, brianyc@rice.edu, dbryant@rice.edu, mmoll@cs.rice.edu,
lichtarge@bmc.tmc.edu, kavraki@cs.rice.edu, kimmel@rice.edu

[§]equal contribution, [*]author is now with the Dept. of Biochemistry and Biophysics, Columbia University, bc2272@columbia.edu

Abstract

The identification of protein function is crucial to understanding cellular processes and selecting novel proteins as drug targets. However, experimental methods for determining protein function can be expensive and time-consuming. Protein partial structure comparison methods seek to guide and accelerate the process of function determination by matching characterized functional site representations, motifs, to substructures within uncharacterized proteins, matches. One common difficulty of all protein structural comparison techniques is the computational cost of obtaining a match. In an effort to maintain practical efficiency, some algorithms employ efficient geometric threshold-based searches to eliminate biologically irrelevant matches. Thresholds refine and accelerate the method by limiting the number of potential matches that need to be considered. However, because statistical models rely on the output of the geometric matching method to accurately measure statistical significance, geometric thresholds can also artificially distort the basis of statistical models, making statistical scores dependant on geometric thresholds and potentially causing significant reductions in accuracy of the functional annotation method. This paper proposes a point-weight based correction approach to quantify and model the dependence of statistical scores to account for the systematic bias introduced by heuristics. Using a benchmark dataset of 20 structural motifs, we show that the point-weight correction procedure accurately models the information lost during the geometric comparison phase, removing systematic bias and greatly reducing misclassification rates of functionally related proteins, while maintaining specificity.

1. Introduction

Hundreds of protein structures are added to the Protein Data Bank [6] each month, many with unknown function [9], highlighting the need for high-throughput methods for functional annotation. Experimental methods, however, are expensive, time consuming, and often require significant expert knowledge. A number of computational methods have been developed to accelerate this process, ranging from amino acid sequence- [2, 10] to protein structure-based methods [4, 5, 13, 25, 30, 35], among others. While these approaches have already been shown to be useful tools to suggest protein function [34, 36], further improvements in sensitivity and specificity are desirable.

Partial structural comparison methods, such as Geometric Hashing [35], JESS [5], PINTS [30], LabelHash [25] and Match Augmentation [13], are based on the underlying biological hypothesis that for a large portion of proteins, their functionality can be traced to just a few active residues, the active site, while the rest of the proteins structure is not directly involved in its function. As such, these methods approach functional annotation by concentrating their attention on protein active sites and identifying instances of greatest geometric and chemical similarity (*matches*) between models of known active sites (*motifs*) and substructures within functionally uncharacterized proteins (*targets*).

Finding a match in a target protein does not guarantee functional similarity. Partial structural comparison methods can identify spurious matches with similar atoms but geometrically dissimilar configurations, which are unlikely to be biologically relevant. To separate matches of functionally related proteins from spurious matches found by chance, statistical models of substructural similarity are coupled with partial structural comparison methods [5, 13, 33, 25]. These models establish, for each motif, a threshold of baseline geometric similarity necessary to imply functional similarity, based on the statistical significance score of a match.

One common limitation of partial structural comparison approaches is the computational expense of identifying the match of greatest geometric and chemical similarity when potential matches have low overall similarity [31]. To identify biologically relevant matches, heuristics-based structural searches [5, 13, 33, 20, 25] have been used to eliminate matches with so little geometric similarity that functional

978-1-4244-2890-8/08/$25.00 ©2008 IEEE

similarity is biologically unlikely. For example, after establishing the chemical compatibility of matched residues, JESS [5], PINTS [33], LabelHash [25] and Match Augmentation [13] use geometric thresholds to rapidly eliminate potential matches if pair-wise distances between residues in optimal alignment exceed specified thresholds.

Unfortunately, eliminating matches based on geometric thresholds, while greatly reducing computing time, may impact the accuracy of statistical significance scores. Statistical models use the distribution of matches between a given motif and a set of targets to evaluate the statistical significance of a match. If matches are eliminated by the geometric thresholds, the resulting statistical model may be distorted. The generated significance scores may depend on the geometric thresholds, resulting in significant misclassification rates of functionally related proteins.

We will illustrate that different geometric thresholds used by structural comparison methods can introduce serious bias. In order to correct this effect without modifying the original structural comparison algorithm, we have developed a statistical framework which takes into account eliminated matches. We used Match Augmentation as an example, but the technique can be generalized to other partial geometric comparison methods. Our bias correction models the information lost during the structural comparison phase of functional annotation and provides substantially reduced misclassification rates for functionally related proteins. On a test dataset of motifs derived from 20 unrelated enzymatic active sites, we illustrate that with our correction, even at strict levels of geometric thresholds, it is possible to achieve sensitivity and specificity previously reserved only for generous thresholds that required large computation times.

Table 1. Dataset. EC classifications and residues chosen for motifs used. Motifs containing functionally documented residues are indicated by *.

PDB ID	EC class	Amino acids chosen	EC class size
16pk*[7]	2.7.2.3	$R_{39},P_{45},G_{376},G_{399},K_{202}$	27
1ady*[1]	6.1.1.21	$E_{81},T_{83},R_{112},E_{130},Y_{264},R_{311}$	22
1ani*[14]	3.1.3.1	$D_{51},D_{101},S_{102},R_{166},H_{331},H_{412}$	82
1ayl	4.1.1.49	$L_{249},S_{250},G_{251},G_{253},K_{254},T_{255}$	19
1b7y[29]	6.1.1.20	$W_{149},H_{178},S_{180},E_{206},$ Q_{218},F_{258},F_{260}	20
1czf	3.2.1.15	$D_{180},D_{201},D_{202},A_{205},$ $G_{228},S_{229},R_{256},K_{258},Y_{291}$	21
1did*[15]	5.3.1.5	$F_{25},H_{53},D_{56},F_{93},W_{136},K_{182}$	153
1dww*[16]	1.14.13.39	$C_{194},V_{346},F_{363},W_{366},$ Y_{367},E_{371},D_{376}	239
1ep0	5.1.3.13	$S_{53},R_{61},H_{64},K_{73},R_{90},D_{172}$	39
1ggm*[3]	6.1.1.14	$E_{188},R_{311},E_{239},E_{341},E_{359},S_{361}$	11
1jg1	2.1.1.77	$E_{97},G_{99},G_{101},D_{160},L_{179},G_{183}$	17
1juk	4.1.1.48	$E_{51},S_{56},P_{57},F_{89},G_{91},$ $F_{112},E_{159},N_{180},S_{211},G_{233}$	12
1kp3	6.3.4.5	$R_{106},F_{139},E_{202},L_{286},R_{288},Y_{331}$	36
1kpg	2.1.1.79	$D_{17},G_{72},G_{74},W_{75},G_{76},F_{200}$	13
1lbf	4.1.1.48	$E_{51},S_{56},P_{57},F_{89},G_{91},$ $F_{112},E_{159},N_{180},S_{211},G_{233}$	12
1nsk	2.7.4.6	$K_{12},P_{13},Y_{52},R_{105},N_{115},H_{118}$	203
1ucn	2.7.4.6	$K_{12},P_{13},G_{92},R_{105},N_{115},H_{118}$	203
2ahj	4.2.1.84	$P_{53},L_{120},Y_{127},V_{190},D_{193},I_{196}$	39
7mht	2.1.1.73	$P_{80},C_{81},S_{85},E_{119},R_{163},R_{165}$	11
8tln*[18]	3.4.24.27	$M_{120},E_{143},L_{144},Y_{157},H_{231}$	61

2. Related Work

Functional annotation through partial structural alignment has been approached by a number of algorithms comparing a variety of structural and chemical properties. These include atom coordinates and amino-acid labels [30, 5, 13, 25, 33, 34], cavity comparison [8, 11], and graph-theoretic [4] approaches. To reduce the computational cost of search and comparison of 3D structures, many atom coordinate-based partial structural alignment methods employ geometric thresholds. For example, to discard matches with so little geometric similarity that they are biologically irrelevant, PINTS uses geometric thresholds of 7.5 Å between C_α atoms and 6.5 Å between C_β atoms [33], by default Match Augmentation constrains C_α distances to ≤ 7 Å [13], and JESS relies upon motif-dependent empirical thresholds [5].

Existing methods A common metric for measuring geometric similarity between a motif and a target substructure is the Least Root Mean Squared Deviation (LRMSD) measured in angstroms and defined as the minimum RMSD over all possible motif-target substructure alignments. It has been noted, however, that geometric matches alone are not sufficient to infer functional similarity, partially because the RMSD score is affected by the number of points in the motif, as well as individual motif geometry [5, 13, 33]. Furthermore, because different motifs have different frequencies of appearance in proteins, as well as in common protein structures, such as the α-helix, determining a single RMSD threshold, applicable to all motifs, that is capable of separating matches in functionally similar proteins from matches in functionally dissimilar proteins is difficult [5] . To address this problem, several statistical models have been developed to establish RMSD thresholds based on statistical significance of matches on a per-motif basis, by accounting for a motifs distinct 3D geometry [5, 13, 33].

For example, JESS [28] assesses the statistical significance of individual matches by comparing them with a reference population of proteins obtained from CATH [27], a multi-level nested categorization of increasingly specific sequence and structure classifications. Matches to every protein in this reference population are computed and a parametric model based on a mixture of Gaussian distributions is used to determine how unusual, or statistically significant, any given match is.

The PINTS method uses a sequentially non-redundant version of Protein Data Bank (NRPDB) as a reference pop-

978-1-4244-2890-8/08/$25.00 ©2008 IEEE

ulation, but imposes strict comparison thresholds to significantly reduce computing time. PINTS employs a parametric model based on extreme value theory to model the left tail of the match distribution and uses RMSD between motif and matches in the NRPDB to estimate thresholds for each motif.

In contrast to what has been done before, this work does not develop a model of protein sub-structural similarity, but rather a statistical correction procedure to be applied to existing approaches in order to eliminate the dependence of the statistical model on geometric threshold parameters. While the effect of our procedure is demonstrated using a specific structural comparison method (Match Augmentation), presented in detail in the methods section, it is our intention to develop a model that can be applicable to any structural comparison approach.

Match augmentation method In our earlier work [13, 11], motifs designed using only C_α atoms were able to successfully identify functionally similar proteins [12, 21, 20]. Motifs employed by the Match Augmentation method [13] are defined as sets $S = \{S_1, ..., S_{|S|}\}$ of $|S|$ points in 3D space, whose coordinates are taken from backbone C_α atoms (Fig. 1). In order to include into consideration residue substitutions, such as substitutions which have been tolerated over the course of evolution [22, 23], each S_i is assigned a set of possible alternate residue *labels* l_{S_i} = GLY, ALA, PRO,.... Defining the motifs in this fashion increases the sensitivity of the approach by performing searches for many closely related (chemically and/or evolutionarily) functional sites at once, at the cost of increased computing time.

In order to identify potential matches for motif S in a target protein, the target also needs to be represented as a set of $|T|$ points $T = \{t_1, ..., t_{|T|}\}$, where each t_i stands for C_α coordinates, and the label set $l(t_i)$ contains only one amino acid, l_{t_i}. A bijection of corresponding motif points in S to a subset of points in T is defined as: $m = \{(s_{a_1}, t_{b_1}), (s_{a_2}, t_{b_2}), ..., (s_{a_{|S|}}, t_{b_{|S|}})\}$ To consider the bijection a match, it must meet two criteria:

Criterion 1: $\forall i$, s_{a_i} and t_{b_i} are label compatible: $l_{t_{b_i}} \in l(s_{a_i})$;

Criterion 2: $\forall i$, $||A(s_{a_i}) - t_{b_i}|| < \varepsilon$, threshold for geometric similarity, where $||A(s_{a_i}) - t_{b_i}||$ stands for Euclidean distance between s_{a_i} and t_{b_i} where motif S is in least root mean square deviation (LRMSD) alignment with a subset of target points, via a rigid transformation A.

The first criterion simply requires all labels (amino acids) in the target to match amino acids in corresponding motif labels lists $l(s_{a_i})$. The second criterion rejects low quality matches without completing full motif-target alignment, by applying the threshold criteria to less computationally expensive partial motif-target alignment(s). The parameter, ε, is the geometric threshold used to discard potential matches with pair-wise distances greater than what is deemed biologically or algorithmically acceptable. However, excluding a significant number of matches from consideration can introduce bias in the statistical model. In general, ε is set empirically and varies in different partial structural alignment methods. For more details see [13].

Statistical hypothesis testing For each motif, the LRMSD cutoff separating statistically significant (structurally similar) matches from the rest of the matches depends on the frequency of a motif's appearance in target proteins, the number of points comprising the motif, and bias caused by incomplete knowledge of all protein structures. The statistical model can be formulated in the terms of a hypothesis test, where the Null Hypothesis (H_0) asserts that *the motif and matched substructure are not structurally similar*, while the Alternative Hypothesis (H_A) states that *the motif is structurally similar or identical to the matched substructure*. One of the advantages of this approach is that it represents the statistical importance of each match as a *p*-value.

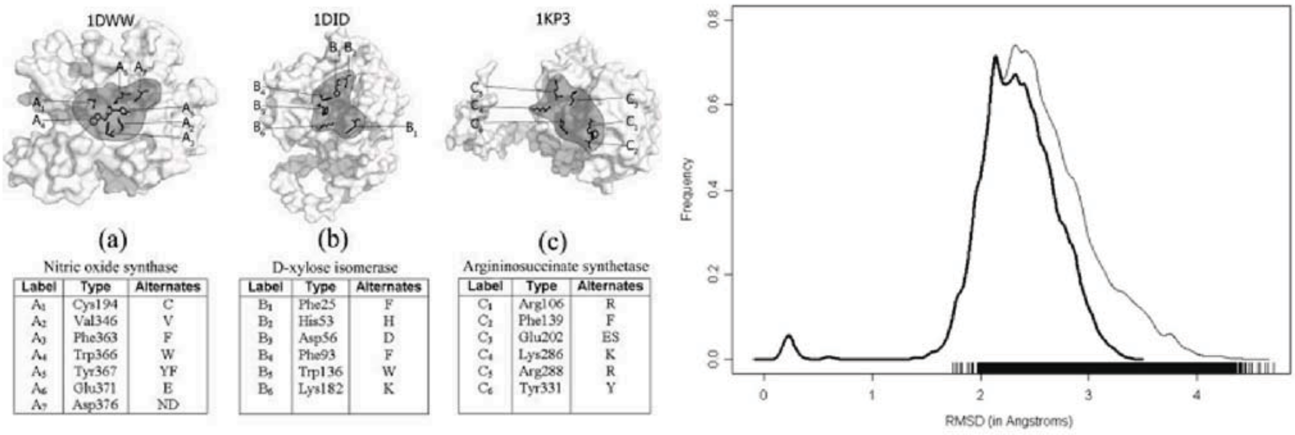
In order to test the Null Hypothesis, which is technically equivalent to computing the *p*-value, one needs to know the distribution of matches (f) of a given motif with respect to *all possible* targets, which can be approximated by the distribution of matches in the representative set of target structures, such as Protein Databank [6], sequentially non-redundant versions of PDB [6], or different levels of CATH fold classification database [27]. Because such distributions for different motifs are known to have a wide range of shapes, in our model we reconstruct them using non-parametric Kernel Density Estimation techniques that utilize a Gaussian kernel with bandwidth chosen by the Sheather-Jones method[19].

3. Methods

Proposed correction model The distribution of matches that pass the search criteria depends on the geometric threshold, ε, defined earlier. However, ideally, the statistical significance score of a match should not reflect the parameters of the geometric comparison algorithm, as it makes the sensitivity and specificity of the method highly dependent on parameters used. To eliminate this dependence, we outline in this section a correction that incorporates eliminated matches into the existing statistical model of sub-structural similarity.

The proposed correction is based on the observation that increasing the geometric threshold parameter (ε) leads to non-random appearances of new (previously discarded) matches. Because these matches were eliminated due to failing geometric thresholds (criterion 2), they tend to appear in the right tail of the LRMSD distribution (Fig. 1). Because we are only interested in matches within the left tail of the distribution, we can represent matches eliminated by geometric thresholds as a point-weight (pwt) at ∞. In the terms of the cumulative distribution function, $\hat{F}_h(x)$, of RMSD distance X, we can therefore write:

$$\hat{F}_h(x) = (1 - pwt) \int_0^x \hat{f}_h(w) dw; x \geq 0 \qquad (1)$$



(a)

Nitric oxide synthase

Label	Type	Alternates
A_1	Cys194	C
A_2	Val346	V
A_3	Phe363	F
A_4	Trp366	W
A_5	Tyr367	YF
A_6	Glu371	E
A_7	Asp376	ND

(b)

D-xylose isomerase

Label	Type	Alternates
B_1	Phe25	F
B_2	His53	H
B_3	Asp56	D
B_4	Phe93	F
B_5	Trp136	W
B_6	Lys182	K

(c)

Argininosuccinate synthetase

Label	Type	Alternates
C_1	Arg106	R
C_2	Phe139	F
C_3	Glu202	ES
C_4	Lys286	K
C_5	Arg288	R
C_6	Tyr331	Y

Figure 1. The left figure details active site motifs from (a) Nitric oxide synthase, PDB ID: 1dww, (b) D-xylose isomerase, PDB ID: 1did, and (c) Argininosuccinate synthetase, PDB ID: 1kp3. Motif coordinates are taken from the backbone C_α atom of the highlighted residues. The right figure shows distributions of matches for a motif derived from 3lzt protein for ε = 4Å (dark line) and ε = 7Å (light line). Vertical hashes (bottom of graph) represent the LRMSD scores of matches eliminated with ε = 4Å but found using relaxed geometric thresholds of 7Å .

where, $\hat{f}_h(x)$ is the Gaussian kernel density estimate of the distribution of geometric matches, and h is the optimal bandwidth of the kernel chosen by the Sheather Jones method [19, 32]. The distribution function $\hat{F}_h(x)$ is *defective* in the sense that $\hat{F}_h(\infty) = (1 - pwt) < 1$. The magnitude of the defect is exactly equal to the point-weight, pwt. Furthermore, it is possible to show (see Appendix) that in the lower RMSD range, which contains the most biologically relevant matches, the p-values obtained using the corrected model will be exact for matches with RMSD $< \frac{\varepsilon}{\sqrt{N}}$, where N is the number of points comprising the motif. This range covers most biologically relevant matches, making pwt correction a very attractive choice.

Estimating model parameters The point-weight correction, outlined in the previous section, models matches eliminated due to geometric thresholds as a point-weight at ∞. These are the matches that would have been identified if the geometric thresholds were relaxed. Thus it is important to separate target proteins capable of producing a match from targets in which no match is possible due to incompatibility of the set of target labels to any set of labels in the motif. We employ an algorithm based on the Halls marriage theorem [17] to identify and exclude target proteins that do not have appropriate labels *a priori*. This ensures that the point-weight represents the proportion of matches eliminated due to geometric constraints.

If the reference protein population can be exhaustively sampled, the point-weight is computed as: $pwt = \frac{N_p}{N}$ where N is total number of proteins in the reference population and N_p is the number of eliminated matches. In practice, only N is known and N_p is obtained by subtracting from N the number of targets in which a match was found. Furthermore, sampling the entire reference population is not necessary and the point-weight can be estimated using the maximum likelihood estimation (MLE) paradigm.

To accurately estimate the distribution of matches, the reference population must be sampled until a specified number of geometric matches, G_s, has been obtained. The likelihood function based on this sampling process is defined as:

$$l(N_p, G_p | G_s, N_s) = \frac{\frac{(G_p)!}{(G_p - G_s)!} \frac{(N_p)!}{(N_p - N_s)!}}{\frac{(N)!}{(N - G_s - N_s)!}} \qquad (2)$$

where G_p, N_p are the unknown total numbers of geometric and eliminated matches in the reference population and G_s, $N_s$are the known counts of geometric and eliminated matches in the sample. Under the constraint that $G_p + N_p = N$, with N known, we can maximize (2) to obtain that the MLE estimate of N_p is the integer satisfying:

$$L_b = \frac{N_s}{N_s + G_s} N - \frac{G_s}{N_s + G_s} < \hat{N}_p < \frac{N_s}{N_s + G_s} N + \frac{N_s}{N_s + G_s} = U_b$$

$$\hat{N}_p = ceiling(L_b) = floor(U_b),$$

with appropriate modifications in the case when L_b and U_b are integers themselves. Thus, the final maximum likelihood estimate of the point-weight is given by:

$$pwt = \frac{floor(U_b)}{N}$$

The maximum likelihood estimation approach provides an asymptotically unbiased estimator of the point-weight that can be computed using quantities already determined in the course of the sampling process.

978-1-4244-2890-8/08/$25.00 ©2008 IEEE

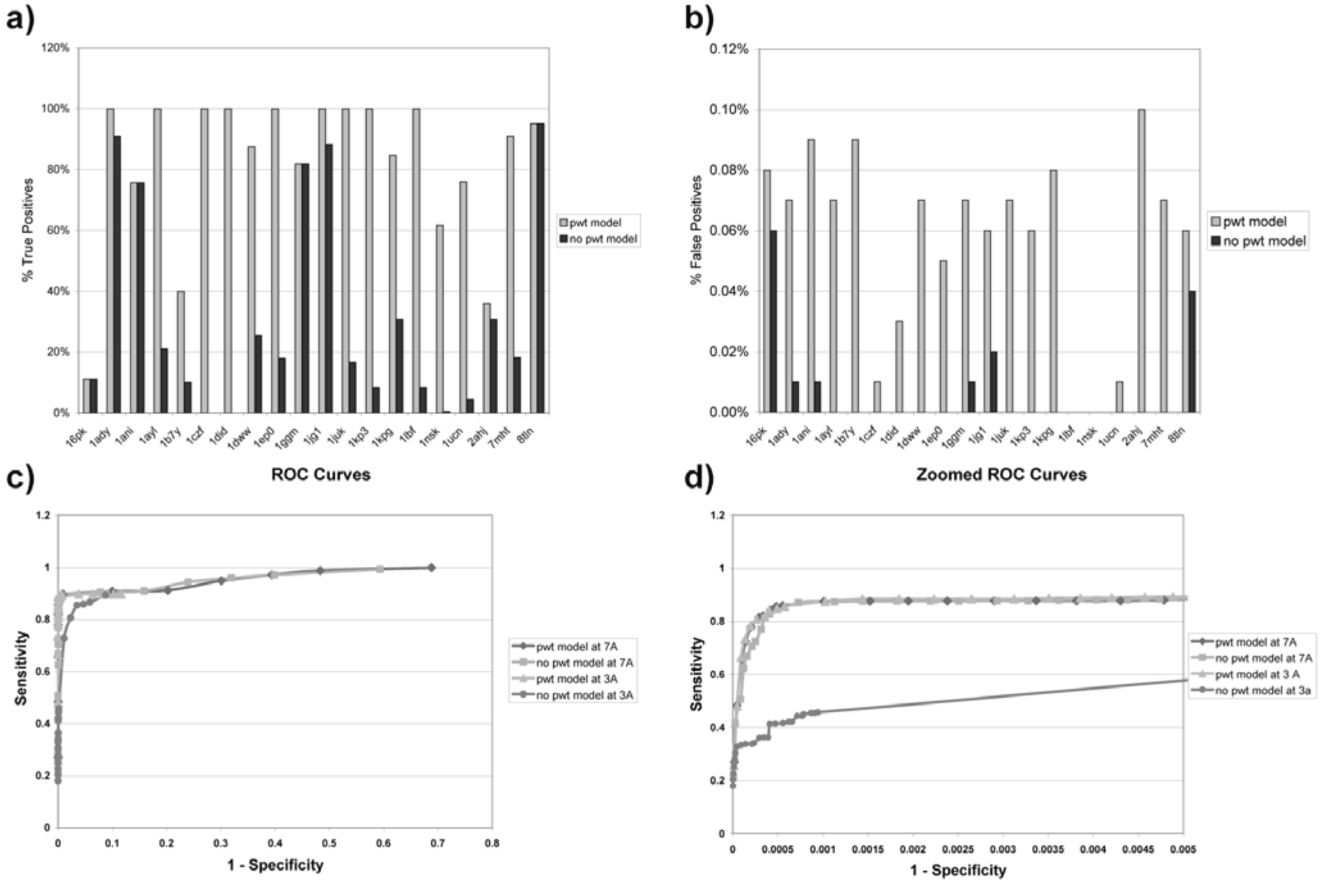


Figure 2. Effect of the point-weight correction on the sensitivity and specificity, expressed in terms of the proportions of True Positives (a) and False Positives (b) identified by the Match Augmentation method with geometric thresholds, ε, set at 3Å . The ROC curves based on sensitivity and specificity values averaged over all 20 motifs (c), with a more detailed look (d).

4. Results

First, we will demonstrate that geometric thresholds can significantly impact the sensitivity of the functional annotation method. Next, we show that the point-weight correction removes systematic bias and greatly reduces misclassification rates of functionally related proteins, without greatly impacting the methods specificity. Finally, we examine in detail misclassification rates of 3 cases where neither the original nor corrected model was able to improve performance.

Dataset The dataset used to illustrate effects of our model consists of 20 motifs representing a range of distinct active sites in non-mutated protein structures. The Enzyme Commission (EC) functional classification [26] was used to identify proteins homologous to each motif (Table 1). Amino acids used to define motifs were selected by choosing those with function documented (see Table 1) and predicted [13, 11] using the Evolutionary Trace Method [11, 23, 24].

Experimental results We computed the distribution of matches at geometric thresholds of ε = 3, 4, 5, 6, and 7 Å using 95% sequence identity filtered version of the Protein Data Bank (as of 12/22/2007) as a reference population. We chose to consider a target match statistically significant, and thus potentially homologous, if its p-value was less than 0.001. For each of the 20 motifs used, we define a match as a True Positive (TP), if it is both statistically significant under our model and it matches one of the members of the EC class corresponding to this motif. A match is considered a False Positive (FP) if it is statistically significant, but it is not among the members of the EC class.

We have observed that the effects of the point-weight correction, in terms of proportions of both TPs and FPs, are strongest when geometric thresholds are strict (Table 2). At 3 Å the corrected model exhibited improvements in sensitivity for 16 out of 20 motifs (Fig. 2a), with 9 motifs achieving 100% sensitivity even at a very restrictive p-value cutoff of 0.001. The loss of specificity associated with the point-weight correction (Fig. 2b) was not significant in comparison with the improvement in sensitivity. In fact, the corrected model, on average, identified 36 more TP

Table 2. Results of computational experiments. Effect of the point-weight correction on average proportions of True Positive and False Positive Matches for various levels of geometric thresholds.

Threshold	% TP matches		% FP matches	
ε	*No PWT*	*PWT*	*No PWT*	*PWT*
3Å	31.76%	82.00%	0.01%	0.06%
4Å	42.75%	82.00%	0.02%	0.05%
5Å	57.99%	82.29%	0.03%	0.06%
6Å	64.93%	82.29%	0.04%	0.06%
7Å	77.66%	82.29%	0.04%	0.06%

matches, corresponding to roughly 50% increase in sensitivity of the method, while introducing on average only 7 FPs, corresponding to 0.05% drop in specificity. The effect of the correction decreases as the geometric thresholds are relaxed, which is due to fewer potential matches being eliminated by the geometric comparison algorithm. In general, the point-weight corrected model outcompetes the non-corrected models even if they are based on more relaxed geometric thresholds (Fig. 2c and 2d).

Our statistical correction allows us to achieve sensitivity and specificity previously possible only for large geometric thresholds, while simultaneously reducing computation time. On the average, running Match Augmentation at 3 Å resulted in a 46% reduction in computing time when compared to 7 Å with little effect on prediction quality. Motifs defined with more points (residues) and higher number of alternate labels received the most benefit: as much as 70% reduction in run time for some cases (1gj1, 16pk, 8tln).

Evaluating sensitivity On the average, for the 20 motifs used in this work, our model exhibited 82% sensitivity. However, in 3 cases, regardless of thresholds used in our approach, the proportion of True Positives was particularly low. In order to investigate these cases, we performed clustering based on RMSD distances between motifs generated from every member of EC class corresponding to entries 6.1.1.20 (1b7y motif), 4.2.1.84 (2ahj motif), and 2.7.2.3 (16pk motif), as presented in Fig. 3a, Fig. 3b, and Fig. 3c respectively. The 1b7y is a dimer consisting of non-symmetric chains A and B, both of which are assigned the same EC classification. The motif chosen as an example in this paper was generated from the active site chain A, and because of this our model was unable to identify homologs of chain B (boxed in Fig. 3b). Similarly, members of the 4.2.1.84 EC class (2ahj motif) cluster into 4 distinct categories, making it difficult to identify a motif capable of representing the entire class, implying either a sub-optimal motif design, possible conformational changes in the active site, or that several structurally different active sites comprise the EC class. No such pattern was visible for the 16pk motif and we presume that low sensitivity there was primarily caused by incorrect choice of motif residues.

5. Conclusions

Partial structural alignment methods, such as JESS, PINTS, and Match Augmentation, are promising computational approaches to the problem of functional annotation of proteins. However, the geometric thresholds used by these approaches to reject biologically irrelevant matches and thus maintain practical efficiency may introduce a systematic bias into the statistical models of sub-structural similarity, which in turn may result in misclassification of functionally similar proteins.

Using a test dataset of 20 structural motifs representing a wide range of distinct active sites, we have demonstrated that the bias introduced by geometric thresholds has a substantial impact on the number of correctly identified homologous proteins. The proposed point-weight correction improved sensitivity, without significantly reducing specificity. In addition, our results suggest that the point-weight correction may make more restrictive levels of geometric thresholds useful without a reduction in accuracy, thus reducing the computing cost of partial structural comparison approaches by an average of 46% and by as much as 70% for some of the motifs used in this work. Eliminating biologically irrelevant matches via geometric thresholds to achieve accelerated runtimes is an effective practice for several existing structural comparison methods. For Match Augmentation, and potentially for other methods utilizing similar thresholds, our statistical correction provides additional sensitivity and the potential for further gains in computational efficiency.

Acknowledgements This work is supported by a grant from the National Science Foundation NSF DBI-0318415 and NSF DBI-054795 to OL and LK. Additional support is gratefully acknowledged from training fellowships the Gulf Coast Consortia (NLM Grant No. 5T15LM07093) to B.C.; from NIH GM079656, GM066099, and March of Dimes Grant FY03-93 to O.L.; from a Whitaker Biomedical Engineering Grant and a Sloan Fellowship to L.K; and from an affiliation agreement between Rice University and Texas Childrens Hospital to V.F. D.B. was supported by the Brown School of Engineering at Rice University. Experiments were run on equipment funded by AMD and EIA-0216467. This work was supported in part by the Rice Terascale Cluster funded by NSF under Grant EIA-0216467, Intel, and HP.

Appendix Let X be the RMSD of a match between a motif consisting of N points and a substructure within a target protein, and let Y be the maximum pair-wise distance between the matched residues. The joint probability density function of X and Y variables is given by $(X,Y) \sim \phi(x,y)$. The marginal distribution of X is $X \sim f_x(s) = \int_{\mathbb{R}^+} \phi(x,y)dy$, and the marginal distribution of Y is given by $Y \sim g_Y(y) =$

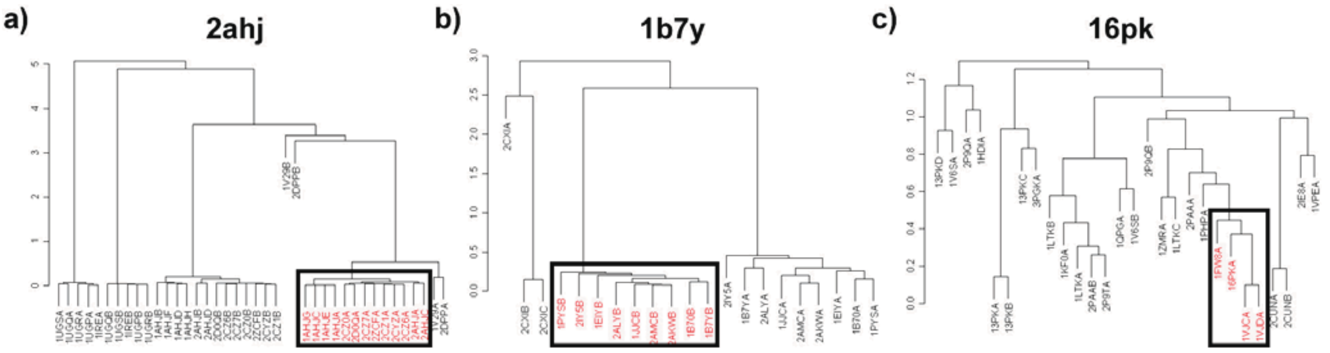


Figure 3. Clustering based on RMSD distances between motifs generated from every member of EC class corresponding to entries 6.1.1.20 (1b7y motif), 4.2.1.84 (2ahj motif), and 2.7.2.3 (16pk motif), A, B, and C respectively. Motifs derived from 2ahj and 1b7y were able to identify only the members of their immediate clusters (boxed), implying either a sub-optimal motif design or that several structurally different active sites comprise the EC class. No such pattern was visible for the 16pk motif and we presume that low sensitivity there is primarily caused by sub-optimal motif design.

$\int_{\mathbb{R}^+} \phi(x,y)dx$. The distribution we ideally would like to employ is $f_X(x)$. However, due to matches eliminated in the course of partial structural alignment, what we observe is the distribution of X conditional on the event $[Y < \varepsilon]$. The distribution $f_{X|Y<\varepsilon}(x)$, can be derived as follows:

$$X|Y < \varepsilon \sim f_{X|Y<\varepsilon}(x) = \frac{d}{dx}\left[\frac{\int_0^x \int_0^\varepsilon \phi(\xi,\eta)d\eta d\xi}{\int_0^\infty \int_0^\varepsilon \phi(\xi,\eta)d\eta d\xi}\right].$$

Note that the denominator $\int_0^\infty \int_0^\varepsilon \phi(\xi,\eta)d\eta d\xi = \int_0^\varepsilon g(\eta)d\eta$ integrates to $1 - pwt_\varepsilon$, whereas the numerator $\frac{d}{dx}\int_0^x \int_0^\varepsilon \phi(\xi,\eta)d\eta d\xi = \int_0^\varepsilon \phi(x,\eta)d\eta$. In other words,

$$X|Y < \varepsilon \sim f_{x|y<\varepsilon}(x) = \frac{\int_0^\varepsilon \phi(x,y)dy}{1 - pwt_\varepsilon}.$$

The proposed corrected distribution $f_\varepsilon(x)$ is given by: $f_\varepsilon(x) = f_{X|Y<\varepsilon}(x) \times (1 - pwt_\varepsilon) + \delta_\infty(x)pwt_\varepsilon$, where $\delta_\infty(x)pwt_\varepsilon$ includes matches eliminated by the partial structural alignment with threshold ε as a point-mass at ∞.

When the corrected distribution is placed into the non-parametric density estimation setting we get:

$$\hat{f}_\varepsilon(x) = \hat{f}^h_{X|Y<\varepsilon}(x) \times (1 - pwt_\varepsilon) + \delta_\infty(x)pwt_\varepsilon$$

where $\hat{f}^h_{X|Y<\varepsilon}(x)$ is the Gaussian kernel density estimate of the distribution of geometric matches, the optimal bandwidth kernel (h) chosen by the Sheather-Jones method [19, 32].

Because the correction proposed above shifts the eliminated matches to point-weight at ∞, for some x 1 the corrected and desired distributions of X will not coincide. In other words: $F_\varepsilon(x) \neq F_X(x), \forall x$. To establish the range of x values for which the corrected model is exact, consider the following pair of inequalities:

$$\frac{1}{\sqrt{N}}\left[\max_{1\le i\le N}(\lambda_i)\right] \le \left(\frac{1}{N}\sum_{i=1}^{N}(\lambda_i)\right)^{\frac{1}{2}} \le \max_{1\le i\le N}(\lambda_i) \qquad (3)$$

OR

$$\frac{Y}{\sqrt{N}} \le X \le Y,$$

where λ_i's are the pair-wise distances between matched points.

Let us assume $x < \frac{\varepsilon}{\sqrt{N}}$. Now let us consider the event $\{X \le x\}$. Under the assumption of $x < \frac{\varepsilon}{\sqrt{N}}$, we get: $\{X \le x\} \subset \left\{X < \frac{\varepsilon}{\sqrt{N}}\right\}$.

However, from (3) we know that $Y \le X\sqrt{N}$, which under our assumption implies $Y < \frac{\varepsilon}{\sqrt{N}}\sqrt{N} = \varepsilon$. This implies that $\{X \le x\} \subset \{Y < \varepsilon\}$. Therefore also: $F_\varepsilon(x) = P(X \le x, Y < \varepsilon) = P(X \le x) = F_X(x)$, when $x < \frac{\varepsilon}{\sqrt{N}}$.

This means that in the lower RMSD range, which contains the most biologically relevant matches, the p-values obtained using the corrected model will be exact for matches with RMSD $< \frac{\varepsilon}{\sqrt{N}}$. For example, when considering matches to a 6-point motif with an ε cutoff of 3Å our p-values are exact for all matches < 1.225Å .

References

[1] A. Aberg, A. Yaremchuk, M. Tukalo, B. Rasmussen, and S. Cusack. Crystal structure analysis of the activation of histidine by Thermus thermophilus histidyl-tRNA synthetase. *Biochemistry*, 36(11):3084–3094, 1997.

[2] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *J. Mol. Biol*, 215(3):403–410, 1990.

[3] J. Arnez, A. Dock-Bregeon, and D. Moras. Glycyl-tRNA synthetase uses a negatively charged pit for specific recognition and activation of glycine. *Journal of Molecular Biology*, 286(5):1449–1459, 1999.

[4] P. Artymiuk, A. Poirrette, H. Grindley, D. Rice, and P. Willett. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J Mol Biol*, 243(2):327–44, 1994.

[5] J. Barker and J. Thornton. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis, 2003.

[6] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *logo*, 58(1 Part 6):899–907.

[7] B. Bernstein, P. Michels, and W. Hol. Synergistic effects of substrate-induced conformational changes in phosphoglycerate kinase activation. *Nature*, 385(6613):275–278, 1997.

[8] T. Binkowski, P. Freeman, and J. Liang. pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic Acids Research*, 32(Web Server Issue):W555, 2004.

[9] S. Brenner. A tour of structural genomics. *Nature Reviews Genetics*, 2(10):801–809, 2001.

[10] D. Brown, N. Kjrishnamurthy, J. Dale, W. Christopher, and K. SJöLander. Subfamily HMMs in Functional Genomics. *Biocomputing 2005: Proceedings of the Pacific Symposium, Hawaii, USA 4-8 January 2005*, 2004.

[11] B. Chen, V. Fofanov, D. Bryant, B. Dodson, D. Kristensen, A. Lisewski, M. Kimmel, O. Lichtarge, and L. Kavraki. Geometric Sieving: Automated Distributed Optimization of 3D Motifs for Protein Function Prediction. *Lecture Notes in Computer Science*, 3909:500, 2006.

[12] B. Chen, V. Fofanov, D. Bryant, B. Dodson, D. Kristensen, A. Lisewski, M. Kimmel, O. Lichtarge, and L. Kavraki. The MASH Pipeline for Protein Function Prediction and an Algorithm for the Geometric Refinement of 3D Motifs. *Journal of Computational Biology*, 14(6):791–816, 2007.

[13] B. Chen, V. Fofanov, D. Kristensen, M. Kimmel, O. Lichtarge, and L. Kavraki. Algorithms for structural comparison and statistical analysis of 3D protein motifs. *Pac Symp Biocomput*, 334:45, 2005.

[14] J. Coleman. Structure and Mechanism of Alkaline Phosphatase. *Annual Reviews in Biophysics and Biomolecular Structure*, 21(1):441–483, 1992.

[15] C. Collyer, K. Henrick, and D. Blow. Mechanism for aldose-ketose interconversion by D-xylose isomerase involving ring opening followed by a 1, 2-hydride shift. *J Mol Biol*, 212(1):211–35, 1990.

[16] B. Crane, A. Arvai, S. Ghosh, E. Getzoff, D. Stuehr, and J. Tainer. Structures of the N-hydroxy-L-arginine complex of inducible nitric oxide synthase oxygenase dimer with active and inactive pterins. *Biochemistry*, 39(16):4608–4621, 2000.

[17] R. Diestel. Graph Theory, volume 173 of Graduate Texts in Mathematics. *Springer, Heidelberg*, 91:92, 2005.

[18] D. Holland, A. Hausrath, D. Juers, and B. Matthews. Structural analysis of zinc substitutions in the active site of thermolysin. *Protein Science*, 4(10):1955, 1995.

[19] M. Jones, J. Marron, and S. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996.

[20] D. Kristensen, B. Chen, V. Fofanov, R. Ward, A. Lisewski, M. Kimmel, L. Kavraki, and O. Lichtarge. Recurrent use of evolutionary importance for functional annotation of proteins based on local structural similarity. *Protein Science*, 15(6):1530–1536, 2006.

[21] D. Kristensen, R. Ward, A. Lisewski, S. Erdin, B. Chen, V. Fofanov, M. Kimmel, L. Kavraki, and O. Lichtarge. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics*, 9:17, 2008.

[22] O. Lichtarge, H. Bourne, and F. Cohen. An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. *Journal of Molecular Biology*, 257(2):342–358, 1996.

[23] O. Lichtarge and M. Sowa. Evolutionary predictions of binding surfaces and interactions. *Current Opinion in Structural Biology*, 12(1):21–27, 2002.

[24] I. Mihalek, I. Reš, and O. Lichtarge. A Family of Evolution–Entropy Hybrid Methods for Ranking Protein Residues by Importance. *Journal of Molecular Biology*, 336(5):1265–1282, 2004.

[25] M. Moll and L. Kavraki. LabelHash: A Flexible and Extensible Method for Matching Structural Motifs. *In the Proceedings of the 2008 IEEE Conference on Computational Systems Bioinformatics*, 2008.

[26] E. Nomenclature. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, NC-IUBMB, 1992.

[27] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[28] C. Porter, G. Bartlett, and J. Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32:D129, 2004.

[29] L. Reshetnikova, N. Moor, O. Lavrik, and D. Vassylyev. Crystal structures of phenylalanyl-tRNA synthetase complexed with phenylalanine and a phenylalanyl-adenylate analogue. *Journal of Molecular Biology*, 287(3):555–568, 1999.

[30] R. Russell. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *Journal of Molecular Biology*, 279(5):1211–1227, 1998.

[31] M. Shatsky. *The common point set problem with applications to protein structure analysis*. PhD thesis, Tel Aviv University, School of Computer Science, 2006.

[32] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

[33] A. Stark, S. Sunyaev, and R. Russell. A Model for Statistical Significance of Local Similarities in Structure. *Journal of Molecular Biology*, 326(5):1307–1316, 2003.

[34] J. Watson, R. Laskowski, and J. Thornton. Predicting protein function from sequence and structural data. *Current Opinion in Structural Biology*, 15(3):275–284, 2005.

[35] H. Wolfson and I. Rigoutsos. Geometric Hashing: An Overview. *IEEE Computational Science & Engineering*, pages 10–21, 1997.

[36] C. Zhang and S. Kim. Overview of structural genomics: from structure to function. *Current Opinion in Chemical Biology*, 7(1):28–32, 2003.