JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 9, 2006
© Mary Ann Liebert, Inc.
Pp. 1565–1573

Fold Recognition via a Tree

YU CHEN[1] and GORDON M. CRIPPEN[2]

ABSTRACT

Recently, we developed a pairwise structural alignment algorithm using realistic structural and environmental information (SAUCE). In this paper, we at first present an automatic fold hierarchical classification based on SAUCE alignments. This classification enables us to build a fold tree containing different levels of multiple structural profiles. Then a tree-based fold search algorithm is described. We applied this method to a group of structures with sequence identity less than 35% and did a series of leave one out tests. These tests are approximately comparable to fold recognition tests on superfamily level. Results show that fold recognition via a fold tree can be faster and better at detecting distant homologues than classic fold recognition methods.

Key words: protein structure classification, multiple alignment.

1. INTRODUCTION

With the progress in structural genomics, the number of protein structures is increasing rapidly. The ongoing Protein Structure Initiatives (PSI) has set its ultimate goal to make structural annotations available for almost every protein sequence (Norvell and Machalek, 2000). Since experimentally determining a three-dimensional (3D) structure is still far more expensive compared to sequencing, much of the structural annotations have to rely on computational work. If all of the unique folds are known, then the structure determination problem can be re-formulated into the fold recognition problem, i.e., finding the most suitable 3D-fold for a given protein sequence.

Machine learning methods including neural network (NN) and support vector machine (SVM) have been widely used in the fold recognition field (Ding and Dubchak, 2001; Han et al., 2005; Xu et al., 2003; Jones, 1999). Those methods improve fold recognition performance greatly by extracting sequence- or structure-based features to construct more comprehensive scores to measure the overall similarity between sequences and structures. However, machine learning methods do not solve the fundamental alignment problem: they either bypass the alignment procedure completely (Ding and Dubchak, 2001) or rely on other methods to generate alignments (Xu et al., 2003; Han et al., 2005).

Structural profiles can be used to bridge 3D structures and 1D sequences and generate alignments. A structure profile is equivalent to a position specific scoring matrix (PSSM), where each position on a 3D structure will have different propensity scores for different amino acids. Usually the propensity score is calculated based on the log-likelihood ratio between the probability of a certain amino acid occupying that

[1]Bioinformatics Program, University of Michigan, Ann Arbor, Michigan.
[2]College of Pharmacy, University of Michigan, Ann Arbor, Michigan.

position versus the probability of that amino acid occurring in nature (Equation (1)).

$$s(AA, pos) = \log \frac{P(AA, pos)}{P(AA)P(pos)} = \log \frac{P(AA|pos)}{P(AA)} \quad (1)$$

There are two approaches to convert a 3D structure into a (single) structural profile. One is the sequence-based approach: after a database query of the sequence of a given structure, a multiple sequence alignment (MSA), including all homologous sequences, can be built. Then a PSSM can be easily calculated based on the MSA. In PSI-BLAST (Altschul et al., 1997), a further database query of the PSSM is used to update the PSSM iteratively. Such sequence based profiles can achieve good performance in fold recognition if sequence identity is high, while their performance drops dramatically when sequence identities are below the twilight zone (20–30%).

Another approach is the structure-based 3D profile approach, which was first introduced by Eisenberg and his colleagues in 1991 (Bowie et al., 1991). Instead of calculating propensity scores based on multiple sequence alignments, they defined a set of environmental states based on structure-derived descriptors such as secondary structure, solvent accessibility, etc. The propensity scores of each amino acid for each environmental state were log-likelihood scores (LLS) calculated based on a survey of known 3D structures (Equation (2)).

$$LLS(AA, ENV) = \log \frac{P(AA, ENV)}{P(AA)P(ENV)} = \log \frac{P(AA|ENV)}{P(AA)} \quad (2)$$

Eisenberg's method is also known as 3D-1D method and the corresponding substitution table containing the above propensity scores is called 3D-1D table. In their initial paper, Bowie et al. (1991) defined 18 environmental states to characterize 3D structures. By incorporating other information such as predicted secondary structure and residue types, the number of environmental states may easily increase to thousands (Mallick et al., 2002).

Instead of defining more environmental states to describe 3D structures, we can use linear combinations of a set of basis environments to create more enriched descriptions. If a multiple structural alignment (MSTA) is available, for each aligned column, the combination of environments of aligned residues may be used to generate a new environment, and the propensity scores of each amino acid in such a hybrid environment can be calculated as:

$$LLS\left(AA, \sum ENV\right) = \log \frac{P\left(AA, \sum ENV\right)}{P(AA)P\left(\sum ENV\right)} = \log \frac{P\left(AA \mid \sum ENV\right)}{P(AA)} \quad (3)$$

In this work, $\sum ENV$ is defined as a combination of equally weighted, non-duplicated, structurally aligned environment states.

In fact, the multiple structural profile approach is not new. A few (but not many) other multiple-structure based fold recognition methods have been reported, including Fugue (Shi et al., 2001), 3D-PSSM (Kelley et al., 2000), and S3 (Zhou and Zhou, 2005).

Fugue (Shi et al., 2001) adopted a very similar way to incorporate MSTA information into fold recognition, where weighted means over single structural profiles for each aligned position were used to build a multiple structural profile (Equation (4)).

$$LLS_m\left(AA, \sum ENV\right) = \sum_i f_i s(AA, ENV_i) = \sum_i f_i \frac{P(AA|ENV_i)}{P(AA)} \quad (4)$$

3D-PSSM (Kelley et al., 2000) also used MSTA information to build multiple structural profiles. However, it differs from other purely structure based fold recognition methods in that structural profiles in 3D-PSSM were derived from PSI-BLAST PSSMs of each single structure's sequence, while structures were only used to generate multiple alignments. Therefore, a 3D-PSSM is actually a concatination of single-sequence based PSSMs.

S3 (Zhou and Zhou, 2005) defined PSSM on fragment level. Each structure was divided into fragments (nonapeptides) and each fragment was used to query the fragment database. Fragment similarities were

measured by both fragment structural similarity and environment similarities, and multiple structure alignments were then generated. Since the alignment is on the fragment level, the number of structures in an MSTA is much larger than for other methods (>225 for each nonapeptide). Therefore those structure-based multiple fragment alignments can be directly used to calculate the PSSMs by Equation (1).

The performance of multiple structural profiles is still being studied. It was reported by the Fugue investigators that profiles generated by multiple structural alignments on family levels improve the recognition performance (Shi et al., 2001). However, further studies showed that multiple structural profiles based on alignments of divergent sequences sometimes may not perform better than single structural profiles (Mizuguchi et al., 2004). Therefore, Fugue3 is currently using a highly redundant fold library containing both single structural profiles and multiple structural profiles.

As the number of available protein folds increases, we think it is time to revisit the multiple structural profile approach. Particularly, we are interested in two questions: the first question is can we make fold recognition faster by using multiple structural profiles? With such a large number of structures, we will have no time to search a large redundant fold library as Fugue3. Secondly, how can we make multiple structural profiles perform better than a cluster of single structural profiles? Intuitively, multiple structural profiles contain more information than single structures. Thus, multiple structural profiles should have better performance. However, if there are too many diverse sequences in a structural cluster, there will be too much noise in the multiple structural profile. If we can reduce the noise, we should be able to get better performances in fold recognition.

A tree-based fold recognition is proposed here because (1) searching along a tree can be fast; and (2 in a tree containing multiple levels of fold clusters, the lower the level is, the less noise will be observed. As a depth-first searching proceeds from root to leaves along the tree, we should observe less and less noise.

One fundamental difference between our tree-based fold recognition approach and other approachs is that we make use of different levels of multiple structural alignments, and each node in the fold tree is represented by a multiple structural profile (in the form of PSSM) rather than a single structural profile based on one representative structure. Since a multiple structural alignment can grasp the common core structure of a whole family, we expect performance improvements over using the single representative structure.

The quality of structural alignments is essential for the multiple-structure-based fold recognition. For the full-length pairwise protein structural alignment, few existing methods ensure environmental similarity in the final alignments. Due to the same reason, although there are several well maintained hierarchical fold classification databases [especially CATH (Orego et al., 1997), SCOP (Murzin et al., 1995), and FSSP (Holm and Sander, 1996)], we can not use them as our fold tree because they are not tailored for fold recognition purposes. In fold recognition, similarities between structures should not only be measured by how structures can be superimposed in 3D space, but also be measured by how corresponding structural profiles agree with each other. It is very important because 3D superpositions will only handle isolated protein chains while neglecting all inter-chain interactions.

Recently, we developed a new pairwise structural alignment method, SAUCE, which can measure both structural and environmental similarities between protein structures and guarantee environmental matching in pairwise structural alignments (Chen and Crippen, 2005). We think SAUCE is especially suitable for the fold recognition problem. An automatic hierarchical clustering of all-to-all SAUCE pairwise alignment results, therefore, has been performed and multiple structural alignments are built at each node of the tree via SAUCE-(IRIS)-TCOFFEE. IRIS is a multiple structural alignment step developed by our group (Chen and Crippen, 2006). The resulting fold trees are used for further recognition experiments.

2. EXPERIMENTS AND RESULTS

2.1. Dataset

A total of 379 chains were selected based on CATH v2.5.1. All of them are single domain proteins with length of 30 to 100 residues. Membrane proteins and proteins containing nonstandard residues were excluded. Each of them belongs to a different S35 family, which means most sequence identities between those chains are less than 35%. Based on CATH v2.5.1, these 379 chains fall into 155 homologous clusters.

2.2. Hierarchical classification of folds

Using SAUCE (Chen and Crippen, 2005), we performed $\frac{379\times378}{2}$ all-to-all pairwise structural comparisons. The greater value between the two SAUCE E-values obtained is used as the distance between two protein structures. A hierarchical clustering using complete linkage was performed using function hclust from R (R Development Core Team, 2005). By choosing a cutoff at 0.1, we partitioned the 379 chains into 154 clusters. Since the complete linkage method was used, all structural pairwise alignments within each cluster should have E-values less than 0.1.

A comparison between our fold classification and CATH classification is shown in Figure 1. We can see our SAUCE classification with cutoff 0.1 is approximately equivalent to homologous families in CATH (or superfamilies in SCOP). We got fewer orphan clusters (clusters containing only one structure) and large clusters, but more medium clusters. Our SAUCE classification in general agrees with the CATH classification except for mostly-alpha structures. One reason is that one alpha helical strand can be superimposed perfectly on any other alpha helical strand, and the environmental differences between those strands are sometimes negligible.

As we can see, SAUCE classification merges/splits CATH families. In one of the most extreme examples, nine CATH homologous families over three CATH topologous families are merged (Fig. 2). In the other case, a CATH homologous family 1.10.10.10 containing 20 structures was split into six SAUCE groups. A further investigation showed that among these 20 structures, some of them are free monomers, some of them are dimers and some of them are complexed with DNA (data not shown). Another example of splitting two different oligomers belonging to the same CATH homologous family has been shown in Chen and Crippen (2005).

2.3. Fold tree

The hierarchical clustering tree can be used as a fold tree. For each node in the tree, a (multiple) structural profile is built. In order to make reliable multiple structural profiles, it is necessary that only highly similar structures are used to build those profiles. Since each of the SAUCE clusters at level 0.1 can be treated as a fold superfamily, we built our fold trees starting from the above 154 SAUCE clusters.

In order to build multiple structural profiles, we need to build multiple structural alignments first. Then multiple structural profiles can be obtained from multiple alignments based on a modified 3D-1D table.

2.4. Multiple structural alignment

Multiple structural alignments are built using TCOFFEE (Notredame et al., 2000). SAUCE alignments have three levels of structural similarity: (a) 3D structurally and environmentally similar, (b) 3D structurally similar only, and (c) environmentally similar only. We used three different weights on these three different types of similarities: 4,000 for (a), 800 for (b), and 100–800 for (c) depending on environmental similarities. We imposed such large weights on structural based alignments that sequence information will generally be disregarded. Then a TCOFEE library is generated for each pairwise structural alignment. We used two gap penalties: open gap penalty -300 and extension penalty -30. It is also possible to use IRIS to refine SAUCE based pairwise alignment libraries before the TCOFFEE assembly step.

2.5. Multiple structural profiles

In the original 3D-1D table obtained (Chen and Crippen, 2005), we have 75 environmental states, each of which has different propensities for different amino acids. Given a multiple structural alignment, aligned environments will be merged to form a new environmental state and the log-likelihood will be recalculated (Equation (3)). For columns containing gaps, pseudo counts based on background frequencies were added. The more gaps in a column, the less specific to amino acids the column will be.

2.6. Threading via the fold tree

We used a simple Smith-Waterman dynamic programming method to align sequences to structural profiles. Given gap opening and extension penalties as 300/30, we found that the raw alignment scores

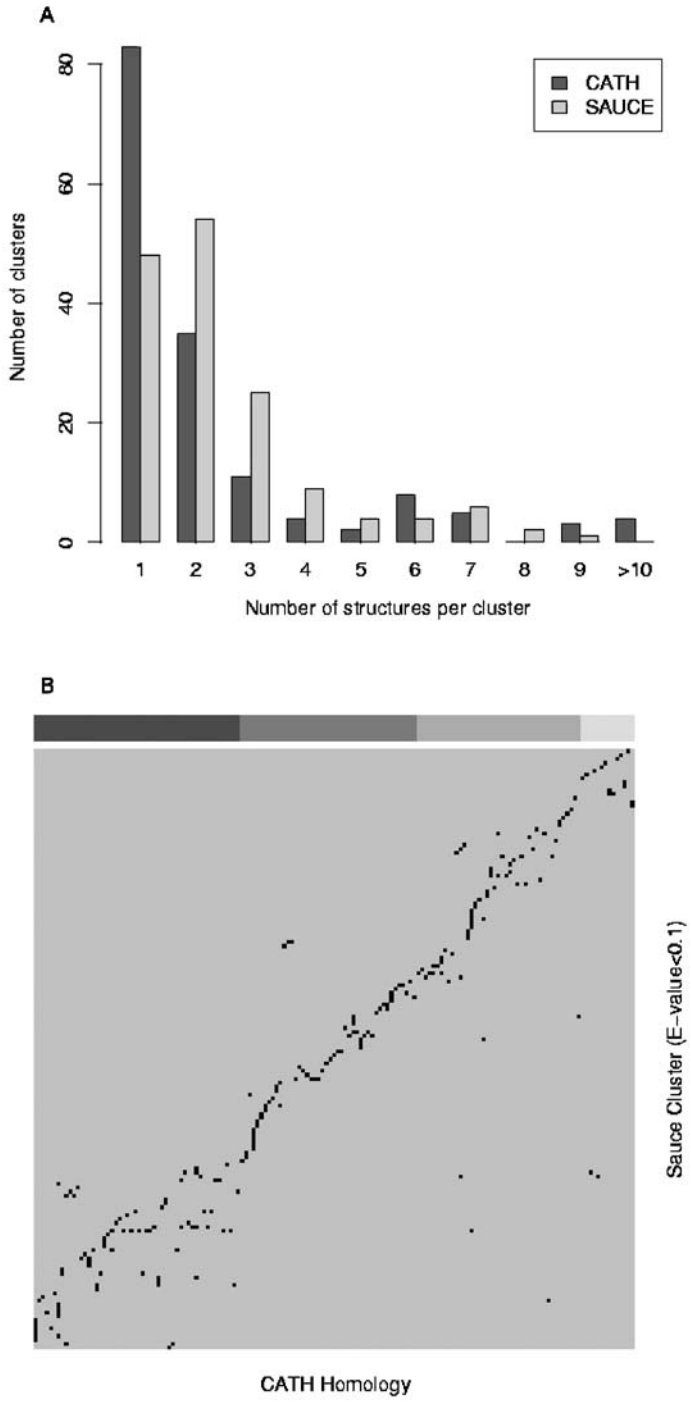

FIG. 1. Comparison between SAUCE classification results and CATH classification. (**A**) Comparison of cluster sizes between SAUCE and CATH v2.5.1 over 379 protein chains. (**B**) Comparison of 154 SAUCE clusters with 155 CATH homologous families. Black dots represent overlapping between corresponding cluster pairs. The bars on the top of the plot represent the four major CATH classes (from left to right: mainly-alpha, mainly-beta, mixed alpha-beta, and few secondary structures).

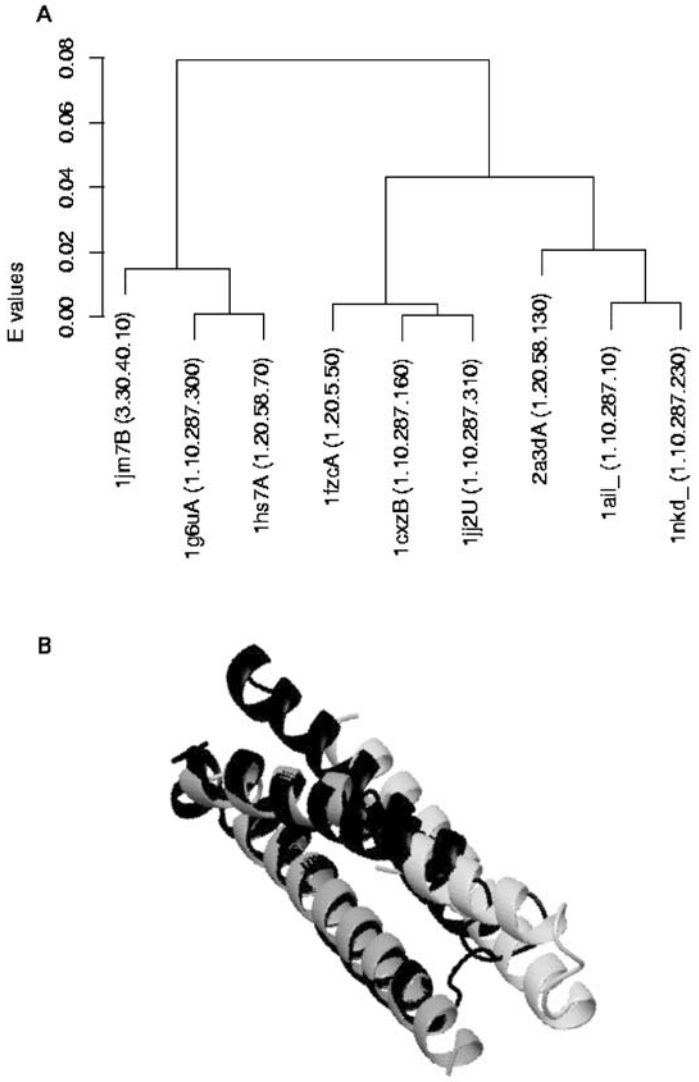

FIG. 2. An example of merging CATH families. (**A**) Hierarchical clustering of nine CATH homologous families/three CATH topologous families by complete linkage method. Similarities between proteins are measured by SAUCE E-values. Leaves are labeled with PDB ids as well as CATH homologous family ids. (**B**) An example of SAUCE pairwise alignment. We found that all these structures contain an environmentally similar helical bundle motif. In this particular example, it is interesting to see that a helical hairpin dimer (1hs7: white/gray) has an environment very similar to that of a three helical bundle monomer (1g6u: black).

follow EVD (similar to SAUCE scores, see Chen and Crippen [2005]) and the EVD parameters can be determined from Equations (5) and (6).

$$\lambda = 0.01610 - 0.00081 \log(MN) \tag{5}$$

$$\mu = -622.3 + 154.1 \log(MN) \tag{6}$$

In order to evaluate whether our fold tree can assist fold recognition, we only used sequences from clusters containing 3 or more structures as queries to search the fold tree containing 154 folds. There are 220 test sequences in such nontrivial clusters out of all 379 proteins used to build the tree. We align each test sequence to its native SAUCE cluster (with and without the native structure) as well as to other nonnative SAUCE clusters. Ranks for native SAUCE clusters were recorded.

Table 1. Comparison of Fold Recognition Performance

Routine ID	Search routines	Max searches	Self recognition		Leave-one-out	
			Top 1	Top 5	Top 1	Top 5
1	Leaves only	379	100.00%	100.00%	22.07%	40.99%
2	Roots only	154	59.46%	70.27%	18.02%	30.18%
3	Leaves and nodes	644	99.10%	99.10%	27.47%	42.79%
4	Nodes only	274	81.98%	90.09%	30.18%	45.94%

Four search routines were used:

(1) *Leaves only*: only use single structural profiles, 379 profiles in total
(2) *Roots only*: only use the highest level multiple structural profiles in the tree, 154 profiles in total
(3) *Leaves and nodes*: a depth-first search is performed from the root (superfamily) to the leaves (single structural profile), 654 profiles altogether in the tree
(4) *Nodes only*: similar to *leaves and nodes* but does not include single structural profiles, totaling 274 profiles in the tree.

Routine 1 is equivalent to conventional (single) fold recognition methods which do not use multiple structural alignment information at all, while routine 3 mostly resembles the idea of redundant fold library in Fugue3, in which both single structural and multiple structural profiles are used.

A depth-first walk was performed to search the tree. Starting from the root, if there was any son node having a more significant E value than the parent, a further search was done in that son's branch. If none of son nodes can perform better than the parent, the search will end at this branch and backtrack.

We did two fold tree tests. The first one is the self recognition test, which included the native structure for the query sequence in the tree. This is a simple test. We want to see whether the native sequence can find the native structure. The other test is the leave-self-out test. We left out the native structure of the query sequence from the tree, rebuilt multiple structural alignments and multiple structural profiles. We want to see whether the sequence can find its native fold superfamily. The second test simulates a real fold recognition test, where the native fold of a sequence is unknown.

The results (Table 1) show that all native structures, if included in the library, are ranked as the best hit by their corresponding sequences (*leaves only* search routine). In most cases, tree walk methods (*leaves and nodes* and *nodes only* routines) will be able to guide the sequence to find the native structure via the tree. The leave-self-out results show that although using only one multiple structural profile per fold cluster is the fastest, the performance is the worst (routine 2). Via our fold depth-first tree searches (routines 3 and 4), the recognition performance is better (correct fold recognition increases around 5%) with a greater speed than conventional threading (routine 1).

3. DISCUSSION

If we only use single structure profiles to do fold recognition, we can see that as the number of single structures in a fold cluster increases, the fold searching space expands, which should result in a better performance in fold recognition (Fig. 3A: P vs. Ns). If we use multiple structural profiles, the situation is more complicated: with more and more sequences added to the multiple structural alignment, the proportion of gaps/misaligned regions may increase. Such noise can sometimes offset the benefits of extra structural information. As we can see from Figure 3B, for small clusters ($Ns<6$), multiple structural profile (nodes only) can lead to an overall better performance than single structural profiles (leaves only). But as the cluster size grows too large, the performance of multiple structural profiles is undermined by noise (probably due to large amount of gaps), and therefore is no better than that of single structural profiles. We also used the number of CATH homologous families (Nc) in a cluster to measure cluster diversity. It seems that the larger the Nc is, the more diverse the sequences will be. As shown in Figure 3, performance of fold recognition drops dramatically for large Nc. The better performance of single structures in large Nc and Ns suggests maybe we need to remodel threading algorithms in the gapped regions.

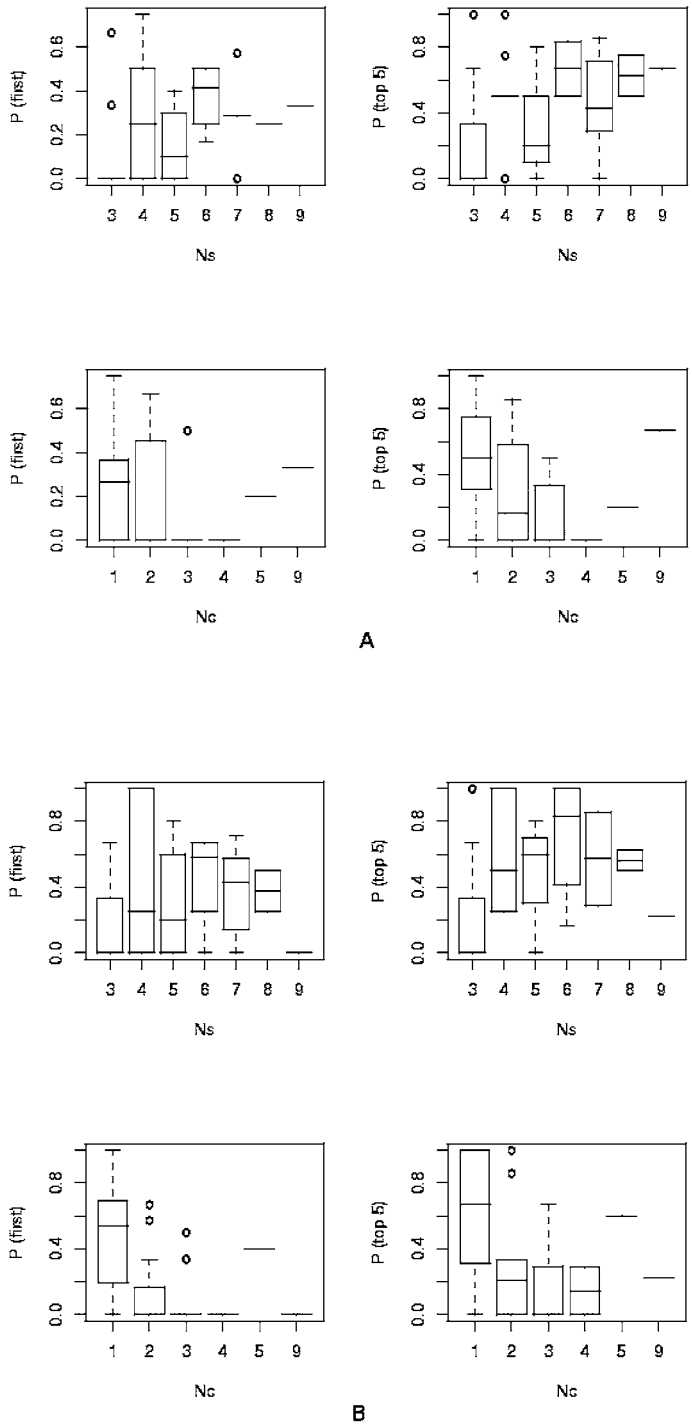

FIG. 3. Relationship between cluster diversity and threading performance. (**A**) Leaves only. (**B**) Nodes only. P (first) is probability that the native cluster has the top rank; P(top 5) is the probability that the native cluster is ranked in the top 5; Ns is the number of structures in the SAUCE cluster; Nc is the number of CATH homologous families in the SAUCE cluster.

ACKNOWLEDGMENT

We thank Dr. David States for computing resources.

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Bowie, J.U., Luthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164–170.

Chen, Y., and Crippen, G.M. 2005. A novel approach to structural alignment using realistic structural and environmental information. *Protein Sci.* 14, 2935–2946.

Chen, Y., and Crippen, G.M. 2006. An iterative refinement algorithm for consistency based multiple structural alignment methods. *Bioinformatics* 22, 2087–2093.

Ding, C.H.Q., and Dubchak, I. 2001. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 17, 349–358.

Han, S., Lee, B., Yu, S., et al. 2005. Fold recognition by combining profile-profile alignment and support vector machine. *Bioinformatics* 21, 2667–2673.

Holm, L., and Sander, C. 1996. Mapping the protein universe. *Science* 273, 595–603.

Jones, D.T. 1999. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287, 797–815.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J.E. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* 299, 499–520.

Mallick, P., Weiss, R., and Eisenberg, D. 2002. The directional atomic solvation energy: an atom-based potential for the assignment of protein sequences to known folds. *Proc. Natl. Acad. Sci. USA* 99, 16041–16046.

Mizuguchi, K., Blundell, T.L., Gweon, H.S., et al. 2005. Homology recognition using environment-specific substitution scores enriched with homologous sequence information. Available at: *www.forcasp.org/paper1760.html*. Accessed September 13, 2006.

Murzin, A.G., Brenner, S.E., Hubbard, T., et al. 2005. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540.

Norvell, J.C., and Machalek, A.Z. 2000. Structural genomics programs at the U.S. National Institute of General Medical Sciences. *Nat. Struct. Biol.* S7, 931.

Notredame, C., Higgins, D.G., and Heringa, J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302, 205–217.

Orengo, C.A., Michie, A.D., Jones, S., et al. 1997. CATH—a hierarchic classification of protein domain structures. *Structure* 5, 1093–1108.

R Development Core Team. 2005. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: *www.R-project.org*. Accessed September 13, 2006.

Shi, J., Blundell, T.L., and Mizuguchi, K. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* 310, 243–257.

Xu, J., Li, M., Kim, D., et al. 2003. RAPTOR: optimal protein threading by linear programming. *J. Bioinform. Comput. Biol.* 1, 95–117.

Zhou, H., and Zhou, Y. 2005. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* 58, 321–328.

Address reprint requests to:
Dr. Gordon M. Crippen
College of Pharmacy
University of Michigan
428 Church St.
Ann Arbor, MI 48109-1065

E-mail: gcrippen@umich.edu