Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions†

Gregory B. Gloor,*,‡ Louise C. Martin,§ Lindi M. Wahl,§ and Stanley D. Dunn‡

Departments of Biochemistry and Applied Mathematics, The University of Western Ontario, London, Ontario, Canada, N6A 5C1

Received February 17, 2005; Revised Manuscript Received March 28, 2005

ABSTRACT: Information theory was used to identify nonconserved coevolving positions in multiple sequence alignments from a variety of protein families. Coevolving positions in these alignments fall into two general categories. One set is composed of positions that coevolve with only one or two other positions. These positions often display direct amino acid side-chain interactions with their coevolving partner. The other set comprises positions that coevolve with many others and are frequently located in regions critical for protein function, such as active sites and surfaces involved in intermolecular interactions and recognition. We find that coevolving positions are more likely to change protein function when mutated than are positions showing little coevolution. These results imply that information theory may be applied generally to find coevolving, nonconserved positions that are part of functional sites in uncharacterized protein families. We propose that these coevolving positions compose an important subset of the positions in an alignment, and may be as important to the structure and function of the protein family as are highly conserved positions.

Some positions in multiple sequence alignments are highly conserved, while others may vary a little or a lot. The conserved positions are obviously important, but mutations to nonconserved positions can also affect protein function. This is not surprising, since the structures and functions of proteins depend on highly complex interactions among their constituent residues, and the net stabilization of the folded state, relative to the unfolded state, is often small. How, then, do the variable positions change during evolution without compromising the activity of the protein, and are some variable positions more important for activity than others? In the laboratory, the detrimental effects of one mutation may be suppressed by a compensating mutation at another position, that is, a second-site suppressor mutation, and it is expected that during evolution the effects of mutations are often counterbalanced in a similar way. This would logically lead to the coevolution of the two positions.

It is reasonable to expect that coevolution should occur between residues that are adjacent in the three-dimensional folded structure of a protein, since compensatory changes between neighboring residues might help to maintain internal volumes, preserve salt bridges, or retain optimal hydrogen bonding. Thus, the analysis of coevolving residues could serve as a tool for identifying proximal residues in the folded structure, a major step toward the in silico determination of protein structure from sequence data. Numerous groups have attempted to demonstrate coevolution or covariance within proteins by analyzing multiple sequence alignments (*1*−*5*). In each case, a number of coevolving residue pairs have been reported, but it has proven difficult to separate the coevolution signal caused by structural or functional linkage from that caused by phylogenetic linkage (*1*−*5*). In some cases, there has been no assessment of the confidence that the signals reflect true coevolution rather than the random noise that will inevitably be present in an alignment containing a limited number of sequences. Furthermore, position pairs that have scored highly have often had no obvious structural relationship.

Tillier and Liu (*5*) used a correction for multiple interdependency to remove the underlying influence of genetic linkage, without needing to model the effect of this linkage, and successfully identified a number of coevolving positions. However, their analysis was biased toward those positions that covary with at most a few others and excluded coevolving groups of positions.

Here, we have analyzed position coevolution by applying information theory to multiple sequence alignments with a goal of understanding the structural or functional relationships between positions exhibiting high mutual information. To extract reliable information, it is essential that the alignment be of the highest possible quality. We therefore collected and aligned sequences using a rigorous, reproducible, but conservative method (Materials and Methods). Our use of information theory has been guided by initial studies of how the number of sequences in the alignment affect mutual information in simulated protein sequences, how the degree of variability, or entropy, at positions affects the mutual information scores, how the scores can best be normalized for that influence, and how to assess the

† This work was supported by grants from the Canadian Institutes of Health Research to G.B.G. and S.D.D., and from the Natural Sciences & Engineering Research Council of Canada and the Premier's Research Excellence Awards (PREA) to L.M.W. L.C.M. was supported through the PREA and Canada Research Chairs Programs.

* To whom correspondence should be addressed. E-mail, ggloor@uwo.ca; phone, 519-661-3526; fax, 519-661-3175.

‡ Department of Biochemistry, The University of Western Ontario.

§ Department of Applied Mathematics, The University of Western Ontario.

10.1021/bi050293e CCC: $30.25
Published on Web 04/22/2005

confidence that apparently high scores reflect mutual information due to structural or functional constraints rather than random noise and contributions from the phylogenetic background (*6*). Importantly, the method makes the assumptions that each position in a multiple sequence alignment is affected equally by phylogenetic linkage, and that the majority of positions in the alignment covary only because of linkage. On the basis of these assumptions, each alignment is used as its own null model for the identification of covarying positions.

Our analysis of 23 protein families shows that, in some cases, significantly coevolving positions are indeed proximal to one another, suggesting that they coevolve for reasons of local structure. However, the most significantly coevolving positions coevolve as groups, and these groups are usually located in or near active sites or binding interfaces. Consequently, coevolving positions in the groups may be either close together or somewhat farther apart, depending to some degree on the size of the ligand-binding interface. We postulate that the coevolution of positions in these groups is particularly strong because it is driven by functional as well as structural factors.

MATERIALS AND METHODS

Collection of Sequence Homologues. Excepting the homeodomain, alignments from the various protein family and protein domain databases typically contained fewer sequences than required for the analysis. Prior modeling showed that more than 125 nonidentical sequences were required for analysis (*6*). We thus selected protein families based on the presumed numbers of sequences that could be collected. The majority were selected from the Clusters of Orthologous Genes (COG)[1] database (*7*) in which we restricted our attention to those protein families which were present in the most species with the least number of paralogues. That is, protein families were chosen based on which were most likely to give orthologous sequences in the multiple sequence alignments. Protein families were further examined to ensure that at least one, and preferably more, structures were available to guide the alignments. This information for each protein family is tabulated as Supplementary Table 1 which can be found on the supplementary data Web site (http://www.biochem.uwo.ca/cgi-bin/MI/index.cgi).

Representative protein sequences were used as queries in a PSI-Blast (*8*) search of the Genbank nonredundant database released either June 16 or Nov 15, 2004. The PSI-Blast was allowed to converge, usually with an *E*-value cutoff of $1e^{-20}$. An *E*-value cutoff of $1e^{-40}$ was used if the search did not converge within 10 iterations. The single sequence from each organism with the lowest *E*-value in the converged data set was identified and kept. Homeodomain sequences were found at the Homeodomain Resource site located at the National Human Genome Research Institute (*9*). Additional information on how each protein family was collected is available in the supplementary data.

Multiple Sequence Alignments. Structure-based multiple sequence alignments were generated by modifying existing VAST (*10*) or Dali (*11*) structural alignments with the Cn3D program (*12*). The structure-guided option of the clustalw program (*13*) was used to align the sequences extracted from the PSI-Blast output using a gap penalty mask derived from the structural alignments. Positions in the structural alignment with well-conserved secondary structures were assigned a gap penalty multiplier of 4; unstructured loop regions and gaps were assigned a multiplier of 1. All other settings in clustalw were the defaults.

The Jalview alignment editor (*14*) was used to edit the alignments as follows: sequences that were truncated at the amino and carboxyl termini with reference to the structural alignments were deleted; an individual sequence was discarded if it was the only sequence that introduced a deletion relative to all others; groups of sequences were removed if they were of a different sequence family than expected. Finally, the sequences were culled with a 90% sequence identity cutoff.

Each sequence in the resulting alignment was checked for its inclusion in the correct COG. Any sequence not belonging to the correct COG was examined manually. In most instances, the sequence was found to be a gene fusion of the correct gene with another or to be a protein that had not been included in the current COG annotation because of recent addition to the database. Sequences that belonged to an incorrect COG which were not obvious gene fusions were removed from the alignment. In some cases, a neighbor-joining tree of the alignment was generated in clustalw (with multiple substitutions and ignore gap positions enabled) to aid in determining if a uniform protein family had been identified.

Calculation of Mutual Information. Shannon's entropy (*H*) is a measure of uncertainty or randomness (*15*). Standard methods were used to calculate entropy for each ungapped position in the alignment and for each pair of ungapped positions in the multiple sequence alignment (this is referred to as the joint entropy). The entropy of a column c in the alignment was determined as shown in the following equation:

$$H_c = -\sum_{i=1}^{20} p(x_i) \log_{20} p(x_i)$$

Here, $p(x_i)$ is the observed frequency of amino acid i occurring at a site. All values were calculated using a $\log_{20}$ scale so that the range of position entropy scores, H_c or H_d, was 0−1 (where c and d represent the columns in the alignment). These extreme values occur when one amino acid is completely conserved or when each of the 20 amino acids occurs in a column with the same frequency. The joint entropy H_{cd} was calculated by the same method using the frequencies of occurrence of each combination of residues in positions c and d. The range of joint entropy scores ranges from the maximum of H_c or H_d to the sum, $H_c + H_d$.

Mutual information (MI) was calculated as follows: $MI_{cd} = H_c + H_d - H_{cd}$ (*15*). The range of MI scores ranged from 0 to the minimum of H_c or H_d. The raw MI values were normalized by dividing by the joint entropy of the positions, H_{cd}, to reduce the influence of entropy on MI. Values of the MI/H_{cd} range from 0 to 1 (*6*, *16*). The mean and standard deviation of the MI/H_{cd} values were determined

[1] Abbreviations: COG, cluster of orthologous genes; *H*, Shannon's entropy; MI, mutual information; SD, standard deviation; *Z*, number of standard deviations from the mean; Å, angstrom; MAP1, methionine aminopeptidase type 1.

for each protein family. Finally, a *Z*-score (the number of standard deviations from the mean) was assigned to each normalized ratio. All values in the output files were keyed to the position number and the residue identity in a reference sequence obtained from NCBI for structural mapping. All files used in this analysis are in tab-delimited format and can be obtained from the supplementary data Web site.

Statistical Analysis. Bootstrapping was used to determine if the data distributions of two data sets overlapped. One million random samples of the size of the smallest data set were selected with replacement from the combined distribution, and the mean value for each sample was tabulated. The fraction of these mean values that is equal to or more extreme than the real value is a measure of the likelihood that the smaller data set could have been derived from the combined sets. We concluded that the data sets were significantly different when this likelihood was small.

RESULTS

Multiple Sequence Alignments and Mutual Information Calculation. Excepting the homeodomain, proteins were chosen for analysis if inspection of the COG (*7*) database indicated that sequence orthologs were likely to be found in most bacterial species. Multiple sequence alignments for 23 different protein families were generated as described in Materials and Methods and in the supplementary data (http://www.biochem.uwo.ca/cgi-bin/MI /index.cgi). The alignments used in this study contained between 129 (inorganic pyrophosphatase) and 237 (homeodomain) sequences. Seventeen of the protein families contained more than 150 sequences. All sequences in the alignments were confirmed to be members of the same orthologous family, as defined by the COG database, or were very closely related to a true member of that protein family. Essential statistics about each protein family are given in the supplementary data. There was an average distribution of about 2.5 sequences per taxonomic order, indicating a broad representation from a wide diversity of sequences in each alignment. This distribution was achieved by ensuring that no two sequences in the alignments were more than 90% identical.

Mutual information between all ungapped positions in these alignments was calculated as described under Materials and Methods. Mutual information (MI) is a measure of reduced uncertainty that is based on information theory (*15*). When applied to sequence alignments, it is the reduction in uncertainty of a pair of positions over what would be seen if the two positions were evolving independently. Thus, MI is high if the two positions are correlated. Obviously, MI is a potentially useful tool for the identification of coevolving positions in protein families and has been used extensively for that purpose (*1*−*5*). However, we have found that pairs of positions display significant background MI because of random pairings of residues when the number of sequences in the multiple sequence alignments is small (*6*). These results show that the practical lower limit is about 125 sequences. In addition, positions with high entropy (nonconserved positions) have more of this background MI than do positions with low entropy. Thus, raw MI values, as used by some others, is a very poor predictor of coevolution (*6*).

Our analysis of in silico evolved positions further revealed that dividing each MI value by the joint entropy of the pair of positions strongly reduced the confounding effect of positional entropy (*6*). The resultant MI/H_{cd} ratio measures the fraction of the maximum possible MI contained by a pair of positions. Interestingly, 1 - this ratio is a true distance measure which satisfies the triangle inequality (*16*).

A second source of background MI derives from the genetic linkage of residues in a given protein. In effect, any pair of positions in a given protein family cannot be independent because they are in the same gene, and thus inherited together. We found that this causes significant background MI even when there are thousands of sequences in the alignment (*6*). However, each position in a given alignment is affected equally by linkage. This means that the background linkage MI signal can be approximated for a given alignment by calculating the mean MI/H_{cd} ratio, if the assumption is made that the majority of positions in the alignment do not share significant MI. This linkage effect is most pronounced with conserved positions ((*6*), http://www.biochem.uwo.ca/cgi-bin/MI /index.cgi?WhyHCutoff). The analysis that follows therefore focuses on only nonconserved ungapped positions that contained less than about 3 bits of information, corresponding to an entropy of 0.3 or greater. This entropy cutoff was chosen because prior modeling of protein evolution and coevolution in silico showed that it was very difficult to separate positions which were coevolving from those that were not if either position was highly conserved (*6*). The same minimum entropy threshold was enforced by Tillier and Liu (*5*) for similar reasons. We note that highly conserved positions are identified easily in multiple sequence alignments using traditional means.

The position pairs with the highest MI/H_{cd} ratios were identified by calculating the number of standard deviations from the mean for each ratio in each alignment as described in Materials and Methods. Prior modeling showed that a *Z*-score of 4 was the minimum value that reliably identified coevolving positions in in silico-generated alignments (*6*), and this value was chosen as our minimum level of significance. There were 599 091 position pairs in the 23 alignments in which both positions were found in a representative PDB file. These position pairs formed the baseline for the studies described below.

Correlation between Maximum Z-score and Inter-Residue Distance. The simplest prediction of coevolution is that coevolving positions in protein families should interact with each other. This interaction could be direct or indirect, but in either case, coevolving position pairs should be located closer to each other in 3D space than the average position pair. This hypothesis was tested by measuring the mean minimum distance between non-hydrogen atoms of all possible residue pairs for the 23 representative structures ($n = 599\,091$ unique pairs) and comparing this to the mean minimum distance between those residue pairs in which one residue was the highest *Z*-score partner of the other (max*Z* partners, $n = 4337$ pairs). For reference, the minimal distances between atoms of residues that are in van der Waals contact are typically between 3 and 4; if the side chains are separated by the side chain of one other residue, the distance is usually increased to at least 7. The black bars in Figure 1 show the distances between all pairs of residues in the 23 protein families, and the dotted bars show the distances for each of the max*Z* partners. These distributions are obviously

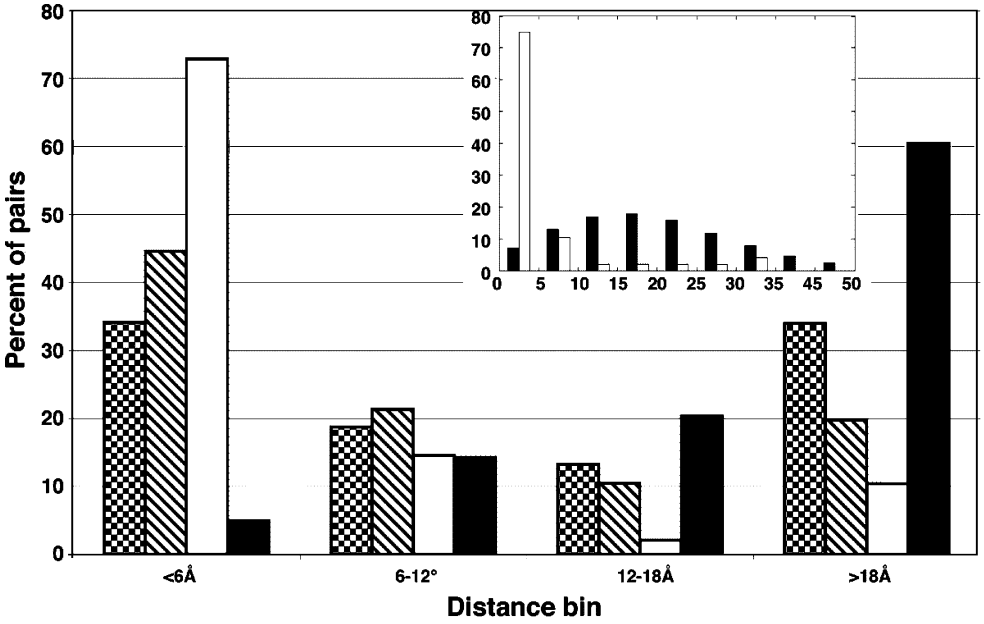

FIGURE 1: The minimum distance between non-hydrogen atoms for each residue pair in the representative structures were placed into bins. The main figure shows the distribution for all pairs as black bars, for max*Z* pairs as the dotted bars, for max*Z* pairs where $Z > 4$ as the striped bars, and for the individually coevolving pairs as the white bars. The inset uses the same axes, but includes more bins to highlight the difference in distribution between the individual and all-pairs data sets.

different. The mean distance between the max*Z*-score partners (14.32) was much smaller than the mean distance between all residue pairs (21.92). The likelihood of choosing 4337 residue pairs at random with a mean distance of 14.32 or less from the data set of all possible residue pairs was tested by random sampling with replacement as described in Materials and Methods. The mean distance of 1×10^6 random samples of 4337 pairs was 21.92 ± 0.165 (SD), and there were 0 random samples with a mean distance less than in the maximum *Z* data set. We conclude that residue pairs sharing a maximum *Z*-score are located much closer in space than would be expected by chance alone ($P \ll 1e^{-6}$).

The coevolution model also predicts that there should be a correlation between the *Z*-score of a position and the distance from its coevolving partner. Therefore, the average distance between max*Z* partners should decrease if partners with high *Z*-scores are examined. There were 248 position-pairs where both positions were nonconserved, where at least one position was another's max*Z* partner and in which *Z* was greater than 4. The striped boxes in Figure 1 show the distance distribution for these pairs. The average distance between these position-pairs in the representative structures was 11.31, which was significantly less than the average distance of all possible pairs ($P \ll 1e^{-6}$), and also significantly less than the average distance of all max*Z* partners ($P = 3.3e^{-5}$).

An even stronger prediction of the model is that positions that have strong direct interactions that are required for protein structure, stability, or function should coevolve only with each other. In most cases, these positions would be expected to make direct contact. We identified 48 non-conserved position pairs (96 residues) that had a maximum *Z*-score of 4 or more with only one other position, with the distribution shown by the white bars in Figure 1. A mean distance of only 7.10 separated these positions. The inset in Figure 1 shows the distance distribution for all the pairs and the individual pair data sets using a larger number of bins to highlight the distinct distributions. It was exceedingly improbable to choose a random sample of 48 residue pairs with an average minimum distance of 7.10 from the entire set of paired distances or the set of max*Z* partners ($P \ll 1e^{-6}$ in each case).

Nearly half (23 of 48) of these residue pairs were within 4.0 of each other, and over 70% (35 of 48) were within 5.0. In comparison, while the 48 residue pairs with the highest *Z*-scores among all pairs in the 23 families had a substantially higher average *Z*-score than the 48 individual pairs (6.7 vs 4.8), just 54% (26 of 48) of this group were within 5.0 of each other. If the *Z*-score cutoff for the individual pairs was raised only slightly, the selectivity for proximal positions was even stronger; 21 of 23 pairs with $Z > 4.5$ were within 5.0. The only exceptions were positions 98 and 120 in chorismate synthase (1QXO (*17*)), which are located at the same oligomeric interface of the tetramer, and positions 53 and 127 in nucleoside diphosphate kinase (1NLK (*18*)), which are located on two separate arms of the substrate-binding site. Interestingly, in at least one nucleoside diphosphate kinase structure (1K44 (*19*)), these latter two positions make a charge interaction across the binding site. We conclude that the identification of pairs of positions that share high *Z*-scores with only the other member of the pair is a good means of finding residues that are close to each other in 3D space. A complete list of these 48 residue pairs is available in the supplementary data.

Correlation between Maximum Z and Distance to a Molecular Interface. There were 352 positions that had a *Z*-score > 4 with one or more other positions. Of these, 256 positions were not in the single-pair data set. We will refer to these as grouped positions. In an attempt to identify why these grouped positions had high *Z*-scores, the corresponding residues were examined in a number of representative structures. Two such structures are shown in Figure 2, other structures may be viewed in the supplementary data. It was readily apparent that many of these residues were located

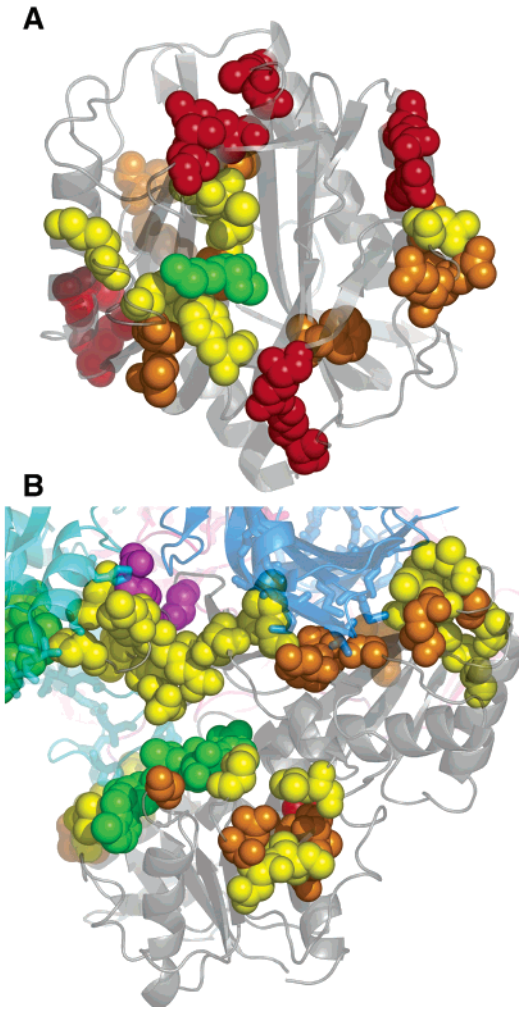

FIGURE 2: Representative structures showing the distribution of residues with *Z*-scores > 4. Shown here are structures for (A) the monomeric methionine amino peptidase 1 (1C24) and (B) tetrameric glyceraldehyde-3-*P*-dehydrogenase (1GD1). Chain A of 1C24 and chain O of 1GD1 are shown as the transparent gray cartoons in both structures; chains P, Q, and R in 1GD1 are the transparent cyan, blue, and pink cartoons. Residues in chain A of 1C24 and chain O of 1GD1 where $Z > 4$ are represented as spheres and are colored according to their group and contact status. Individual max*Z* pairs in which the residues in the pair contact are colored red; those that are not in contact are colored magenta. Residues that have *Z*-scores > 4 with three or more others are colored in yellow. All other residues with $Z > 4$ are orange if they do not belong to either category. Residues in chains P, Q, and R of 1GD1 that correspond to those shown as spheres in chain O are shown as sticks to illustrate the interface location of the positions in all four subunits. Ligands are shown in spacefill and colored green.

close to either the ligand-binding site of the molecule or to an oligomerization interface. For example, many of the grouped residues in the monomeric methionine amino peptidase (orange and yellow spheres in Figure 2A) are located at or near the substrate-binding pocket (*20*). A more complex example is glyceraldehyde-3-*P*-dehydrogenase (*21*) (Figure 2B). The tetrameric structure shown here has grouped residues primarily in the chain O substrate-binding pocket, along the interfaces with chains P and R and extending into the chain R substrate-binding pocket. The residues shown as sticks are the grouped residues in the other three subunits, and highlights that these residues are often found at the subunit interfaces. Grouped residues from chain R extend

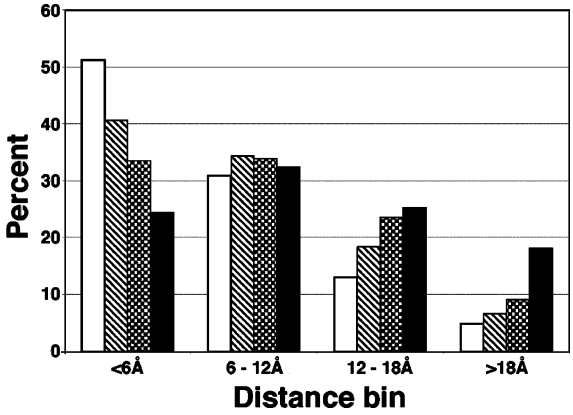

FIGURE 3: The minimum distance to a ligand-binding or oligomerization interface from residues in the 22 representative structures were placed into bins. The figure shows the distribution for all residues as black bars, for residues with $Z > 4$ as the dotted bars, for residues with $Z > 4$ that exclude the single pairs as the striped bars, and residues with $Z > 4$ with three or more others as the white bars.

into the chain O substrate-binding pocket. Results such as these led to the hypothesis that grouped residues might be located near interaction and ligand-binding surfaces. Similar representations of other molecules can be viewed in the supplementary data.

Among the 22 alignments for which the representative structures either contained a ligand or else were multimeric, there were a total of 4652 ungapped individual positions. The distance distribution of these residues to either a ligand-binding or oligomerization interface is shown as the black bars in Figure 3. The mean minimum distance to an interface from each of these residues was 11.19. The mean minimum distance to an interface of the 352 residues with *Z*-scores of 4 or more was 9.57 (dotted bar in Figure 3). The likelihood of selecting 352 residues from the 4652 interface distances with this average distance or less was very low ($P < 1e^{-6}$). Interestingly, we noted that the 96 residues identified previously in which both partners had a high *Z*-score with only one other residue had a mean minimum distance of 12.16 from these interaction surfaces (not shown). This distance was not substantially different from that expected by chance ($P = 0.067$). Therefore, these 96 residues were not contributing to the smaller average distance seen in the entire data set of high *Z* scoring positions. The distance distribution of the 256 residues which excludes these residues is shown by the striped bar in Figure 3.

This suggested that there was another set of residues with high *Z*-scores that approached the interfaces closely, and the data set was explored further to identify this subset of residues.

Figure 4 shows a diagram of the *Z*-score connections for positions with $Z > 4.0$ in the reference methionine amino peptidase alignment. Small groups of linked positions are in the upper left; isolated pairs are in the upper right; most of the residues within these sets are located on or near the surface of the protein, and are usually in contact with, or are located close to, their partner residues. The most prominent feature of the diagram, however, is a larger network with multiple connections among most members. In many cases, these positions have high *Z*-scores with several others and cannot be identified as having a clearly

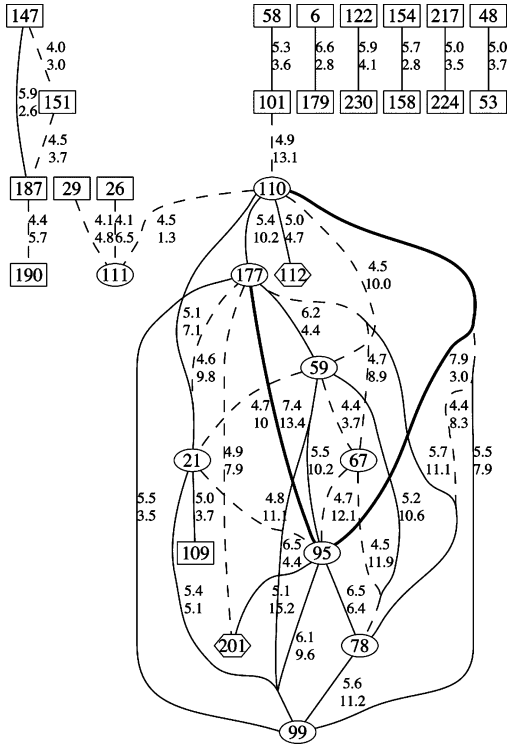

FIGURE 4: *Z*-score linkages in the methionine aminopeptidase alignment. Linkages between positions in the MAP1 alignment with residue numbers from a representative structure (1C24 (*20*)). Residues that make contact with their partner are enclosed in rectangular boxes; residues at or near the active site are enclosed in ovals, and the hexagonal nodes represent the residues for which neither of the other associations holds. Bold lines represent *Z*-scores greater than 7; solid lines are *Z*-scores between 5 and 7, and dashed lines represent *Z*-scores between 4 and 5. The top number to the right of each line is the *Z*-score; the bottom number is the distance of closest approach between the two residues.

higher *Z*-score with any single position. Connection diagrams such as this led us to investigate the properties of such high *Z*-score groups.

We found that the 123 nonconserved residues which shared *Z*-scores > 4 with three or more others (highly grouped residues) in the 22 alignments were often located very close to these interfaces, with a mean distance of 7.52. Their distance distribution is shown as the white bars in Figure 3. Their mean distance compared favorably with the mean distance (7.41, ($P = 0.41$)) of the 625 strongly conserved residues with a bit score > 3.9 ($H < 0.1$) from the same interfaces, but was significantly greater than the mean distance of 5.90 of the 180 absolutely conserved residues ($P = 0.009$). All three of these distances were significantly less than the mean distance of all residues from these interfaces ($P \ll 1e^{-6}$ in each case). Interestingly, the highly grouped positions had a much higher mean *Z*-score than did the paired positions (5.99 vs 4.60, $P \ll 1e^{-6}$), suggesting that the coevolution interactions between highly grouped positions were generally greater than between individually paired positions. We conclude that highly grouped, nonconserved residues represent a novel marker for protein interfaces. A complete list of these highly grouped residues is available in the supplementary data.

Comparison of MI and Mutagenesis Results. The results presented above led to the hypothesis that positions with high *Z*-scores may be important for protein structure or function, even though they were not conserved. This predicts that there should be a correlation between *Z*-scores and the importance of a residue for proper protein function. One way to test this is to determine the correlation between the effect of a mutation on the position and its *Z*-score. Data for the homeodomain and the *Escherichia coli* ϵ subunit are summarized below.

A recent review of the literature by D'Elia et al. (*22*) tabulated homeodomain missense mutations associated with human disease. They identified 27 positions in the 60 amino acid homeodomain that had a mutant phenotype when changed by a missense mutation; 11 of these had more than one independent mutation, and 7 of these had an entropy of > 0.3 and are analyzed here. The residues found to have multiple substitutions in this analysis correlated very well with residues known to be important for homeodomain function, and the mean maximum *Z*-score for these 7 residues (3, 5, 31, 42, 50, 52, and 54) was 3.4. The mean maximum *Z*-score for the 13 nonconserved residues in our alignment in which only one mutant was recovered was 2.1, and the mean maximum *Z*-score for the remaining 30 residues was 1.8. A Student *t*-test was used to compare each class with the others. This test showed that the single mutant and nonmutant classes were indistinguishable. The double mutant class was very different from the single and no-mutant classes, either separately or combined (combined classes, reject H_0, with $P = 2.3 \times 10^{-7}$). No difference was found between the mean entropies of these three sets. We conclude that those residues with high maximum *Z*-scores are more likely to be associated with a phenotype when mutated.

The *Escherichia coli* ATP synthase ϵ subunit has been extensively mutated in vitro, and the effects of these mutations have been characterized (*23*−*25*). These data are exceptional because information exists on both mutations that cause a functional change and those that do not. Table 1 shows the position and nature of each residue change, the effect of the mutation in a variety of assays, and the maximum *Z*-score for the residue. There is a notable difference in the maximum *Z*-score achieved by those residues that have a phenotypic effect when mutated and those that do not. Seven of the 12 residues that cause a measurable phenotype when mutated have a maximum *Z*-score of 3 or greater, while only one of the six mutants that do not cause a phenotype when mutated achieve this maximum score. The mean scores of the two groups are also different, being 3.3 and 2.4. A Student *t*-test was used to test the null hypothesis that these two means were the same, and H_0 was rejected with $P < 0.04$. As was observed in the homeodomain, the mean entropies of the two groups were the same. These data suggest that positions with high maximum *Z*-scores are more likely to change the activity of the protein upon in vitro mutagenesis than those that do not.

DISCUSSION

We used information theory to identify nonconserved amino acid residue pairs that coevolve for reasons of structure or function in a diverse set of protein families including enzymes, DNA binding proteins, and a structural member of a large protein complex. The majority of the position pairs with maximum *Z*-scores above 4 were at or near the active site of the protein. The fraction of positions at these interfaces

Table 1: Maximum Coevolution Scores for ϵ Residues Mutated in Vitro[a]

residue change	maximum coevolution score	functional changes
S65A	4.5	IDLM
S65C	4.5	I
D81A	4.5	LMP
F16A	4.0	AGM
E70A	3.7	IDLM
K46A	3.5	L
E31A	3.4	DIM
T77A	3.4	DGLM
K54A	3.2	nr[b]
T82A	2.9	I
R85A	2.8	IL
Q24A	2.8	nr[b]
S10A	2.4	AGM
S10C	2.4	I
Y63A	2.4	DGLMP
E29A	2.3	nr[b]
R51A	2.3	nr[b]
S108C	2.0	nr[b]
T43A	1.8	L
T43C	1.8	I
E21A	1.6	nr[b]

[a] Results compiled from ref *23*−*25*. Data for effects of all mutations on all functions are not available. Functions affected are indicated as follows: A, assembly of F_1F_0-ATPase; D, inhibition of membrane-bound ATPase by dicyclohexylcarbodiimide; G, cell growth; I, inhibition of F_1-ATPase by ϵ subunit; L, stimulation of membrane-bound ATPase by lauryl dimethylamine oxide; M, membrane-bound ATPase activity; P, ATP-driven proton translocation in membranes. Changes greater than 20% were considered significant. [b] nr, none reported.

could be enriched considerably by restricting the analysis only to positions that had a high *Z*-score with 3 or more others. This suggests that these positions tend to coevolve coordinately. A subset of the coevolving residues not near the active site or subunit interfaces were found to coevolve with only one other, and the amino acid side chains of these pairs were often found in contact with each other, suggesting that they are coevolving primarily to maintain local structure.

For the large majority of individual pairs in the families that we have examined, the residues are located at or near the surface of the protein structure, with at least one residue being solvent-exposed. Often, they are found near the beginning or end of a secondary structural element; this is particularly well-illustrated by the five isolated pairs of coevolving residues in methionine aminopeptidase I (see Figures 2, 4, and 5).

Two of the pairs are located at the C-terminal ends of helices and may assist in the formation of capping or termination motifs (*26*). Residues Tyr48 and Gln53, which interact through both side-chain and backbone atoms, occupy the C5 and Ccap positions of a helix that started at Ser37. Within this region, the carbonyl of Asp47 is hydrogen-bonded to the NH groups of both Asn51 and Glu52, while the carbonyl of Tyr48 forms H-bonds with the NH groups of both Glu52 and Gln53, distorting the helix and potentially fostering its termination. The following turn contains neither proline nor glycine residues. Residues Gln154 and Glu158, which also exhibit extensive interactions, occupy the C6 and C2 positions of another helix. In this case, the turn following the helix contains a Schellman-type motif with glycine in the C′ position. The third pair of coevolving residues that are close together in sequence, Met217 and Lys224, make hydrophobic contact from adjacent strands in a short antiparallel β-sheet. The carbonyl group of Met217 is hydrogen-bonded to the NH of Gly220 located in the turn joining the two strands.

The two other residue pairs highlighted in Figure 3C are more distant in sequence. Residue Lys6 is just before the first long helix that extends from Pro8 to Val32, while its salt-bridge partner, Glu179 is one of the first residues of an extended strand on the edge of an antiparallel β-sheet. The side chain of Leu230, the first residue of an extended strand that passes through the center of the structure, interacts with Cα of Gly122, located near the beginning of a long helix that extends from residue Thr118 to Val139.

That these particular position pairs gave far higher *Z*-scores than most pairs that are in contact implies that they are particularly important for the structure or conformation of the protein, either by defining the limits of secondary structural elements or else by participating in the establishment and maintenance of relationships between secondary structural elements that are distant in sequence. We therefore propose that positions identified by this method as isolated pairs are particularly important for protein folding and structural stability.

Coevolution of clusters of positions, which are usually not in contact, has been previously reported by other groups (*27*). Here, we have determined that these clustered positions tend to be located near binding regions or active sites, leading to a possible explanation of why such groups of noncontacting positions would coevolve. We think that the coevolution is related to the stricter constraints on residues within functional areas, even if they are not directly involved in catalytic mechanisms. These constraints are illustrated by the prevalence of highly conserved residues within these areas (see blue, cyan, and green residues in Figure 5), and it is our hypothesis that residues in coevolving groups (yellow and orange residues in Figure 5), which comprise most of the other residues near the active site, may be almost as important as the conserved residues. The primary difference would be in whether a change in one of the positions can be compensated by an additional change in another position. For some positions, such as those directly involved in catalysis, no compensatory second site mutations may be possible; such residues would be highly conserved. However, at other positions, some variation may be allowed provided an overall volume or shape is maintained; these positions would be free to vary within particular bounds. If the initial compensatory mutation left suboptimal functionality, additional changes would arise. Since these mutable positions will be interspersed with highly conserved positions in the functional region, one can envision how a network of coevolving positions, many of which are not in contact, could arise. The structural relationship between conserved positions and group-coevolving positions can be seen in the cutaway view shown in Figure 5.

In the MAP1 structure with the bound transition state analogue methionine phosphinate (1C24.PDB) residues Phe177, Cys59, Thr99, and Lys67 make a contiguous cluster, with the first three forming a part of the binding site for the substrate methionyl side chain (in Figure 5, the bound transition state analogue methionine phosphinate is shown in magenta). One could hypothesize that these residues contribute either to the efficiency or to the specificity of

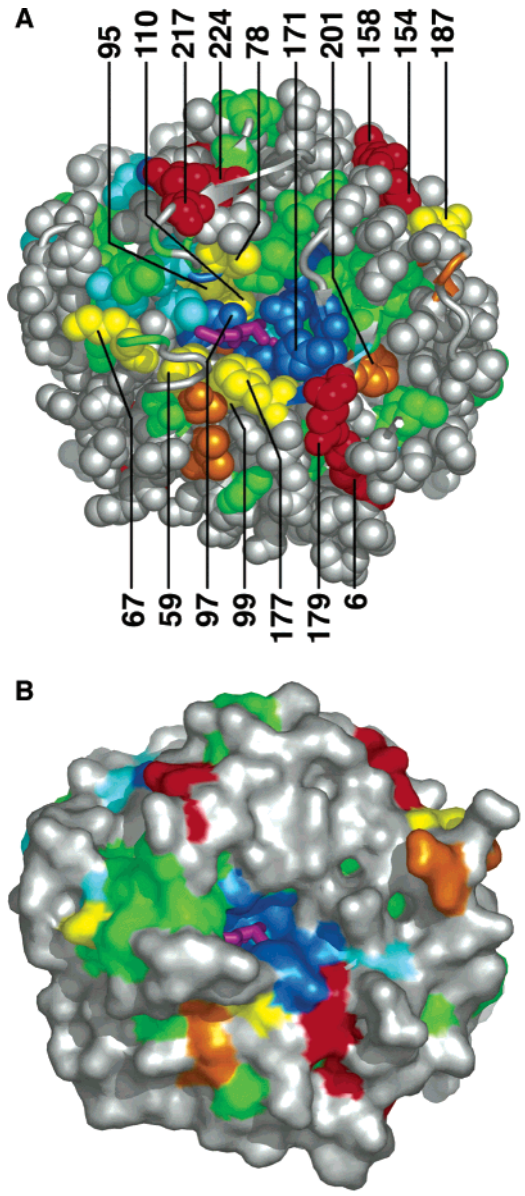

FIGURE 5: Two views of MAP1 highlighting conserved and coevolving residues near the active site. Shown here are two views of *E. coli* methionine aminopeptidase bound with the transition state analogue methionine phosphiniate. The top image (A) shows a cutaway view of the molecule where residues obscuring the binding pocket are shown in cartoon format, and all others are represented as spheres. The bottom image (B) shows a surface representation of the molecule from the same perspective. Residues are colored according to the properties of the corresponding positions in the multiple sequence alignment: completely conserved residues in blue (entropy = 0); highly conserved residues (*H* between 0 and 0.1) in cyan; conserved residues (*H* between 0.1 and 0.3) in green; residues that share a reciprocal high *Z*-score with only one partner in red; residues that share a high *Z*-score with three or more others in yellow; and residues that share a high *Z*-score with 2 others, or with one other that is part of the coevolution group, in orange; nonconserved residues that do not have high *Z*-scores in gray. The ligand is shown as the purple stick figure. Residues mentioned in the Discussion are identified.

substrate recognition. Residues Asn95, Ala21, Thr109, and Ser110 form a separate contiguous grouping. Rather than interacting with the substrate, these residues lie behind the conserved residues, Asp97 and Asp108, that ligate a metal ion cofactor at the active site. As seen in Figure 4, Phe177 and Asn95 seem to be particularly important; besides having many connections, MI/H_{cd} for this pair is unusually high ($Z = 7.4$) despite a separation of more than 13 Å. It is notable that the shortest path between the two residues passes through only the substrate-binding site and the conserved residue Asp97. One other member of the coevolving cluster, Cys78, is separated from the two contiguous groups only by well-conserved residues and the substrate methionine side chain, while another member, Phe201, is separated from Phe177 only by conserved residues His171 and Gly172, and from Ser110 only by well-conserved residues Glu235 and Ile203.

It is notable that the coevolution scores for connections within the group associated with the functional domain are higher than those between singly coevolving pairs. We have found more complicated networks in larger proteins or complexes (not shown, connectivity graphs available in the supplementary data). In contrast to the groups, the isolated pairs described above are usually distant from functional regions and are usually near the protein surface in areas where constraints are lower as evidenced by the lower frequency of highly conserved positions. This can also be seen in Figure 5B, which shows a surface representation of the same protein. Here, it is clear that the conserved and group coevolving residues are largely buried, but the single coevolving pairs (shown in red) are surface exposed. In this situation, highly coevolving groups are less likely to develop, and a single key interposition interaction can result in a pair that coevolves in relative isolation.

We used a *Z*-score of 4 or more to identify coevolving positions because our earlier studies with in silico evolved sequences indicated that this cutoff provided a good combination of sensitivity and selectivity (*6*). The evident physical and functional relationships among the pairs and clusters that were found strongly support a coevolutionary basis for their high scores. However, one must expect that many other truly coevolving positions were not identified because their scores were less than 4. This is supported by the observation, shown in Figure 1, that the average distance between max*Z* partners is far less than expected by chance in this set of proteins. The residues in over a third of these pairs are separated by less than 6 Å. Thus, we expect that at least some residue pairs with *Z*-scores of less than 4 are coevolving to various degrees.

Unfortunately, two major factors besides structural or functional coevolution will contribute to the mutual information in multiple sequence alignments, even after normalization by H_{cd}. One is the signal arising because of the finite number of sequences in the alignment. The second is the phylogenetic relationships among the organisms represented in the alignment; the sequence similarities that allow construction of evolutionary trees from the alignments will also produce a mutual information signal in our analysis. This latter signal is largely removed by the scoring scheme used here, which assumes that the phylogenetic signal is relatively constant across the alignment. Among pairs showing coevolution values between 3 and 4 ,we presently cannot distinguish those that are truly coevolving from those that show high signals for the other reasons. Some pairs are close together in the protein structure, but some are not.

While the results presented above strongly suggest that this approach will be useful, it should be noted that there is as yet no direct experimental proof that coevolving positions are important contributors to protein structure or function.

Such results would strengthen the theoretical and practical applications of the method considerably.

There are four major reasons that the method used here was successful. First, we ensured that the multiple sequence alignments contained a sufficient number of diverse sequences. Analysis of alignments produced through simulated evolution showed that the alignments should contain at least 125 sequences so that the coevolution signal exceeds the background noise caused by small sample sizes (*6*). This condition is not met for many entries in the protein family databases, and could be one reason that Tillier and Liu found only a handful of coevolving positions in their study of the entire PFAM database (*5*). Second, the multiple sequence alignments were based on structural information, and nonorthologous sequences were removed by hand to ensure the alignments were of high quality. Oliviera et al. (*1*) also recognized this problem and used custom-generated alignments in their study. The method described here could identify the many of the same coevolving positions using lower quality alignments; however, in these instances, lower *Z*-scores were derived (not shown). Thus, while the method is robust, the confidence in assigning covarying positions depends on the quality of the alignment, as will any method using such alignments. High quality, structure-guided alignments are becoming more readily available and are beginning to augment alignments found in protein family databases. For example, the Conserved Domain Database at NCBI is continuing to add value to sequence alignments by including structural information (*28*). Third, we normalized the MI by dividing it by H_{cd}, providing a value related to a true distance measure. This reduced the influence of entropy on the coevolution score. This normalization was not used in any of the previous studies; others have attempted to account for some of the influence of entropy on MI scores by stratifying the data for residue variability (*1*) or by using a generic model of amino acid substitution (*29*). Using the refinements described above, we developed a general method to identify positions in multiple sequence alignments that coevolve significantly more than expected by chance and demonstrated the utility of the approach in a variety of protein classes. Finally, unlike Tillier and Liu, we did not make the assumption that positions coevolve primarily in pairs or small groups (*5*).

The method used here will likely prove to be useful to those who are studying a protein, domain, or structure of unknown function as in structural genomics projects. We have shown that residues with high *Z*-scores may be generally involved in the major activity of the protein or domain, or may form an important structural element in the protein. These residues are more likely to give a strong phenotype when mutated and are usually closer to the ligand of the protein than residues with lower maximum coevolution scores. The ability to segregate coevolving positions into coevolving groups, that generally do not contact each other, and into pairs that are likely to form close contacts with their coevolving partner, may prove useful in de-novo protein structure prediction. The method may also be useful in identifying the proper binding partner of a given protein. Preliminary investigations along this line are encouraging as positions between noninteracting protein families do not share significant MI by this method (not shown). As more genome sequences become available and methods to discriminate mutual information due to structure and function are further developed we anticipate that the application of information theory to identify coevolutionary relationships among positions in proteins will become an increasingly powerful and important bioinformatic approach.

ACKNOWLEDGMENT

We thank Drs. Brian Shilton, Gary Shaw, and David Litchfield for their comments that improved the manuscript. We also thank Drs. Duncan Murdoch and Kaizhong Zhang for their insights at early stages of this project.

REFERENCES

1. Oliveira, L., Paiva, A. C., and Vriend, G. (2002) Correlated mutation analyses on very large sequence families, *ChemBioChem 3*, 1010−1017.
2. Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W., and Dress, A. W. (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis, *Mol. Biol. Evol. 17*, 164−178.
3. Clarke, N. D. (1995) Covariation of residues in the homeodomain sequence family, *Protein Sci. 4*, 2269−2278.
4. Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis, *Proc. Natl. Acad. Sci U.S.A. 90*, 7176−7180.
5. Tillier, E. R., and Lui, T. W. (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments, *Bioinformatics 19*, 750−755.
6. Martin, L. C., Gloor, G. B., Dunn, S. D., and Wahl, L. M. Using information theory to search for co-evolving residues in proteins, *Bioinformatics*, submitted for publication.
7. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003) The COG database: an updated version includes eukaryotes, *BMC Bioinf. 4*, 41.
8. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res. 25*, 3389−3402.
9. Banerjee-Basu, S., Moreland, T., Hsu, B. J., Trout, K. L., and Baxevanis, A. D. (2003) The homeodomain resource: 2003 update, *Nucleic Acids Res. 31*, 304−306.
10. Gibrat, J. F., Madej, T., and Bryant, S. H. (1996) Surprising similarities in structure comparison, *Curr. Opin. Struct. Biol. 6*, 377−385.
11. Holm, L., and Sander, C. (1996) Mapping the protein universe, *Science 273*, 595−603.
12. Wang, Y., Geer, L. Y., Chappey, C., Kans, J. A., and Bryant, S. H. (2000) Cn3D: sequence and structure views for Entrez, *Trends Biochem. Sci. 25*, 300−302.
13. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003) Multiple sequence alignment with the Clustal series of programs, *Nucleic Acids Res. 31*, 3497−3500.
14. Clamp, M., Cuff, J., Searle, S. M., and Barton, G. J. (2004) The Jalview Java alignment editor, *Bioinformatics 20*, 426−427.
15. Ash, R. B. (1965) *Information Theory*, Interscience Publishers, New York.
16. Bardera, A., Feixas, M., and Boada, I. (2004) Normalized similarity measures for medical image registration, in *International Symposium in Medical Imaging* (Fitzpatrick, J. M., and Sonka, M., Eds.) pp 108−118, SPIE, San Diego, CA.
17. Maclean, J., and Ali, S. (2003) The structure of chorismate synthase reveals a novel flavin binding site fundamental to a unique chemical reaction, *Structure 11*, 1499−1511.
18. Williams, R. L., Oren, D. A., Munoz-Dorado, J., Inouye, S., Inouye, M., and Arnold, E. (1993) Crystal structure of *Myxococcus xanthus* nucleoside diphosphate kinase and its interaction with a nucleotide substrate at 2.0 Å resolution, *J. Mol. Biol. 234*, 1230−1247.

19. Chen, Y., Morera, S., Mocan, J., Lascu, I., and Janin, J. (2002) X-ray structure of *Mycobacterium tuberculosis* nucleoside diphosphate kinase, *Proteins 47*, 556−557.
20. Lowther, W. T., Zhang, Y., Sampson, P. B., Honek, J. F., and Matthews, B. W. (1999) Insights into the mechanism of *Escherichia coli* methionine aminopeptidase from the structural analysis of reaction products and phosphorus-based transition-state analogues, *Biochemistry 38*, 14810−14819.
21. Skarzynski, T., Moody, P. C., and Wonacott, A. J. (1987) Structure of holo-glyceraldehyde-3-phosphate dehydrogenase from *Bacillus stearothermophilus* at 1.8 Å resolution, *J. Mol. Biol. 193*, 171−187.
22. D'Elia, A. V., Tell, G., Paron, I., Pellizzari, L., Lonigro, R., and Damante, G. (2001) Missense mutations of human homeoboxes: a review, *Hum. Mutat. 18*, 361−374.
23. Tang, C., and Capaldi, R. A. (1996) Characterization of the interface between gamma and epsilon subunits of *Escherichia coli* F1-ATPase, *J. Biol. Chem*, *271*, 3018−3024.
24. Xiong, H., and Vik, S. B. (1995) Alanine-scanning mutagenesis of the epsilon subunit of the F1-F0 ATP synthase from *Escherichia coli* reveals two classes of mutants, *J. Biol. Chem. 270*, 23300−23304.
25. Xiong, H., Zhang, D., and Vik, S. B. (1998) Subunit epsilon of the *Escherichia coli* ATP synthase: novel insights into structure and function by analysis of thirteen mutant forms, *Biochemistry 37*, 16423−16429.
26. Aurora, R., and Rose, G. D. (1998) Helix capping, *Protein Sci. 7*, 21−38.
27. Pritchard, L., and Dufton, M. J. (2000) Do proteins learn to evolve? The Hopfield network as a basis for the understanding of protein evolution, *J. Theor. Biol. 202*, 77−86.
28. Kann, M. G., Thiessen, P. A., Panchenko, A. R., Schaffer, A. A., Altschul, S. F., and Bryant, S. H. (2005) A structure-based method for protein sequence alignment, *Bioinformatics 21*, 1451−1456.
29. Wollenberg, K. R., and Atchley, W. R. (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap, *Proc. Natl. Acad. Sci. U.S.A. 97*, 3288−3291.

BI050293E