ELSEVIER

Comparing protein structures: A Gaussian-based approach to the three-dimensional structural similarity of proteins

G.M. Maggiora, D.C. Rohrer, and J. Mestres

Computer-Aided Drug Discovery, Pharmacia Corporation, Kalamazoo, MI 49007-4940, USA

This study describes a new method for comparing three-dimensional protein structures based on an optimal alignment of their steric fields. The method is based upon the use of spherical Gaussian functions located on individual atoms. This representation generates a flexible description of the underlying fold geometry of proteins that can be adjusted by changing the 'width' of the Gaussians. Reducing the width sharpens the representation and leads to a more 'atomlike' description; increasing the width creates a fuzzier representation that preserves the general shape features of the chain fold but with a consequent loss in atomic resolution. The width used in this study is based upon the features of individual atoms and provides a representation that is quite robust with respect to the variety of geometric features typically encountered in the alignment process. In addition, a post-alignment analysis is performed that generates sequence alignments from the corresponding structure alignments. An example, based on five mammalian and fungal matrix metalloproteinase crystal structures (human fibroblast collagenase, neutrophil collagenase, stromelysin, astacin, and adamalysin), illustrates a number of features of the Gaussian-based approach. © 2001 by Elsevier Science Inc.

INTRODUCTION

Modern genomics technologies are generating vast amounts of sequence data on a wide variety of proteins. While the function of many these proteins is known, a substantial number of unknown functions remain. In addition, many proteins have incorrectly assigned functions. A major impediment to the proper assignment of a protein's function is generally its low level of sequence similarity to other proteins of known function. Fortunately, the problem of low sequence similarity is ameliorated in many cases by the availability of structural information, since three-dimensional structural similarity is preserved within functional families even when sequence similarity is not. In recent years, modern crystallographic and NMR techniques have lead to substantial increases in the number of available protein structures, but robust methods for aligning protein structures are required to take advantage of this information. This study describes a method, which is based upon the use of atom-centered Gaussian functions to describe the overall shape of a protein. The Gaussians provide a 'soft' or 'fuzzy' description of the protein, and this property gives the method its robustness. A preliminary version of this work was presented earlier.[1]

Many methods exist for determining structural similarity of proteins. The oldest, and most venerable, is based upon minimizing the root-mean-square-distance (RMSD) of a given set of atoms (generally the α-carbons) of one protein with the corresponding set of atoms in the other protein. However, determining the atom correspondences requires performing a sequence alignment, which can be quite difficult when proteins of low sequence similarity are being aligned. Moreover, since structural similarity is preserved far more than sequence similarity, determining the sequence alignment first is 'putting the cart before the horse.' Thus, since the early 1990s, a significant amount of effort has gone into developing sequence-independent methods. A sampling of these methods (that is by no means a complete) is provided in the references.[2–32] The basic methodology, which is implemented in the program MIMIC,[33] is described in the Methods section and includes discussions of the 'soft' Gaussian-based representation of protein structure; the similarity function used for both pairwise and multiple molecule alignments; the optimization procedure employed for overlaying structures; and an example of one type of post-alignment analysis that deals with structure-derived sequence alignments. The Results and Discussion section deals with the findings of our studies of five $\alpha + \beta$ proteins in the

Corresponding author: G.M. Maggiora, Pharmacia Corporation, 301 Henrietta Street, Kalamazoo, MI 49007-4940, USA

E-mail address: gerald.m.maggiora@am.pnu.com

Journal of Molecular Graphics and Modelling 19, 168–178, 2001

655 Avenue of the Americas, New York, NY 10010

1093-3263/01/$–see front matter
PII S1093-3263(00)00129-7

family of matrix metalloproteinases (MMPs) from mammalian and fungal sources (corresponding PDB identifiers are given in parentheses): Human Fibroblast Collagenase (1HFC),[34] Neutrophil Collagenase (1MNC),[35] Stromelysin (2SRT),[36] Adamalysin (1IAG),[37] and Astacin[38] (1AST). In addition to discussing pairwise and multiple structure alignments, we also present a description of a procedure for deriving sequence alignments derived from three-dimension structure matches. We conclude with a brief summary of our main points and a discussion of future applications.

METHODS

An Adaptable Gaussian-Based Representation of Three-Dimensional Protein Structure

Consider a set of K proteins

$$P = \{A,B, \ldots ,K\}; \tag{1}$$

where each protein, say A, is made up of a set of n_A amino acids,

$$A = \{A_1,A_2, \ldots ,A_{n_A}\}. \tag{2}$$

The approach described here is based on the use of spherically-symmetric Gaussian functions ('Gaussians'),

$$g_k^{A_i}(\mathbf{r}) = c_k \cdot \exp(d_k|\mathbf{r} - \mathbf{R}_k|^2) \tag{3}$$

where $g_k^{A_i}(\mathbf{r})$ is located on the k-th atom of the i-th amino-acid residue of protein A. The constants c_k and d_k, which influence the magnitude and width of each Gaussian, are assigned as discussed in our earlier work,[33] but are fully adjustable and can be tailored to meet the needs of a particular study. These functions possess many desirable properties that make them ideally suited for this study.[39] Spherical Gaussian functions have also been used by Grant and Pickup[40] to described the Van der Waals volumes of small molecules, but the form of their molecular representation is different than that used here. Each amino-acid residue, A_i, can be described by a linear combination of Gaussians,

$$G_{A_i}(\mathbf{r}) = \sum_{k\in A_i} g_k^{A_i}(\mathbf{r}). \tag{4}$$

Note that any subset of atoms of an amino-acid residue appropriate to a given study may be used. For example, if only a 'low-resolution' alignment is needed, a single Gaussian located on the α-carbon of each residue will most likely be sufficient. This is represented by $|A_i| = 1$, where $|A_i|$ is the number of Gaussians used to represent the i-th amino-acid residue. Including just the 'backbone' atoms [—N—C_α—C—] gives $|A_i| = 3$, while including all of the 'main chain' atoms [—N—C_α—(C═O)-] gives $|A_i| = 4$. These latter two representations provide increased 'resolution,' but at the expense of additional computation. Interestingly, using all atoms including those of the sidechains, $|A_i| = N_i$, where N_i is the number of atoms in the i-th residue, does not appear to improve structure alignments significantly. For all of the studies described here $|A_i| = 4$ is used. A Gaussian-based description of an entire protein molecule is then obtained by combining each of the residue representations in a similar fashion to Equation 4:

$$G_A(\mathbf{r}) = \sum_{A_i\in A} G_{A_i}(\mathbf{r}). \tag{5}$$

Written in terms of the individual Gaussians, Equation 5 becomes

$$G_A(\mathbf{r}) = \sum_{A_i\in A} \underbrace{\sum_{k\in A_i} g_k^{A_i}(\mathbf{r})}_{i\text{th residue}}. \tag{6}$$

Thus, each $G_A(\mathbf{r})$ essentially provides a 'soft' or fuzzy representation of the three-dimensional structure of the protein. Equation 6 exhibits, in an implicit manner, the fundamental flexibility of the Gaussian-based procedure presented here. Although the summations given in Equation 4 and Equation 5 are taken over each amino-acid residue and then over the entire set of residues, respectively, the summation can be easily rearranged in many different ways. For example, the backbone atoms can be considered separately from the sidechain atoms such that

$$G_A(\mathbf{r}) = \sum_{A_i\in A} \sum_{\alpha\in\{bb,sc\}} \sum_{k_\alpha\in A_i} g_{k_\alpha}^{A_i}(\mathbf{r}), \tag{7}$$

where *bb* and *sc* represent 'backbone' and 'sidechain' atoms, respectively. The pairwise similarity of two proteins represented in this manner results in a set of pairwise backbone–backbone, sidechain–sidechain, and backbone–sidechain similarities, which additively yield the overall pairwise similarity of the two proteins.

Similarity Indices for Structurally Aligning Proteins

The three-dimensional structures of two proteins can be compared using the following similarity measure[41]

$$\Omega_{AB} = \int G_A(\mathbf{r}) \cdot G_B(\mathbf{r}) d^3\mathbf{r}. \tag{8}$$

This can be accomplished by translating and rotating the two proteins to maximize the value of Ω_{AB}, which is basically a measure of the structural overlap of the proteins. Generally, this measure is normalized by the self-similarities, which are constant and depend only on the structures of the individual proteins. The normalized measure, called a similarity index, which was originally introduced by Carbó et al.,[41]

$$Sim(A,B) = \frac{\Omega_{AB}}{\sqrt{\Omega_{AA}} \cdot \sqrt{\Omega_{BB}}}, \tag{9}$$

is bounded by zero and unity, i.e., $0 \le Sim(A, B) \le 1$. This index is related in form to a cosine function, and is only one of a large family of related indices.[42] A distinct advantage of the Gaussian-based approach is that the integrals in Equation 9 (see also Equation 8) can all be written in simple, closed form. This can be seen by substituting Equation 6 into Equation 8 and regrouping terms, which yields

$$\Omega_{AB} = \sum_{A_i\in A} \sum_{B_j\in B} \left[\sum_{k\in A_i} \sum_{\ell\in B_j} \int g_k^{A_i}(\mathbf{r}) \cdot g_\ell^{B_j}(\mathbf{r}) d^3\mathbf{r} \right], \tag{10}$$

where the integral over Gaussians can is given by

$$\int g_k^{A_i}(\mathbf{r}) \cdot g_\ell^{B_j}(\mathbf{r}) d^3\mathbf{r} \sim \left(\frac{\pi}{d_k + d_\ell}\right) \times \exp\left(-\frac{d_k d_\ell}{d_k + d_\ell}|\mathbf{R}_k - \mathbf{R}_\ell|^2\right). \quad (11)$$

While the above methodology can be generalized to handle multiple matches, it is computationally inefficient. Thus, a simple average of the pairwise similarities that we employed in our earlier work[43,44] is used here to handle multiple matches

$$\langle Sim(A,B,\ldots,M)\rangle = \left[\frac{2}{|P| \cdot (|P| - 1)}\right] \cdot \sum_{I,J\in\{P\}\, I\neq J} Sim(I,J), \quad (12)$$

where $P = \{A, B, \ldots, M\}$ is the set of proteins being compared, and $|P|$ is the number of elements in the set P. Previous work on small molecules[43,44] showed that Equation 12 produces results that are in excellent agreement with those obtained by the more complete formulation of multiple match similarity.

Optimization Procedure

A key element of our method involves optimizing the similarity function given in Equation 9. As *Sim*(*A*,*B*) is a nonlinear function of the relative positions and orientations of the proteins being aligned, determination of the global maximum of the function can be problematic. A modified form of the systematic search procedure implemented in MIMIC[33] is used to increase the likelihood that the global maximum is obtained. In the procedure, one of the proteins, called the reference protein, is held fixed while the other, called the adapting protein, is moved until one of the many possible optima of *Sim*(*A*, *B*) is found. More specifically, the centers of mass of the proteins are first superimposed. The adapting protein is then systematically rotated in 45° increments, and the similarity is evaluated at each of the 208 unique points thus generated. A rank-ordered list of the similarity values is produced, and the positions corresponding to the 'best' values are chosen as starting points for a standard gradient-based optimization. Each optimization converges to a maximum, but not necessarily the same maximum, of the similarity function, so that many solutions are typically obtained. Generally, in practical applications of the procedure the starting geometries corresponding to only the top five initial similarities are considered. In all cases examined to date, the starting geometry corresponding to the largest initial similarity value converges to the global maximum of *Sim*(*A*, *B*). Moreover, in the case of protein alignments, as opposed to small molecule alignments, the best solution generally stands out from the remaining solutions. However, as pointed out by Zu-Kang and Sippl,[7] secondary solutions obtained by structure alignment algorithms may also be of value, especially those lying close to the primary solution.

Post-Alignment Analysis—Structure-Based Sequence Matching

Although *Sim*(*A*,*B*) provides a measure of the similarity of protein *A* to protein *B*, it is sometimes useful to compute the

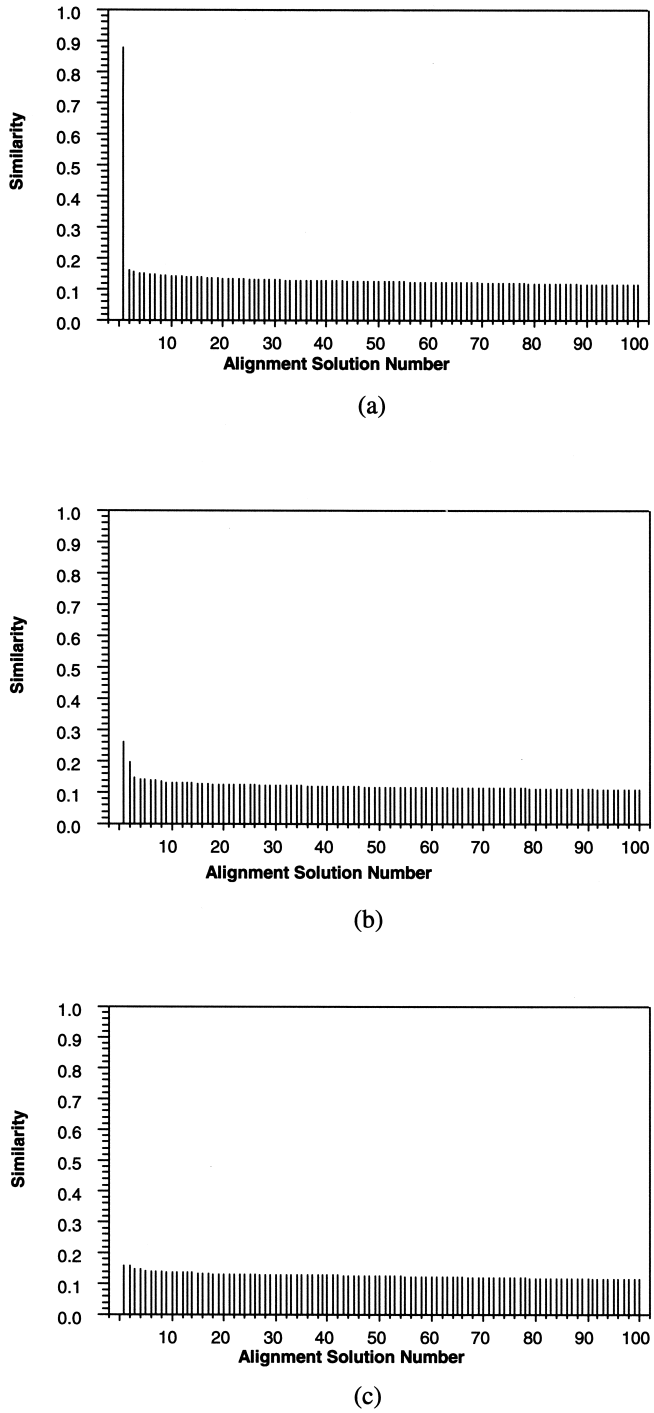


Figure 1. Similarity values corresponding to multiple alignment solutions for specific pairwise matches. (a) 1HFC and 1MNC: The primary solution, with a value of ~0.85, stands out from the 'background' of the remaining solutions, all of whose values are all less than ~0.16. (b) 1HFC and 1IAG: The primary and secondary solutions, with values of ~0.25 and ~0.19, standout from the 'background' of the remaining solutions, all of whose values are less than ~0.15. (c) Pairwise similarity solutions for the MMP actinidin (2ACT), an α + β protein, and the β protein trypsin (1TRY).

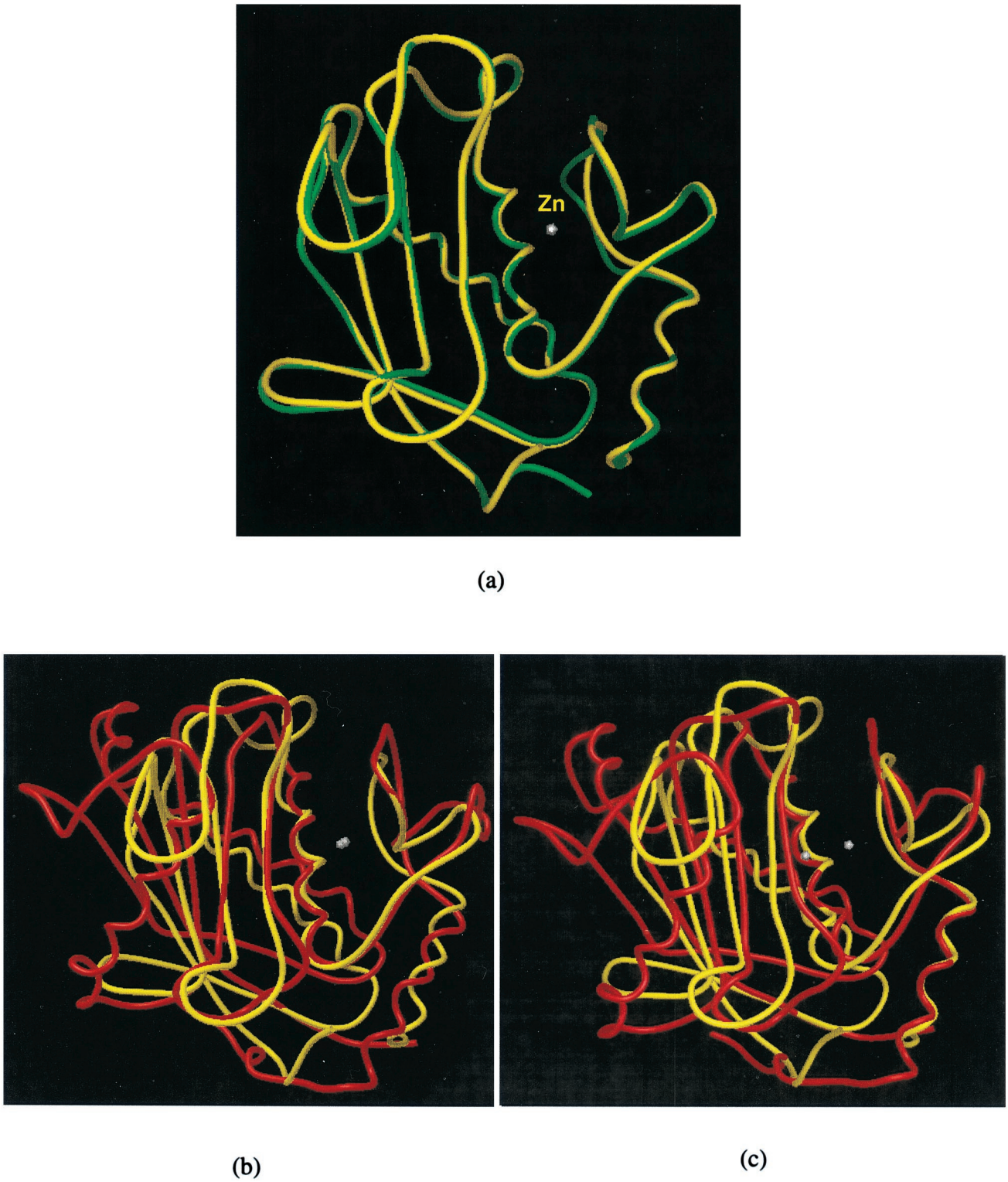


Figure 2. Graphical depictions of the best similarity-based pairwise alignments. (a) 1HFC and 1MNC with a structural similarity of about 85%. (b) 1HFC and 1IAG with a structural similarity of about 25%, corresponding to the 'best' alignment solution. (c) 1HFC and 1IAG with a structural similarity of about 19% corresponding to the 'second best' alignment solution. Note the catalytic zinc ion located within the region of the helix-turn-β-strand motif in all three depictions.

inter-residue similarities between amino acids on the two proteins as will be illustrated in the section on converting three-dimensional overlays to their corresponding sequence matches. Inter-residue similarity indices are obtained simply by modifying Equation 8 and Equation 9 as follows:

$$\Omega_{A_iB_j} = \int G_{A_i}(\mathbf{r}) \cdot G_{B_j}(\mathbf{r}) d^3\mathbf{r} \tag{13}$$

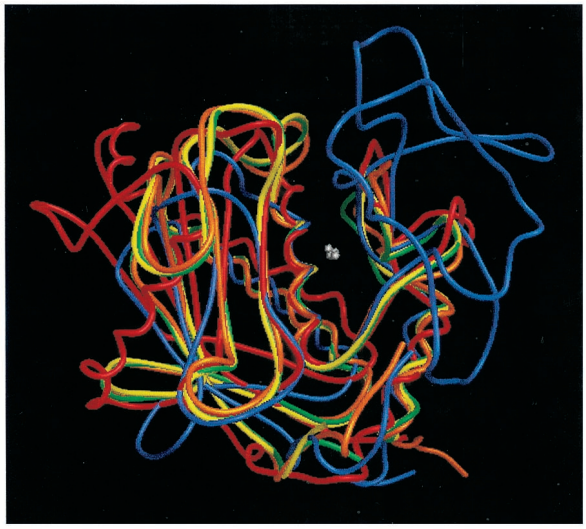
$$Sim(A_i,B_j) = \frac{\Omega_{A_iB_j}}{\sqrt{\Omega_{A_iA_i}} \cdot \sqrt{\Omega_{B_jB_j}}}, \tag{14}$$

where $Sim(A_i,B_j)$ corresponds to the residue-normalized similarity between the *i*-th and *j*-th residues of proteins *A* and *B*, respectively; i.e., $0 \leq Sim(A_i, B_j) \leq 1$. The following procedure is used to determine pairwise sequence alignments from their corresponding structure alignments. First, compute all pairwise inter-residue similarities of Protein *A* and Protein *B*, $Sim(A_i, B_j)$, using Equation 13 and Equation 14. Second, determine the residue in Protein *B* that is maximally similar to each residue in Protein *A*. This may result in some conflicts; that is, more than one residue in Protein *A* may have a maximum similarity relationship with the same residue in Protein *B*. Third, if no conflicts exist, then the two residues are considered to be matched. If a conflict exists, then the match is taken to be the two residues with the largest inter-residue similarity. This leaves some residues in Protein *A* that are now unmatched. Fourth, each unmatched residue in Protein *A* must now either be matched to a residue in Protein *B* or be considered to be part of a 'loop' in Protein *A* that has no matches to residues in Protein *B*; i.e., $Sim(A_i, B_j) \approx 0$. Note that the unmatched residues in Protein *A* should be taken in sequential order during the final phase of the sequence alignment procedure to ensure that crossovers among matched residues do not occur, since crossovers destroy the normal ordering of the residues. For example, consider the crossover that occurs when residue A_{32} is matched to residue B_{36}, denoted by $A_{32} \sim B_{36}$, and residue A_{33} is matched to residue B_{33}, denoted by $A_{33} \sim B_{33}$. This would lead to a shuffling in the order of residues on Protein *B* such that B_{36} would now come before B_{33}, which is incorrect. The above procedure yields a sequence alignment in which each pair of aligned residues are given a score equal to their inter-residue similarity. Several examples, based on the pairwise structural superposition of {1FNC,1MNC}, are given in the Results and Discussion section that illustrate the application of inter-residue similarity to structure-derived sequence alignment and to the analysis of 'similarity regions' in aligned proteins.

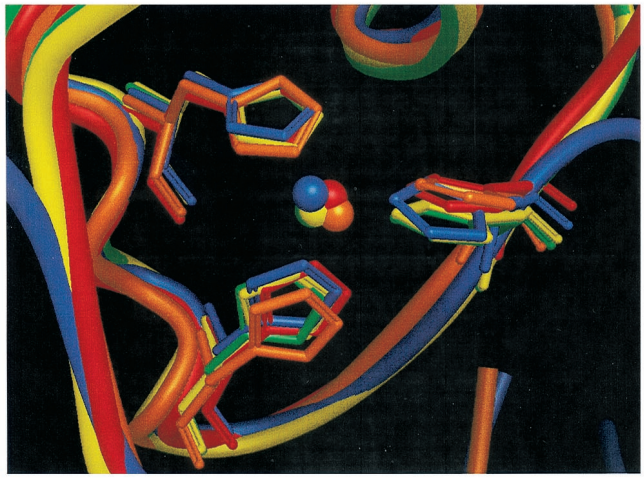
Table 1. Pairwise sequence and structural similarities

Proteins	Sequence Identity (Percent)	Pairwise Similarity (Percent)	Pairwise Similarity Multi-Molecule[a] (Percent)
1HFC-1MNC	66.5	85.4	85.4
1HFC-2SRT	62.5	54.2	54.1
1MNC-2SRT	57.1	54.8	54.7
1AST-2SRT	22.8	22.1	21.7
1HFC-1AST	18.9	25.3	25.2
1IAG-2SRT	16.5	23.9	22.9
1MNC-1AST	16.4	24.9	24.8
1IAG-1AST	14.1	17.4	16.5
1MNC-1IAG	13.4	26.1	25.8
1HFC-1IAG	12.4	25.9	25.9

[a]Pairwise structural similarity derived from a multi-molecule alignment.



(a)



(b)

Figure 3. (a) A multi-molecule alignment of five MMPs (1HFC, 1MNC, 2SRT, 1AST, 1IAG) corresponding to an overall structural similarity of about 35%. Note the catalytic zinc ion located within the region of the helix-turn-β-strand motif, which is well conserved for all five MMPs. (b) A closeup view of the multi-molecule alignment given in Figure 3a, showing details of the catalytic site including the zinc ion and the three His residues responsible its binding.

RESULTS AND DISCUSSION

Pairwise Alignments

Five matrix metalloproteinases (MMPs) from mammalian and fungal sources—Human Fibroblast Collagenase (1HFC), Neutrophil Collagenase (1MNC), Stromelysin (2SRT), Adamalysin (1IAG), and Astacin (1AST) are examined in this work. The results of all of the studies reported here were obtained using Gaussians placed on the mainchain atoms -$C_{\acute{a}}$-(C=O)-N- of each amino-acid residue, *i.e.*, $|A_i| = 4$, for $i = 1,2, \ldots, |A|$; the

```
1HFC  -PRWEQTHLT YRIENYTPDL PRADVDHAI EKAFQLWSNV TPLTFTKVSE
1MNC  GPKWERTNLT YRIRNYTPQL SEAEVERAI KDAFELWSVA SPLIFTGISQ

1HFC  GQADIMISFV RGDHRDNSPF DGPGGNLAH AFQPGPGIGG DAHFDEDERW
1MNC  GEADINIAFY QRDHGDGSPF DGPNGILAH AFQPGQGIGG DAHFDAEETW

1HFC  TNNFREYNLH RVAAHELGHS LGLSHSTDI GALMYPSYTF SG--DVQLAQ
1MNC  TNTSANYNLF FVAAHEFGHS LGLAHSSDP GALMYPNYAF RETSNYSLPQ

1HFC  DDIDGIQAIY GRS
1MNC  DDIDGIQAIY G--
```

Figure 4. A sequence alignment of 1HFC and 1MNC obtained from a multi-sequence match of MMPs (see Dhanaraj et al).[47] The part of the matched sequences marked lying within the box corresponds to the region where the structure-based alignment differs from that obtained from a sequence match (see text and Figure 7 for details).

zinc ion is not used in any of the alignments. The most similar MMPs, 1HFC and 1MNC, have a sequence identity of 66.5%, so that aligning these two proteins should not present a problem for virtually any structure-matching algorithm. The set of alignment solutions for this pair of proteins is shown in Figure 1a. From the figure it is apparent that the alignment solution with the maximum value, *Sim*(1FHC,1MNC) ≈ 0.82, stands well above all other solutions and corresponds clearly to the best alignment solution. This is not always the case, as is seen in the next example that deals with two proteins in the same family with a low level of sequence identity (*vide infra*). Figure 2a shows the structural alignment produced by MIMIC corresponding to the best alignment solution. Note the position of the catalytic zinc ion located in near the center of the helix-turn-β-strand motif in the foreground of the figure. As is seen in the figure, the best alignment solution affords an excellent match of the two proteins in this key helix-turn-β-strand region. The {1HFC,1IAG} pair possesses the lowest sequence identity of any of the pairs of proteins examined in this study, about 12% as indicated in Table 1. This value definitely lies in the "twilight zone" and should provide a stringent test for any structure alignment method. In this case, the best alignment solution has a considerably lower value, *Sim*(1HFC,1IAG) ≈ 0.26. As is seen in Figure 1b the two best alignment solutions, which are close in value, nevertheless stand out from the remainder of solutions, with the greater similarity value again corresponding to the best structural match as is seen in Figure 2b and Figure 2c. As was true in the case of the

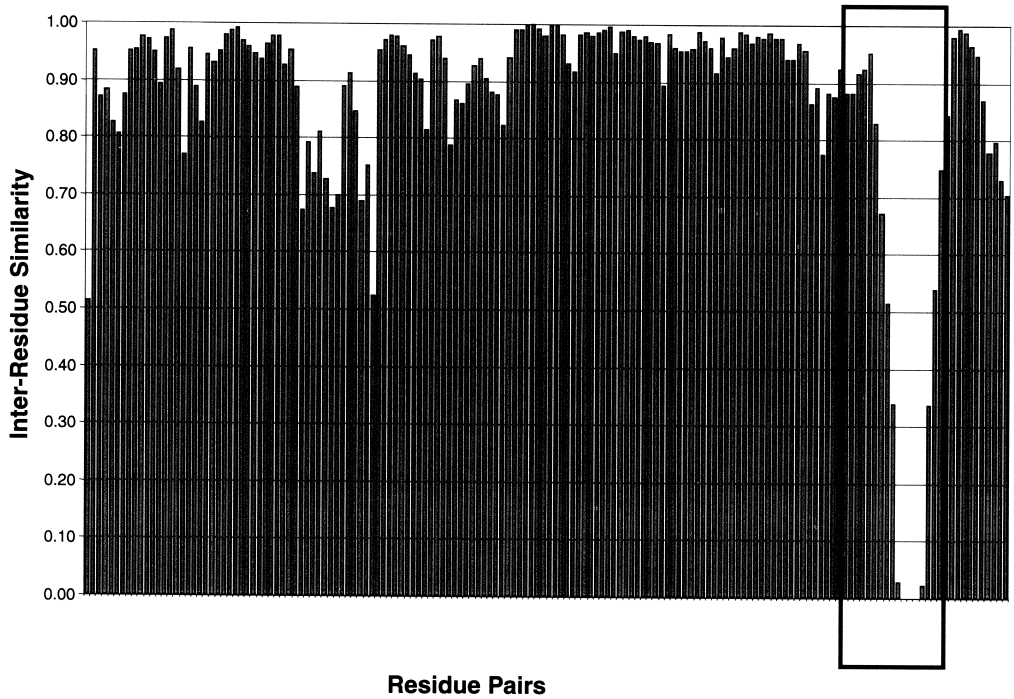


Figure 5. A histogram displaying the inter-residue structural similarities derived from the best pairwise alignment solution for 1HFC and 1MNC. The residues lying within the box correspond to those lying within the box in Figure 4 and shown in greater detail in Figure 6. See Figure 7 for a three-dimensional view of this region. Note that the N-terminus of the protein lies in the left most edge of the abcissa, and the residue numbers, which have been removed for clarity, increase from left to right.

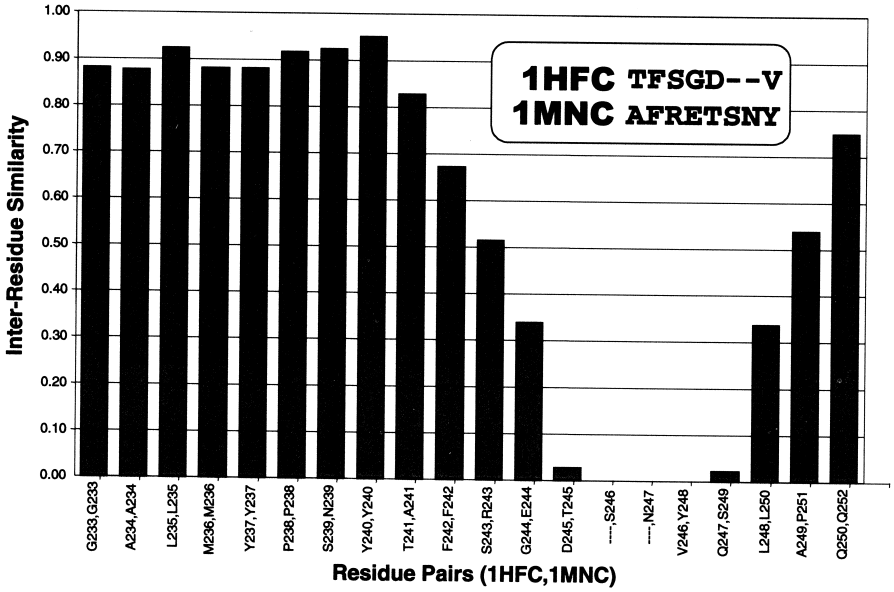


Figure 6. A closeup view of the region lying within the box in Figure 5. The sequence match in the box is derived from the structural alignment and differs from that shown in Figure 4.

{1HFC,1MNC} pair, the important helix-turn-β-strand motif is again structurally well matched. This observation is crucial since it indicates that important, conserved substructural regions can be identified even in the presence of considerable 'structural noise.' In contrast to the best alignment solution, the secondary solution, which corresponds to a similarity of about 0.19, shows improved alignment of the β-sheet regions of the two proteins, with a corresponding loss of alignment in the important helix-turn-β-strand motif. The results of all the pairwise structure alignments of the five MMPs examined in this work are summarized in Table 1. To investigate the possibility that the superpositioning of two structurally unrelated proteins might, nevertheless, lead to alignment solutions of significant magnitude, two proteins, trypsin (1TRY) [45] and 2ACT,[46] were superposed using MIMIC. As 1TRY is a β-protein and 2ACT is an $\alpha + \beta$ protein it is not expected that a structurally meaningful superposition can be obtained. Thus, it is also not expected that a significant similarity value will be obtained as well. This is definitely the case as is seen from the resulting alignment solutions shown in Figure 1c. From the figure, it is clear that no solutions stand out. Basically, all of them are of comparable and relatively small magnitude, and can be considered as 'background noise.'

Multi-Protein Alignments

Simultaneous structural alignment of multiple molecules was shown to be of importance for producing meaningful structural superpositions in the case of small molecules.[33,43,44] We performed a similar procedure to investigate whether the simultaneous alignment of multiple proteins would yield similar results to those obtained by pairwise alignments. Figure 3a shows the results obtained in this study for a simultaneous, multi-protein superposition of all five MMPs. A striking feature of the multi-protein alignment is the tight clustering of the catalytic zinc ions, which were not used in the alignment, in the important helix-turn-β-strand motif of the active site of the MMPs. This shows that multi-protein superpositioning does, indeed, produce meaningful structure alignments. A comparison of the pairwise similarity values obtained from a multi-molecule alignment with the corresponding values obtained solely by pairwise superpositioning are also given in Table 1. It is clear from the table that the values produced by either approach do not differ significantly from each other. Thus, for this set of proteins at least, it appears to make little difference whether pairwise or multi-molecule alignments are used, a situation in distinct contrast to that observed for small molecules,[43,44] where multi-molecule alignments produce much

Figure 7. The best pairwise alignment of 1HFC and 1MNC. The region marked in red corresponds to the residues depicted in the box in Figure 5 and the inset in Figure 6.

Table 2. Inter-residue similarities of specific 1HFC and 1MNC residues

Residue Number	1HFC Residue	1MNC Residue	Inter-Residue Similarity	BLOSUM Substitution Matrix[a]
132	H	R	0.988	0
135	E	K	0.959	1
143	N	V	0.890	−3
151	K	G	0.891	−2
160	M	N	0.977	−2
166	G	R	0.972	−3
180	N	I	0.989	−3
189	P	Q	0.932	−1
207	F	S	0.960	−3
208	R	A	0.954	−2

[a]The values of the BLOSUM amino acid substitution matrix are given by 'log odds' ratio rounded to the nearest integer.[50]

more consistent overall results. This is not to say, however, that the current data on five MMPs represent a general characteristic of protein molecules compared with small molecules. More work will have to be done to determine whether the observations here are truly general. Figure 3b shows a close-up view of the 'catalytic-zinc' binding site with its three key His residues, which are part of the helix-turn-β-sheet motif, depicted explicitly. As noted earlier, neither the catalytic zinc ions nor the His sidechains were used in the superpositioning procedure, clearly indicating that using only the four mainchain atoms, nevertheless, produces structurally meaningful alignments.

Post-Alignment Analysis—Structure-Based Sequence Alignment

An important result obtained from post-alignment analysis is the sequence match derived from the corresponding three-dimensional pairwise structure alignment of two proteins. As has been noted by Zu-Kang and Sippl,[7] structure-based sequence alignments are far more reliable than alignments based solely on sequence matching methods, even those utilizing multiple sequence matches (*vide infra*). A simple example, based upon two mammalian MMPs (1HFC and 1MNC) with greater than 66% sequence identity, illustrates the structure-based approach. Figure 4 depicts the sequence match of 1HFC and 1MNC obtained by Dhanaraj et al.,[47] based upon a multiple sequence match of seven MMP sequences. The residues lying within the box define the region where the structure-derived procedure yields results in disagreement with those obtained purely by sequence matching. The histogram in Figure 5 shows the inter-residue similarity values obtained using the algorithm described in the Methods section, where the residue labels are omitted from the figure for clarity. In distinction to the results obtained from purely sequence-based procedures, where the inter-residue similarities of matched residues all have value unity by default, the inter-residue similarities derived here from structure-based alignments tend to fluctuate. Due, however, to the high level of sequence identity of these two proteins, they also have substantial pairwise structural similarity (~85 percent, see Table 1). Also, as seen in Figure 2a, the mainchains of these proteins overlap significantly throughout their entire lengths. Thus, it is expected that most of the inter-residue similarities will be large. In fact, the mean and standard deviation of the entire set of inter-residue similarities are 87% and 17%, respectively. These values appear quite consistent with the histogram in Figure 5. Focusing on the region within the box in Figure 5, which corresponds to the region where the two mainchains separate, it is clear that the value of the inter-residue similarity drops precipitously. This is illustrated graphically in Figure 6, which also shows the particular inter-residue matchings derived from the structure-based sequence alignment along the abscissa. As is seen in the figure, two of the 1MNC residues, S246 and N247, are not paired with any residue along the 1HFC chain. Thus, the inter-residue similarities of these amino acids are zero with respect to any residue on the 1HFC chain. This is illustrated in Figure 7, where it is clear that the regions of the two chains depicted in inset, deviate significantly from one another. Table 2 compares the inter-residue structural similarity with the BLOSUM amino-acid substitution matrix developed by Henikoff and Henikoff[48] for a few selected residue pairs obtained from the alignment of the{1HFC,1MNC} pair. The larger the BLOSUM value the greater the likelihood that two residues are substitutes for one another. For example, a value of −4 indicates a very low probability that the either of the residues will substitute for the other. From the table it is clear that amino acids with low scores of, for example, ~ −3 nevertheless show reasonably high inter-residue structural similarities, generally greater than 0.96. Recall that the inter-residue structural similarities considered here relate only to mainchain atoms so that the nature of the sidechain does not come into play. Thus, Table 2 shows that the similarity of the chain fold can be retained even when amino acids of very different substitution character are located in comparable spatial positions in the aligned proteins. Inter-residue similarities corresponding to a given pairwise alignment can also be used to visualize regions of high inter-residue structural similarity by color coding the individual residues: bright red corresponds to the highest value, dark blue to the lowest value, and intermediate colors to similarities lying between these extremes. Results for the {1HFC,1MNC} pair are given in Figure 8a and Figure 8b. From the amount of red in the figure it is clear that both proteins are not only structurally

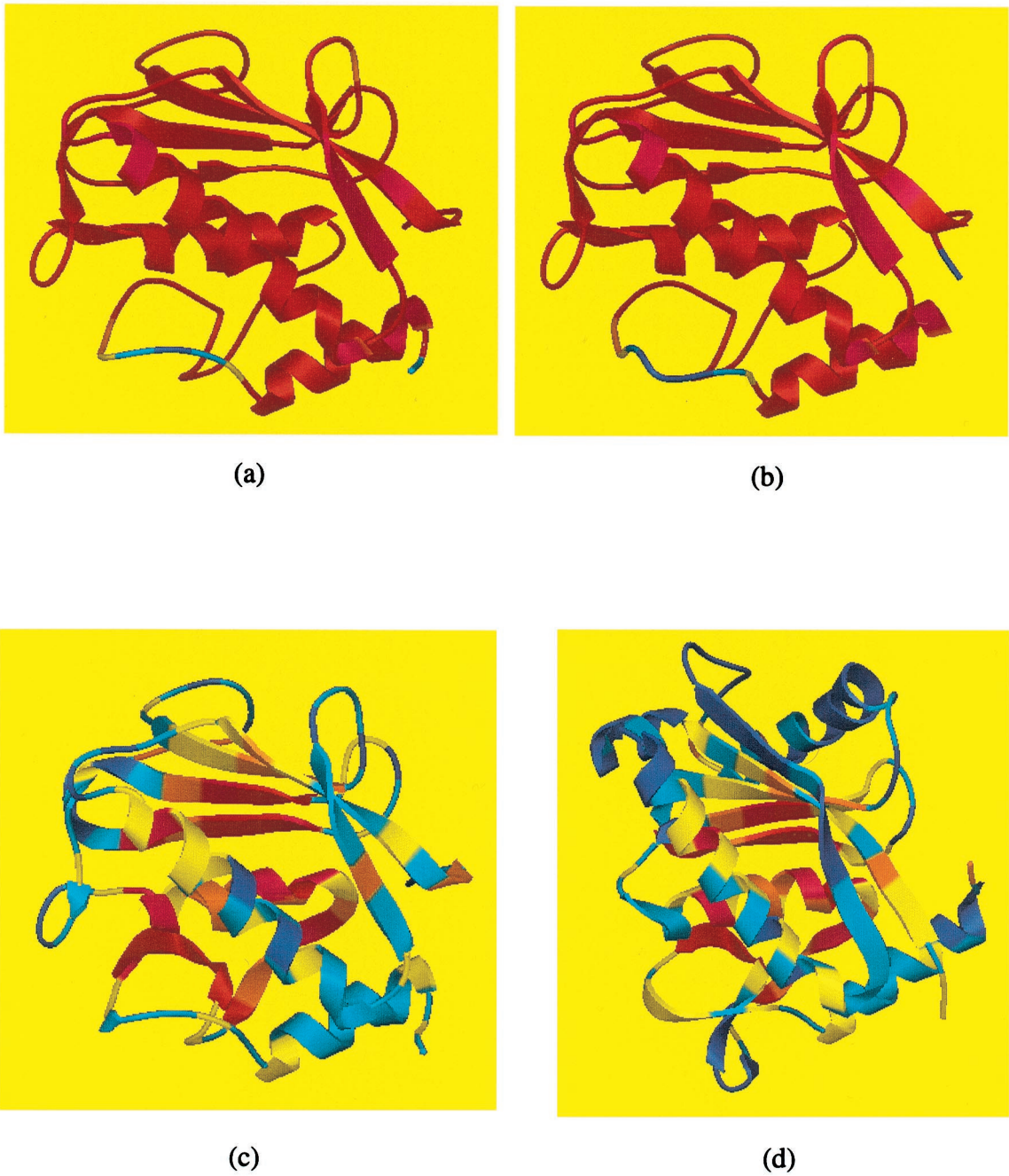


Figure 8. Inter-residue similarities derived from pairwise alignments. Red corresponds to the highest inter-residue similarity value and blue to the lowest. (a) Inter-residue similarities of 1HFC obtained from its pairwise alignment with 1MNC. (b) Inter-residue similarities of 1MNC obtained from its pairwise alignment with 1HFC. Note that the structurally-conserved helix-turn-β-strand motif is located in the background and lies below the prominent helix positioned diagonally across the figure in both (a) and (b). (c) Inter-residue similarities of 1HFC obtained from its pairwise alignment with 1IAG. (d) Inter-residue similarities of 1IAG obtained from its pairwise alignment with 1HFC. Note that the structurally-conserved helix-turn-β-strand motif is located in the background and lies below the prominent helix positioned diagonally across the figure in both (c) and (d). In contrast to the{1HFC,1MNC} pair, the helix-turn-β-strand motif in the{1HFC, 1IAG} pair is not as structurally-conserved.

similar overall, but also similar with respect to their mainchain geometries on a residue-by-residue basis. This is in accord with the histogram in Figure 5 for these same two proteins. In contrast, Figure 8c and Figure 8d depict the inter-residue similarities for the {1HFC,1IAG} pair, which has very low sequence identity (~12%) and structural similarity (~24%).

Here the inter-residue similarities, not surprisingly, are generally of lower value and much less uniform than was the case for the {1HFC,1MNC} pair. Note that the important helix-turn-β-strand motif lies in the background of the figure for both proteins. Its striking red color indicates that, as expected, it is well conserved with respect to these two proteins.

SUMMARY AND FUTURE WORK

This study describes the application of a Gaussian-based approach to the alignment of three-dimensional protein structures. The procedure employs a similarity function that essentially measures the overlap of the structures, each protein being represented by a set of Gaussians located on the four mainchain atoms of each amino-acid residue. The Gaussians provide a 'soft' representation of the shape or 'steric field' of the proteins; thus, the procedure is quite robust to small changes or differences in structure. This was illustrated by the ability of the procedure to identify the important, conserved substructural features or motifs for proteins within the same family even when the proteins possess low sequence identity and/or are of significantly different sizes. The form of the similarity function allows its 'decomposition' in terms of, for example, individual atoms or amino-acid residues. This characteristic makes it possible to carry out a number of post-alignment analyses. For instance, computing the inter-residue similarities provides the means for visually characterizing specific regions of high and low structure similarity among proteins and for deriving sequence alignments directly from similarity-based three-dimensional structural superpositions. As illustrated in the previous section, sequence alignments produced in this way do not always agree with those produced by pure sequence alignment methods, even those employing multiple sequence matching methods. The work presented here is based on the program MIMIC,[33] which was used initially for treating small molecules. Thus, many of its characteristics are not optimized for macromolecules. One of the most important characteristics, if large numbers of proteins are to be handled, is the time it takes to carry out an a priori single pairwise structural alignment, which is ~25 s per alignment. While this is not burdensome for small sets of proteins, it can become so when, say, a substantial fraction of the PDB is treated. Thus, efforts to speed up the alignment process are a top priority in ongoing efforts to continually improve MIMIC. With regard to post-alignment analyses, we intend to implement both molecular steric and electrostatic potential (MEP) similarity fields to provide enhanced visualization of molecular similarity. However, implementing an MEP similarity field procedure in MIMIC raises several issues that must be dealt with. It is well known that protein electrostatics are significantly influenced by the aqueous, ionic environment in which the protein resides. Thus, characterizing protein electrostatics *in vacuo* is not sufficient, and the influence of the solvent environment must be accounted for in some approximate manner. We intend to employ continuum electrostatics as embodied in the linear Poisson-Boltzmann equation[49] to compute the MEPs of the proteins and, from that, their MEP similarities. The second issue that must be dealt with arises from the fact that the MEP, unlike the steric-field function, is not positive definite. This will be dealt with by computing the MEP similarity for regions of positive MEP, negative MEP, and for regions where the MEPs on the two proteins are of opposite sign, as was done in the case of small molecules.[33,43] The method presented here for 'translating' the information in a three-dimensional structural alignment into its corresponding sequence alignment works, but it is difficult to automate and generalize beyond pairwise sequence alignments. Recently, it was pointed out to us[50] that the problem can be formulated in terms of a dynamic programming algorithm. This will be a high priority in our future work. An advantage of the dynamic algorithm approach is its possible generalization to handle multi-sequence matches in a manner similar to those used to treat multiple sequence alignments based only upon sequence data.

ACKNOWLEDGMENTS

The authors would like to thank Jim Blinn for his programming assistance and helpful comments on many aspects of this work.

REFERENCES

1 Mestres, J., Rohrer, D.C., and Maggiora, G.M. Gaussian-based approaches to protein structure similarity. In: *Molecular Modeling and Prediction of Bioactivity*, Gundertofte, K., and Jorgensen, F.S., Eds., Kluwer Academic Publishers, New York, 2000, pp. 83–88
2 Šali, A., and Blundell, T.L. The definition of general topological equivalences in protein structures: a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* 1990, **212**, 403–428
3 Overington, J.P. Comparison of three-dimensional structures of homologous proteins. *Curr. Opin. Struct. Biol.* 1992, **2**, 394–401
4 Holm, L., and Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 1993, **233**, 123–138
5 Orengo, C. Classification of protein folds. *Curr. Opin. Struct. Biol.* 1994, **4**, 429–440
6 Alexandrov, N.N., and Fischer, D. Analysis of topological and nontopological structural similarities in the PDB: New examples with old structures. *Proteins* 1996, **25**, 354–365
7 Zu-Kang, F., and Sippl, M.J. Optimum superposition of protein structures: ambiguities and implications. *Folding and Design* 1996, **1**, 123–132
8 Godzik, A. The structural alignment of proteins: Is there a unique answer? *Prot. Sci.* 1996, **5**, 1325–1338
9 Taylor, W.R., and Orengo, C. Protein structure alignment. *J. Mol. Biol.* 1989, **208**, 1–22
10 Subbarao, N., and Haneef, I. Defining topological equivalences in macromolecules. *Prot. Engineer.* 1991, **4**, 877–884
11 Rose, J., and Eisenmenger, F. A fast, unbiased comparison of protein structures by means of the Needleman-Wunsch algorithm. *J. Mol. Evol.* 1991, **32**, 340–354
12 Vriend, G., and Sander, C. Detection of common 3-dimensional structures in proteins. *Proteins* 1991, **11**, 52–58
13 Zhu, Z.-Y., Šali, A., and Blundell, T.L. A variable gap penalty function and feature weights for protein 3-D structure comparisons. *Prot. Engineer.* 1992, **5**, 43–51
14 Pascarella, S., and Argos, P. A databank merging related protein structures and sequences. *Prot. Engineer.* 1992, **5**, 121–137
15 Alexandrov, N.N., Takahashi, K., and Go, N. Common spatial arrangements of backbone fragments in homologous and nonhomologous proteins. *J. Mol. Biol.* 1992, **225**, 5–9

16 Russell, R.B., and Barton, G.J. Multiple protein sequence alignment from tertiary structure comparison assignment of global and residue confidence levels. *Proteins* 1992, **14**, 309–323
17 Bachar, O., Fischer, D., Nussinov, R., and Wolfson, H. A computer vision based technique for 3-D sequence independent structural comparisons of proteins. *Prot. Engineer.* 1993, **6**, 279–288
18 Grindley, H.M., Artymiuk, P.J., Rice, D.W., and Willett, P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.* 1993, **229**, 707–721
19 Alexandrov, N.N., and Go, N. Biological meaning, statistical significance, and classification of local similarities in non-homologous proteins. *Prot. Sci.* 1994, **3**, 866–875
20 Boutonnet, N.S., Rooman, M.J., Ochagavia, M.-E., Richelle, J., and Wodak, S.J. Optimal protein structure alignments by multiple linkage clustering: Application to distantly related proteins. *Prot. Engineer.* 1995, **8**, 647–662
21 Diederichs, K. Structural superposition of proteins with unknown alignment and detection of topological similarity using a six-dimensional search algorithm. *Proteins* 1995, **23**, 187–195
22 Alexandrov, N.N. SARFing the PDB. *Prot. Engineer.* 1996, **9**, 727–732
23 Schuchardt, J., Schneider, G., Reichelt, J., Schomburg, D., and Wrede, P. Local structural motifs of protein backbones are classified by self-organizing neural networks: An assessment of the significance of 3-D structural similarity. *Prot. Engineer.* 1996, **9**, 833–842
24 May, A.C.W. Pairwise iterative superposition of distantly related proteins. *Prot. Engineer.* 1996, **9**, 1093–1101
25 Falicov, A., and Cohen, F.E.A surface of minimum area metric for the structural comparison of proteins. *J. Mol. Biol.* 1996, **258**, 871–892
26 Koehler, R.T., Villar, H.O., Bauer, K.E., and Higgins, D.L. Ligand-based protein alignment and isozyme specificity of glutathione S-transferase inhibitors. *Proteins* 1997, **28**, 202–216
27 Carugo, O. and Eisenhaber, F. Probablistic evaluation of similarity between pairs of three-dimensional protein structures utilizing temperature factors. *J. Appl. Cryst.* 1997, **30**, 547–549
28 Brown, N.P., Orengo, C.A., and Taylor, W.R. A protein structure comparison methodology. *Computers Chem.* 1996, **20**, 359–380
29 Chew, P., Huttenlocher, D., Kedem, K., and Kleinberg, J. Fast detection of common geometric substructure in proteins. *J. Comput. Biol.* 1999, **6**, 313–325
30 Sauder, J.M., Arthur, J.W., and Dunbrack, R.L. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* 2000, **40**, 6–22
31 Szustakowski, J.D., and Weng, Z. Protein structure alignment using a genetic algorithm. *Proteins* 2000, **38**, 428–440
32 Shindyalov, I.N., and Bourne, P.E. An alternative view of protein fold space. *Proteins* 2000, **38**, 247–260
33 Mestres, J., Rohrer, D.C., and Maggiora, G.M. MIMIC: A molecular-field matching program. Exploiting the applicability of molecular similarity approaches. *J. Comput. Chem.* 1997, **18**, 934–954
34 Spurlino, J.C., Smallwood, A.M., Carlton, D. ., Banks, T.M., Vavra, K. J., Johnson, J. S., Cook, E.R., Falvo, J., Wahl, R.C., Pulvino, T. A., Wendoloski, J. J., and Smith, D.L. 1.56 A structure of mature truncated human fibroblast collagenase. *Proteins* 1994, **19**, 98–109
35 Stams, T., Spurlino, J.C., Smith, D.L., Wahl, R.C., Ho, T. F., Qoronfleh, M.W., Banks, T.M., and Rubin, B. Structure of human neutrophil collagenase reveals large S1′ specificity pocket. *Nat. Struct. Biol.* 1994, **1,** 119–123
36 Gooley, P.R., O'Connell, J.F., Marcy, A.I., Cuca, G.C., Salowe, S.P., Bush, B. L., Hermes, J.D., Esser, C.K., Hagmann, W.K., Springer, J.P., and Johnson, B.A. The NMR structure of the inhibited catalytic domain of human stromelysin-1. *Nat. Struct. Biol.* 1994, **1**, 111–118
37 Gomis-Ruth, F.X., Kress, L.F., and Bode, W. First structure of a snake venom metalloproteinase: a prototype for matrix metalloproteinases/collagenases. *EMBO J* 1993, **12** 4151–4157
38 Bode, W., Gomis-Ruth, F.X., Huber, R., Zwilling, R., and Stocker, W. Structure of astacin and implications for activation of astacins and zinc-ligation of collagenases. *Nature* 1992, **358,** 164–167
39 Bishop, C.M. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995
40 Grant, J.A., and Pickup, B.T. A Gaussian description of molecular shape. *J. Phys. Chem.* 1995, **99**, 3503–3510
41 Carbó, R., Leyda, L., and Arnau, M. How similar is one molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* 1980, **17**, 1185–1189
42 Maggiora, G.M., Petke, J.D., and Mestres, J. A general analysis of field-based molecular similarity indices. *J. Chem. Inf. Comput. Sci.* , submitted
43 Mestres, J., Rohrer, D.C., and Maggiora, G.M. A molecular-field-based similarity study of non-nucleoside HIV-1 reverse transcriptase inhibitors. *J. Comput.-Aided Molec. Design* 1999, **13**, 79–93
44 Mestres, J., Rohrer, D.C., and Maggiora, G.M. A molecular -field-based similarity study of non-nucleoside HIV-1 reverse transcriptase inhibitors. 2. The relationship between alignment solutions obtained from conformationally rigid and flexible matching. *J. Comput.-Aided Molec. Design* 2000, **14**, 39–51
45 Rypniewski, W.R., Hastrup, S., Betzel, C., Dauter, M., Dauter, Z., Papendorf, G., Branner, S., and Wilson, K.S. The sequence and X-ray structure of the trypsin from *Fusarium oxysporum. Protein Eng* 1993, **6**, 341–348
46 Baker, E.N., and Dodson, E.J. Crystallographic refinement of the structure of actinidin at 1.7 angstroms resolution by fast fourier least-squares methods. *Acta Crystallogr., Sect. A* 1980, **36,** 559
47 Dhanaraj, V., Ye, Q.Z. Johnson, L.L.. et al. X-ray structure of a hydroxamate inhibitor complex of stromelysin catalytic domain and its comparison with members of the zinc metalloproteinase family. *Structure* 1996, **4**, 375–386
48 Henikoff, S., and Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* 1992, **89**, 10915–10919
49 Gilson, M.K., and Honig, B. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins* 1988, **4**, 7–18
50 Sean Eddy – personal communication