Protein Engineering vol.11 no.9 pp.739–747, 1998

Protein structure alignment by incremental combinatorial extension (CE) of the optimal path

Ilya N.Shindyalov and Philip E.Bourne[1,2,3]

San Diego Supercomputer Center, PO Box 85608, San Diego, CA 92186, [1]Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 and [2]The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

[3]To whom correspondence should be addressed

A new algorithm is reported which builds an alignment between two protein structures. The algorithm involves a combinatorial extension (CE) of an alignment path defined by aligned fragment pairs (AFPs) rather than the more conventional techniques using dynamic programming and Monte Carlo optimization. AFPs, as the name suggests, are pairs of fragments, one from each protein, which confer structure similarity. AFPs are based on local geometry, rather than global features such as orientation of secondary structures and overall topology. Combinations of AFPs that represent possible continuous alignment paths are selectively extended or discarded thereby leading to a single optimal alignment. The algorithm is fast and accurate in finding an optimal structure alignment and hence suitable for database scanning and detailed analysis of large protein families. The method has been tested and compared with results from Dali and VAST using a representative sample of similar structures. Several new structural similarities not detected by these other methods are reported. Specific one-on-one alignments and searches against all structures as found in the Protein Data Bank (PDB) can be performed via the Web at http://cl.sdsc.edu/ce.html.

Keywords: alignment/protein structure/combinatorial extension

Introduction

There is no exact solution to the protein structure alignment problem, only the best solution for the heuristics used in the calculation. As such, various heuristic approaches has been suggested (for reviews see Holm and Sander, 1994; Gibrat *et al.*, 1996; Godzik, 1996). The crucial question then becomes, how good is a particular choice of heuristic and what is the computational cost of the search? In this work a new algorithm is proposed which is fast and robust in finding an accurate 3D structure alignment, including cases with low structure homology. Existing algorithms for structure alignment mostly attempt global optimization of the alignment path for some similarity measure using dynamic programming (Orengo *et al.*, 1992), Monte Carlo (Holm and Sander, 1993), 3D clustering (Vriend and Sander, 1991; Fischer *et al.*, 1992) or graph theory (Alexandrov, 1996). Dynamic programming approaches solve the optimization task exactly, but are dependent on the target function, which may not reflect critical information about the alignment of other parts of the molecule. Monte Carlo and 3D clustering algorithms allow a better choice of target function but are sensitive to the optimization protocol. Further, the search space for these algorithms may be large. The combinatorial extension (CE) algorithm proposed here provides a significant reduction in the search space and empirically establishes a reasonable target function for the heuristics used. The target function assumes that first, the alignment path is continuous when including gaps, and second, there is one optimal match. CE will not resolve so-called 'nontopological' similarities (Alexandrov and Fischer, 1996), where the order of polypeptide fragments in the structure alignment does not follow their order in the sequence. There is an analogous approach to CE previously used for multiple sequence alignments (Johnson and Doolittle, 1986).

Various protein properties can be used to measure structure similarity. Recently a comprehensive set of properties was introduced and tested for structure comparison (Ponomarenko *et al.*, manuscript submitted). Various properties can also be used with the combinatorial extension algorithm, for example: (i) structure superposition as rigid bodies; (ii) inter-residue distances, (iii) environmental properties (for example, exposure, secondary structure), (iv) conformational properties (for example, bond angles, dihedral angles and orientation with respect to the protein center of mass). This paper is limited to an analysis using (i) and (ii), however, the use of (iii) and (iv) is work in progress (Shindyalov,I.N. and Bourne,P.E., manuscript submitted).

Materials and methods

Definition of the alignment path

The alignment between two protein structures A and B of length n^A and n^B, respectively, is considered the longest continuous path P of AFPs of size m in a similarity matrix, S, of size $(n^A - m)\cdot(n^B - m)$ representing all possible AFPs that conform to the criteria for structure similarity. One of the following three conditions should be satisfied for every two consecutive AFPs i and $i+1$ in the alignment path:

$$p^A_{i+1} = p^A_i + m \text{ and } p^B_{i+1} = p^B_i + m \quad (1)$$

or

$$p^A_{i+1} > p^A_i + m \text{ and } p^B_{i+1} = p^B_i + m \quad (2)$$

or

$$p^A_{i+1} = p^A_i + m \text{ and } p^B_{i+1} > p^B_i + m \quad (3)$$

where p^A_i is the AFP's starting residue position in protein A at the i^{th} position in the alignment path; similarly for p^B_i. Condition (1) describes two consecutive AFPs aligned without gaps and conditions (2) and (3) represent two consecutive AFPs aligned with gaps inserted in proteins A and B, respectively.

Combinatorial extension of the alignment path

The alignment path is constructed from AFPs of fixed size m (8 is a reasonable choice as shown empirically, see below). That is, one fragment of length m from the first protein and another fragment from the second protein form a pair if they satisfy a similarity criterion described below. The first AFP starting the path can be selected at any position within the similarity matrix S, consecutive AFPs are added such that

conditions (1–3) are satisfied. To limit the gap size, conditions (2) and (3) are enhanced by the addition of the following two conditions, respectively:

$$p^A_{i+1} \leqslant p^A_i + m + G \quad (4)$$

and

$$p^B_{i+1} \leqslant p^B_i + m + G \quad (5)$$

where G is the maximum allowable size of the gap (30 is a reasonable choice as determined empirically, see below). Similarities which require gaps longer than G may be misrepresented or missed by the algorithm. Compute time is proportional to G, thus G should be kept as small as possible, but not too small as to have a negative impact on the search resolution.

Heuristics for similarity evaluation and path extension

There are several alternative alignment strategies that differ in computation time and accuracy. In the course of this study we limit the evaluation of similarity to the following three distance measures:

(i) distance D_{ij} calculated using an 'independent' set of inter-residue distances, where each residue participates *once and only once* in the selected distance set:

$$D_{ij} = \frac{1}{m}\left(\left|d^A_{p^A_i p^A_i} - d^B_{p^B_i p^B_i}\right| + \left|d^A_{p^A_i + m-1, p^A_j + m-1} - d^B_{p^B_i + m-1, p^B_j + m-1}\right| + \sum_{k=1}^{m-2}\left|d^A_{p^A_i + k, p^A_j + m-l-k} - d^B_{p^B_i + k, p^B_j + m-l-k}\right|\right) \quad (6)$$

(ii) distance D_{ij} calculated using a full set of inter-residue distances, where all possible distances except those for neighboring residues are evaluated:

$$D_{ij} = \frac{1}{m^2}\left(\sum_{k=0}^{m-1}\sum_{l=0}^{m-1}\left|d^A_{p^A_i + k, p^A_j + l} - d^B_{p^B_i + k, p^B_j + l}\right|\right) \quad (7)$$

(iii) r.m.s.d. obtained from structures optimally superimposed as rigid bodies using least-squares minimization (Hendrickson, 1979).

Where:

D_{ij} denotes the distance between two combinations of two fragments from proteins A and B defined by two AFPs at positions i and j in the alignment path where $i \neq j$ (Figure 1a). In the case of a single AFP, that is where $i = j$, the distance is given as D_{ii} (Figure 1b).
p^A_i denotes AFP's starting residue position in protein A at the i^{th} position in the alignment path; similarly for p^B_i.
d^A_{ij} denotes the distance between residues i and j in the protein A based on the coordinates of C_α atoms; similarly for d^B_{ij}.
m denotes the size of the fragment.

Distance measure (i) is used to evaluate the combination of two AFPs, one already in the alignment path and one to be added, and distance measure (ii) is used to evaluate a single AFP, i.e., how well two protein fragments forming an AFP match each other. Distance measure (iii) is used as the last step in selecting the few best alignments and for optimizing gaps in the final alignment (see the next section).

The following three major path extension strategies can be used when adding the next AFP to the alignment path:

(i) Consider all possible AFPs that extend the path and satisfy the similarity criteria.
(ii) Consider only the best AFP (described subsequently) which extends the path and satisfies the similarity criteria.
(iii) Use some intermediate strategy.

The first strategy defines an exhaustive combinatorial search for the optimal path, while the second strategy defines a limited search among the best paths. It is shown that the second strategy is sufficient to reveal structure similarities when combined with some path evaluation heuristics. The second strategy is far superior in performance.

Another important aspect is the selection of starting point for the alignment path. We normally consider all possible starting points in the similarity matrix S which satisfy the similarity criteria. In searching for the alignment of maximum length, all starting points not leading to an alignment of length greater than the length of the longest alignment found thus far are discarded. This saves computational time, but limits matches to one per polypeptide chain.

Extension of the alignment path is based solely on the distance criteria. Neither the size of the gap, nor the statistical significance of the alignment path is considered at this point in the analysis. The longest alignment path is now evaluated for statistical significance (represented as a z-score). This is done by evaluating the probability of finding an alignment path of the same length with the same or smaller number of gaps and distance from a random comparison of structures using a non-redundant set (Hobohm *et al.*, 1992). The alignment path for the random set is calculated in the same way as an alignment path for two structures of interest. The z-score of a particular alignment (z) is calculated by numerically solving Equation 8 for z using a normal distribution with an average value of 0 and a standard deviation of 1:

$$\rho(0_j 1, -z) = \rho(D^{\text{av}}_i, D^{\text{sd}}_i, D^{\text{obs}}) \cdot \rho(G^{\text{av}}_i, G^{\text{sd}}_i, G^{\text{obs}}) \quad (8)$$

where:
z is the z-score of alignment

$$\rho(\mu,\sigma,x) = \begin{cases} 2\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy, & \text{if } x < \mu \\ 1, & \text{otherwise} \end{cases}$$

i is the number of the AFP in the alignment path
D^{obs}, G^{obs} is the observed distance score (according to Equation 11) and number of gaps for the alignment path under consideration, respectively.
D^{av}_i, D^{sd}_i is the sample average and standard deviation, respectively for the distance score for paths of length i in a random comparison of structures taken using a variety of polypeptide chains and starting points.
G^{av}_i, G^{sd}_i is the sample average and standard deviation, respectively for the gap score for paths of length i in a random comparison of structures.
The following heuristics have been utilized in deciding whether a path should be extended. Decisions are made at three levels:

(i) single AFP
(ii) AFP against the path
(iii) whole path

This results in the following three conditions, respectively:

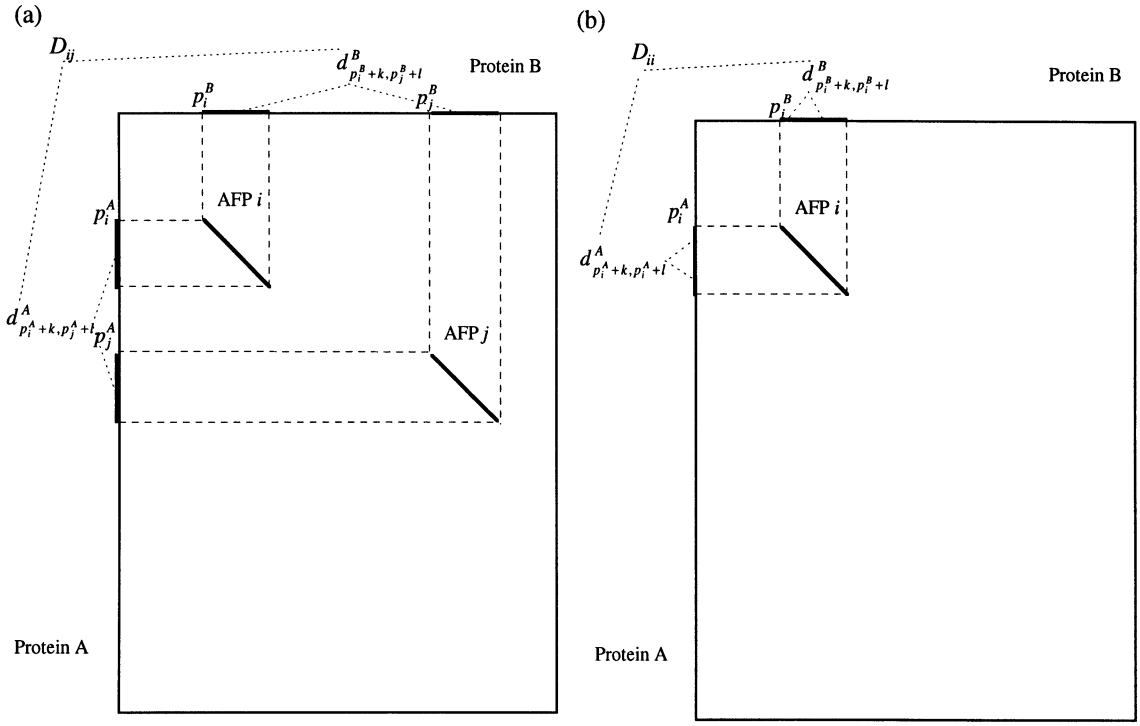

Fig. 1. Calculation of distance: (**a**) D_{ij} for alignment represented by two AFPs i and j from the path; (**b**) D_{ii} for single AFP i from the path.

$$D_{\mathrm{nn}} < D_0 \tag{9}$$

$$\frac{1}{n\text{–}1}\sum_{i=0}^{n-1} D_{in} < \mathrm{D}_1 \tag{10}$$

$$\frac{1}{n^2}\sum_{i=0}^{n}\sum_{j=0}^{n} D_{ij} < D_1 \tag{11}$$

where D_{ij} is the distance between aligned fragments defined by the AFPs i and j in the alignment path and n is the next AFP to be considered for addition to the alignment path of n–1 AFPs in length. D_0 and D_1 are similarity thresholds with typical values of D_0 = 3 Å and D_1= 4 Å.

It was shown empirically that the most accurate alignment occurs when the selection of the best AFP and extension of the path is done in three steps: (i) all candidate AFPs are selected based on condition 9; (ii) the best AFP is chosen based on condition 10; and (iii) the decision to extend or terminate the path is made based on condition 11.

Optimization of the final path

A final optimization has been added which contributes up to 2 Å improvement in the r.m.s.d. between two protein structures. It is only applied to alignments with z-scores above a certain threshold (normally 3.5) and is implemented in three steps: (i) the 20 best paths at the end of the search are evaluated based upon r.m.s.d. and the best one selected; (ii) each gap in this single alignment is evaluated for possible relocation in both directions up to *m/2* positions, where *m* is the AFP size, and if the r.m.s.d. of superimposed structures (Hendrickson, 1979) indicates improvement, then modified gap boundaries are adopted; and (iii) iterative optimization using dynamic programming (Needleman and Wunch, 1970) is performed on the distance matrix calculated using residues from the two superimposed structures. The gap penalty is 5 for initiation and 0.5 for extension with the elements of the distance matrix $M_{ij} = d_0 - d_{\mathrm{ij}}$, where d_0 is a constant at every optimization cycle. Optimization begins at d_0 = 2 and d_0 is incremented by 0.5 in every cycle. Optimization continues until either of two conditions is satisfied: (i) alignment length is less than 95% of alignment length before optimization; (ii) r.m.s.d. is less than 110% of r.m.s.d. at the cycle when condition i was first satisfied. Others have performed the same form of iterative optimization, for example see Feng and Sippl (1996). Terminal gaps have not been penalized. Iteration attempts to increase the alignment length found previously while keeping the r.m.s.d. at about the same level. Such optimization did not have significant impact on the computation time for database searches since it is performed only in a limited number of cases where the z-score is sufficiently high. A z-score of 3.5 and above corresponds to a probability of 10^{-3} or smaller. Thus, for a search sample of 1000 structures, one would match by chance. Given that the number of different folds in the PDB (Bernstein *et al.*, 1977) is estimated to be close to 1000 (Chothia, 1992), there is likely to be a single error or less in each search of the complete PDB.

Results

Test case: phycocyanin versus colicin A

The initial goal was to empirically determine the best values for various parameters used in a CE based structure comparison to balance accuracy and sensitivity against computational cost. We used the known structure similarity (Fischer *et al.*, 1996) between bacterial toxin colicin A (PDB code 1COL:A) and phycocyanin (PDB code 1CPC:L), a member of the globin superfamily, as a test case for the CE alignment algorithm (Tables I and II; Figures 2, 3 and 4). [Protein polypeptide chains are identified by their four character PDB codes followed by a colon and the chain identifier. Proteins with single,

Table I. Results of the structure alignment of phicocyanin (1CPC:L) to colicin A (1COL:A) using different alignment parameters

No.	Distance method	Path method	D_0 (Å)	D_1 (Å)	N^A	N^G	R.m.s.d. (Å)	Time (s)	Number of combinations
1	Ind.	Best	3.0	3.0	112	30	3.49	9	1 319 536
2	Ind.	Best	3.0	4.0	120	25	4.34	12	1 207 369
3	Ind.	Best	3.0	5.0	120	25	3.91	18	1 534 820
4	Ind.	Best	4.0	5.0	120	25	3.91	20	1 698 527
5	All	Best	3.0	4.0	120	27	4.16	84	1 300 003
6	All	Best	3.0	5.0	120	23	3.90	122	1 586 622
7	Ind.	Best	2.5	2.5	88	21	3.20	8	1 303 371
8	Ind.	All	2.5	2.5	112	29	3.25	16	2 901 811
9	All	Best	2.5	2.5	104	26	3.30	35	944 263
10	All	All	2.5	2.5	112	23	3.45	171	2 489 546

Distance method: Ind., only a subset of 'independent' distances (one for each residue) is used; All, all distances are used. Path method: Best, path extension with only the best scoring AFP is considered; All, all possible AFPs are considered. D_0, and D_1 are similarity thresholds in Å. N^A is the number of aligned positions. N^G is the number of non-aligned positions. R.m.s.d. is the difference in the two structures based on C_α positions after the optimization of gaps has been calculated. Time (s) is for execution on a single Sun Microsystems Ultra Sparc II processor (248 Mhz).

Table II. Results of a structure alignment of phicocyanin (1CPC:L) to colicin A (1COL:A)

No.	Fragment size (m)	N^A	N^G	R.m.s.d. (Å)	Time (s)	No. of combinations
1	4	116	32	3.72	40	2 759 440
2	6	114	35	3.62	29	1 930 654
3	8	120	25	3.91	20	1 207 369
4	10	120	26	3.89	16	924 544
5	12	120	26	4.07	18	875 217
6	16	120	29	3.68	15	780 596
7	24	96	16	4.15	8	263 117
8	36	108	22	4.33	7	17 551

For all calculations an 'independent' set of distances was used and the 'best' path. Similarity thresholds are $D_0 = 3.0$ Å and $D_1 = 4.0$ Å. N^A is the number of aligned positions. N^G is the number of non-aligned positions. R.m.s.d. is the difference in the two structures based on C_α positions after the optimization of gaps. Time (s) is for execution on a single Sun Microsystems Ultra Sparc II processor (248 Mhz).

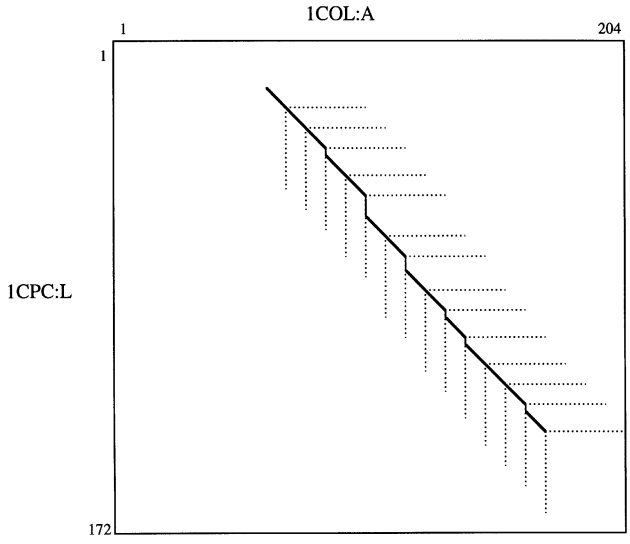

Fig. 2. Structure alignment of phycocyanin (1CPC:L) to colicin A (1COL:A). The solid line represents the optimal path built from AFPs. The dotted line represents the search area at every step of path extension.

unassigned polypeptide chains are represented by an underscore (_).] With polypeptide chain lengths of 204 and 172, respectively, this represents a typical example with respect to chain length. Tests with significantly variant chain lengths and large variations in known homology (not shown) did not cause the optimum values of these parameters to change significantly. [Test was performed on the following pairs of similar protein chains from (Fischer *et al.*, 1996): (1MDC:_, 1IFC:_), (1BBH:A, 2CCY:A), (2SAS:_, 2SCP:A), (1AAJ:_, 1PAZ:_), (2OMF:_, 2POR:_), (3CD4:_, 2RHE:_), (1TAH:A, 1TCA:_), (1BGE:B, 2GMF:A), (2SIM:_, 1NSB:A).] A larger value of m for long chains would reduce search times, but at the price of sensitivity. As will be shown subsequently search times with $m = 8$ balance sensitivity against computational cost and yet still permit interactive access via a Web site. The following properties of the alignment algorithm have been considered: (i) independent versus the full set of inter-residue distances; (ii) all possible AFPs versus the best AFP in the path extension; (iii) various similarity thresholds D_0 and D_1 (ranging from 2.5 to 5.0 Å); (iv) various fragment sizes m (4–36). All the results presented in Tables I and II are calculated without optimization of the final path so as to evaluate the capabilities of a pure AFP-based alignment (Figure 2). After optimization of the final path (described subsequently) the alignment calculated

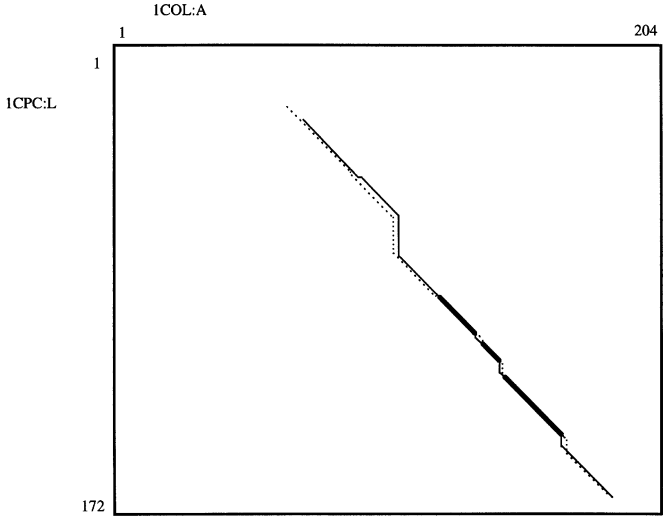

Fig. 3. Structure alignment of phycocyanin (1CPC:L) to colicin A (1COL:A). The thick solid line represents alignment overlap both before and after optimization. The dotted and thin solid lines represent alignments found before and after optimization, respectively, where they do not overlap.

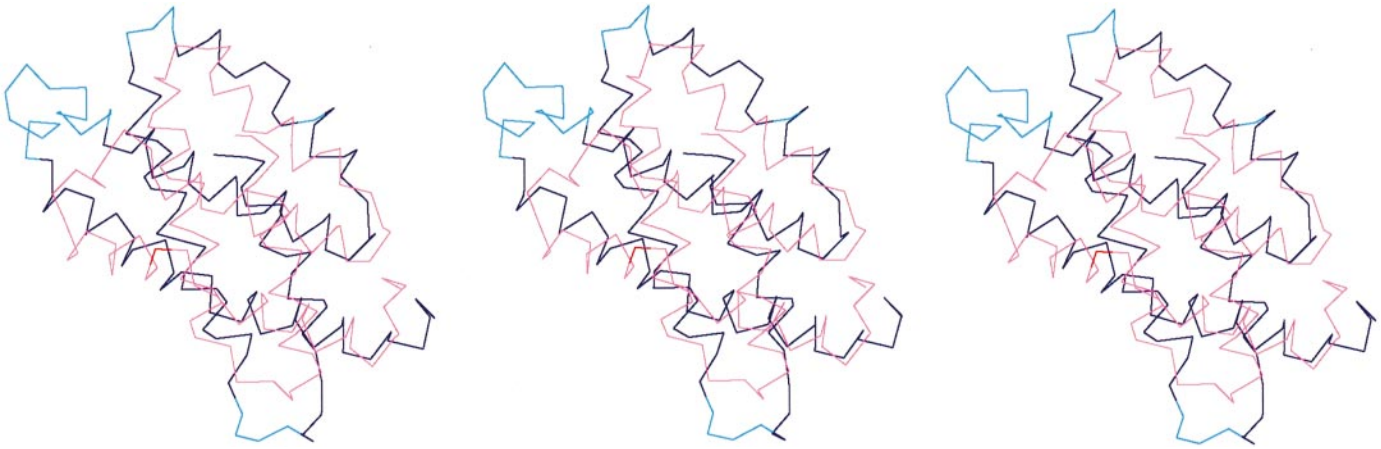

Fig. 4. Stereo projection of the final alignment and superposition of phycocyanin (1CPC:L) to colicin A (1COL:A). Phycocyanin is in dark blue and colicin A in purple. The offset colors, light blue and red, respectively, represent insertions in the respective structure.

Table III. Search for similarity to the quaternary complex of cAMP-dependent protein kinase (**1ATP:E**) using the PDB (September 1997)

No.	Chain (size)	N^A	N^G	R.m.s.d. (Å)	Z Score
1	1APM:E(350)	336	0	0.3	7.9
2	1CDK:A(350)	336	0	0.4	7.9
3	1YDR:E(350)	336	0	0.5	7.9
4	1CTP:E(350)	303	0	1.5	7.4
5	1PHK:_ (298)	255	28	2.5	7.2
6	1KOA:_ (491)	258	20	2.7	7.1
7	1KOB:A(387)	260	20	2.8	7.1
8	1AD5:A(438)	237	31	2.5	7.0
9	1CKI:A(317)	260	47	2.8	6.9
10	1CSN:_ (298)	249	37	2.4	6.8
11	1ERK:_ (364)	254	55	2.6	6.8
12	1FIN:A(298)	253	69	2.2	6.8
13	1GOL:_ (364)	254	55	2.6	6.8
14	1JST:A(298)	253	69	2.4	6.7
15	1IRK:_ (306)	244	69	3.3	6.5
16	1FGK:A(310)	251	54	3.5	6.2
17	1FMK:_ (452)	245	19	2.8	6.2
18	1WFC:_ (366)	240	72	3.1	5.6
19	1KNY:A(253)	112	79	4.3	3.9
20	1TIG:_ (94)	54	3	4.2	3.9

Chains with identical sequences were excluded, i.e., one protein chain representing all chains with identical sequences was used as follows: **1ATP:E** for (2CPK:E); 1**CDK:A** for (1CDK:B, 1CMK:E); **1YDR:E** for (1YDS:E, 1YDT:E); **1KOB:A** for (1KOB:B); **1AD5:A** for (1AD5:B, 2HCK:A, 2HCK:B); **1CKI:A** for (1CKI:B, 1CKJ:A, 1CKJ:B); **1CSN:_**(2CSN:_); **1FIN:A** for (1FIN:C, 1HCK:_, 1HCL:_); **1JST:A** for (1JST:C 1JSU:A); **1FGK:A** for (1FGK:B). N^A is the number of aligned positions. N^G is the number of non-aligned positions.

with $D_0 = 3$ Å and $D_1 = 4$ Å yields a r.m.s.d. of 3.25 Å for a sequence length of 116 residues (Figure 3). Figure 2 illustrates search zones at every step of the combinatorial extension of the path. Figure 3 represents paths before and after optimization. It is interesting to note that although both paths look very similar the r.m.s.d. between the paths is 1.37 Å, while the number of matches differs only by four. Figure 4 shows a stereo projection of the C_α traces after final alignment.

Comparison of independent and full sets of inter-residue distances indicates no significant difference in r.m.s.d. For example, for the case $D_0 = 3$ Å and $D_1 = 4$ Å (items 2 and 5 in the Table I) the difference in r.m.s.d. is only 0.2 Å. Similarly, there was no significant difference in the case $D_0 = 3$ Å and $D_1 = 5$ Å (items 3 and 6 in Table I). However, the difference in computation time was more than sixfold in both cases. Thus, use of the independent set of distances provides a significant advantage in computation time for little loss in accuracy.

A comparison using the best AFP versus all possible AFP combinations (items 7 versus 8 and 9 versus 10 in Table I) indicates some loss in the number of aligned positions for a two- to fivefold increase in speed.

The effect of different threshold values, D_1, is seen from a comparison of items 1, 4 and 7 from Table I. The length of the alignment increases from 88 to 120 when the threshold increases from 2.5 to 5.0 Å. At the same time the r.m.s.d. goes up from 3.2 to 4.3 and then goes down again to 3.9. The latter decrease in r.m.s.d. can be explained as follows. An alignment between two structures includes areas of high and low similarity and at a certain point on the alignment path the overall similarity may drop below the threshold leading to a termination of the path. A higher D_1 may allow the detection of a path which has been terminated in the middle at a lower threshold value but which overall has a better r.m.s.d.

Increasing D_0 and D_1 increases the computational cost, but finds longer alignments. The choice of higher thresholds offsets the effect of analyzing less combinations (compare items 1 and 8 in Table I), and results in a loss of only 0.2 Å in accuracy. The loss in accuracy associated with the choice of a faster search strategy is eventually compensated by optimization of the final path, resulting in an accuracy of 3.2 Å for the final alignment. Optimization of the final path would also help in those cases where the optimal alignment has dissimilarity in the middle. While the complete path is not detected during path extension, it will likely be completed during the optimization stage.

The effect of different AFP sizes is illustrated in Table II. With longer AFPs there are fewer places where gaps can be adopted which decreases the accuracy, but also the computational cost. With shorter AFPs the significance of a single AFP and of a comparison between two AFPs drops because shorter fragments have a higher chance of an accurate match. Empirically it has been determined that a fragment of size 8 balances

both requirements and has an acceptable computational cost. As stated previously, a dynamic AFP size (m) could be used to balance speed of computation against the sensitivity of finding an alignment. However, this is not deemed necessary for searching the current PDB.

Detecting members of the same protein family

The goal was to determine how well CE could detect members of the same protein family using a single member as a target probe. We used as a test the protein kinases for which over 30 structures are available in the PDB (Smith *et al.*, 1997).

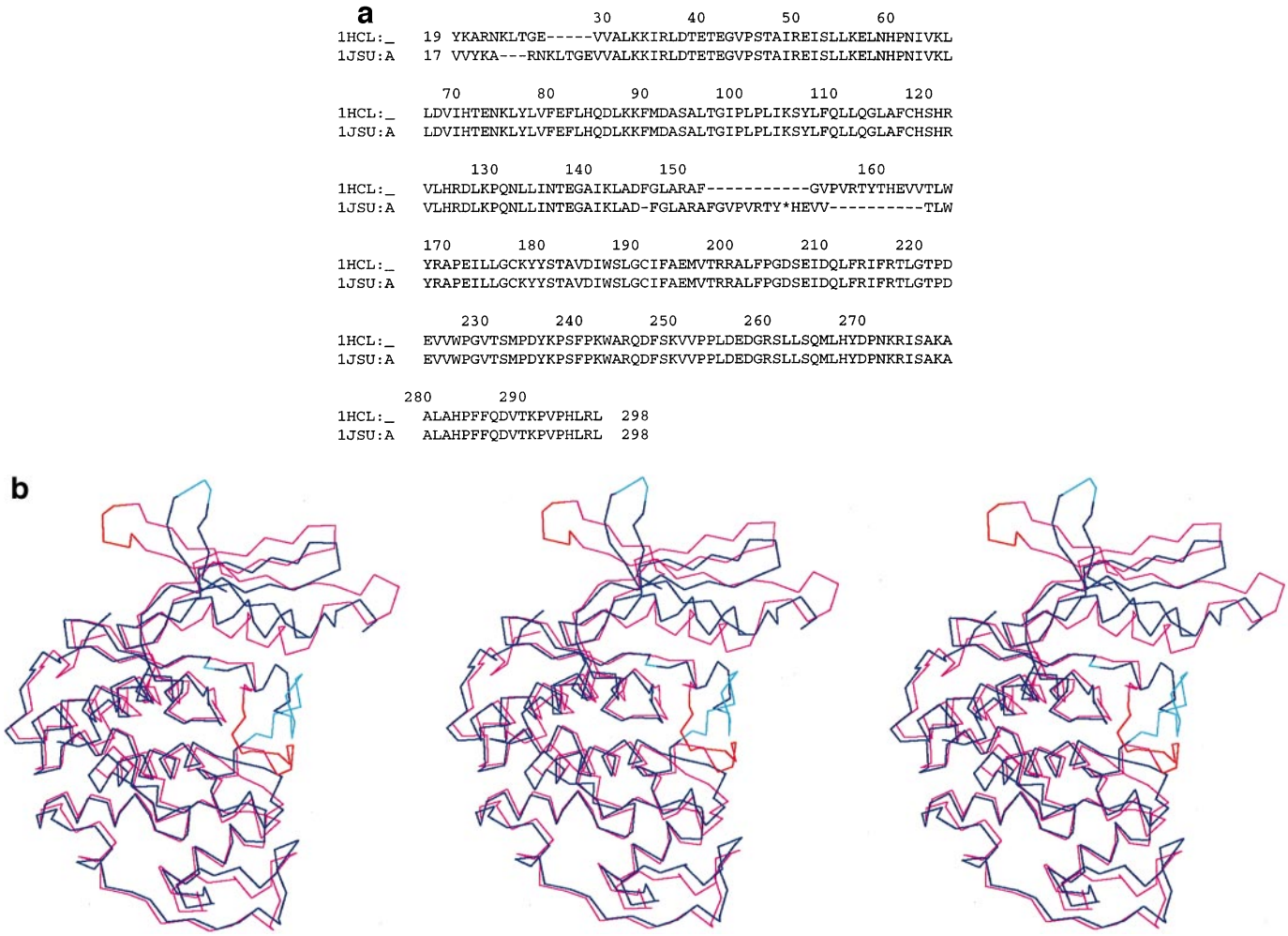

Fig. 5. Protein structure alignment and superposition of the catalytic subunit of two cyclin-dependent protein kinases in the open (1HCL:_) and closed (1JSU:A) conformations, (**a**) sequence (**b**) stereo projection of the structures. The open conformation is in purple and the closed conformation in blue. The offset colors, red and light blue, respectively, represent insertions in the respective structure.

Table IV. Selected results of a search for similarity with selected polypeptide chains (Chain 1) which are classified by SCOP (Murzin *et al.*, 1995) as a '4-helical up-and-down bundle'

Chain 1 query (size)	Chain 2 result (size)	N^A	N^G	R.m.s.d. (Å)	N^H	SCOP 'fold'	SCOP 'protein'
2ASR:_(142)	1OCC:C (261)	124	20	2.7	4	Not assigned	Cytochrome c oxidase from bovine
2ASR:_ (142)	1MMO:D (512)	116	9	5.9	4	Ferritin-like	Methane monooxygenase hydrolase, β and α subunits
2ASR:_ (142)	2BRD:_ (247)	116	13	4.3	4	Membrane all-alpha	Bacteriorhodopsin
256B:A (106)	1AEP:_ (161)	92	6	4.1	4	Apolipophorin – III	Apolipophorin-III
256B:A (106)	1CIY:_ (590)	93	17	4.3	4	Toxins' membrane translocation domains	Delta-endotoxin (insectocide), N-terminal domain
256B:A (106)	1AGS:A (221)	94	14	5.2	4	Glutathione *S*-transferases, C-terminal domain	Glutathione *S*-transferase
2ASR:_ (142)	1LKI:_ (180)	92	19	5.5	4	4-helical cytokines	Leukemia inhibitory factor (LIF)
2ASR:_ (142)	1FPS:_ (348)	118	43	5.0	4	Isoprenyl diphosphate synthases	Farnesyl diphosphate synthase

N^A is the number of aligned positions, N^G is the number of non-aligned positions and N^H is the number of aligned helices.

The results of a search against the complete PDB using the quaternary complex of cAMP dependent protein kinase in a closed conformation (1ATP:E) as a probe structure is presented in Table III. These results were subsequently compared to VAST (Madej *et al.*, 1995) and Dali (Holm and Sander, 1993) using databases of August, 1997. All three methods find the major family members. Differences result from recent structures being absent from Dali and VAST (the database used by CE is updated nightly from the PDB archives), for example, 1WFC:_. Further minor differences are found in the relatively weak similarities, which are not part of the family, but may be interesting from a biological perspective, for example, other nucleotide binding proteins. Three examples of structure homology found by one of the three methods, but not the other two are:

CE: 1KNY:A(253) – 112/4.3 Å; 1TIG:_ (94) – 54/4.2 Å;
VAST: 1MAE:H(373) – 56/2.9 Å; 2BBK:J(355) – 58/2.8 Å
Dali: 2MHR(118):_ – 82/3.9 Å; 2BRD:_ (221) – 100/4.0 Å

(The overall length of the polypeptide chain is given in parentheses followed by the number of residues that match and the r.m.s.d. for that match.) Implications of these findings, that is, differences in the twilight zone of structural similarity are discussed subsequently.

Comparing members of the same protein family

Within a given protein family it is possible to highlight similarities and differences in structural features with CE. Figure 5 illustrates the structural alignment of two cyclin-dependent protein kinases, the uncomplexed monomer (1HCL:_) in the open state and the complex with cyclin and P27 (1JSU:A) in the closed state. While the sequences of the uncomplexed and complexed state are almost identical with 96.2% homology (Figure 5a), there are significant conformational differences (Figure 5b). Differences are found in both the active site (center of Figure 5b) where P27 mimics ATP binding and in the small N-terminal lobe (top of Figure 5b) where P27 binds in an extended conformation (Russo *et al.*, 1996).

Detecting a protein fold

The goal was to recognize a particular protein fold, namely a 4-helical up-and-down bundle. Proteins known to exhibit that fold were chosen using the 'structure classification of proteins' (SCOP) resource (Murzin *et al.*, 1995). The fold is represented in SCOP by a set of 54 protein chains, subsequently limited to 24 based on sequence identity. (**1NFN:_**, **1LPE:_**, **1NFO:_**, **1LE2:_**, **1LE4:_**, **2ASR:_**, **1WAS:_** (1WAT:A, 1WAT:B), **2LIG:A** (2LIG:B, 1LIH:_), **256B:A** (256B:B, 1APC:_), **2CCY:A** (2CCY:B), **1BBH:A** (1BBH:B), **1CGN:_**, **1CGO:_**, **1CPQ:_** (1NBB:A, 1NBB:B, 1RCP:A, 1RCP:B), **1CPR:_**, **2HMQ:A** (2HMQ:B, 2HMQ:C, 2HMQ:D, 2HMZ:A, 2HMZ:B, 2HMZ:C, 2HMZ:D), **1HMD:A** (1HMD:B, 1HMD:C, 1HMD:D, 1HMO:A, 1HMO:B, 1HMO:C, 1HMO:D), **1HRB:_**, **2MHR:_**, **2TMV:P**, **1VTM:P**, **1CGM:E**, **1BUC:A** (1BUG:B), **3MDD:A** (3MDD:B, 3MDE:A, 3MDE:B). Representative set of '4-helical up-and-down bundle' chains is shown in bold, all other chains with identical sequence are given in parenthesis.) The search was performed using a probe subset of three query chains: the ligand binding domain of the aspartate receptor (2ASR:_), cytochrome B562 (256B:A), and the tobacco mosaic virus viral coat protein (2TMV:P). All 24 target chains were found with a z-score of 4.0 and above. Additionally, a number of polypeptide chains classified as different folds by SCOP had significant similarity to the 3 probe polypeptide chains (Table IV). Similarities involving chains not yet classified in SCOP are excluded. Similarities are to polypeptide chains in a closely related classification. For example, apolipophorin-III (1AEP:_) is a five-helix bundle, where four of the helices overlap and the order of the helices correspond to the probe structure. These similarities support the notion of a small number of classifications (Orengo *et al.*, 1994) where each distinct classification (e.g., all alpha) has a spectrum of folds (e.g., 4- and 5-helix bundles).

Table V. Examples of similarities found by CE and not detected by Dali (Holm and Sander, 1993) and VAST (Madej *et al.*, 1995)

No.	Chain 1(size)	Chain 2(size)	N^A	N^G	R.m.s.d. (Å)	Z Score
1	1LIS:_ (136)	1CIY:_ (590)	112	20	4.0	5.3
2	1CFP:A(92)	4ICB:_ (76)	64	9	2.6	4.2
3	1RPA:_ (342)	1HIW:A(133)	72	19	3.5	4.2
4	1HYP:_ (80)	1MZM:_ (93)	72	16	3.7	4.1
5	1CLC:_ (639)	1HOE:_ (74)	64	17	3.4	3.9
6	1UTG:_ (70)	1NOX:_ (205)	56	2	3.4	3.9
7	1FAR:_ (52)	1PTQ:_ (50)	40	4	1.8	3.7
8	1KUM:_ (108)	1TUL:_ (108)	64	16	3.6	3.7
9	1PYI:A(96)	1PYC:_ (71)	40	1	2.3	3.7
10	1VIH:_ (71)	1PYT:A(94)	56	8	3.2	3.7

N^A is the number of aligned positions and N^G is the number of non-aligned positions.

Table VI. Comparison of structure alignments for 10 'difficult' structures from (Fischer *et al.*, 1996) obtained by three methods: **Dali** (Holm and Sander, 1993), **VAST** (Madej *et al.*, 1995), and **CE**

No.	Chain 1(size)	Chain 2(size)	VAST N^A/r.m.s.d.(Å)	Dali N^A/r.m.s.d.(Å)	CE N^A/r.m.s.d.(Å)
1	1FXI:A	1UBQ:_	48/2.1	–	–
2	1TEN:_	3HHR:B	78/1.6	86/1.9	87/1.9
3	3HLA:B	2RHE:_	–	63/2.5	85/3.5
4	2AZA:A	1PAZ:_	74/2.2	–	85/2.9
5	1CEW:I	1MOL:A	71/1.9	81/2.3	69/1.9
6	1CID:_	2RHE:_	85/2.2	95/3.3	94/2.7
7	1CRL:_	1EDE:_	–	211/3.4	187/3.2
8	2SIM:_	1NSB:A	284/3.8	286/3.8	264/3.0
9	1BGE:B	2GMF:A	74/2.5	98/3.5	94/4.1
10	1TIE:_	4FGF:_	82/1.7	108/2.0	116/2.9

Table VII. Timings for CE on a single Sun Microsystems Ultra Sparc II processor (248 Mhz)

Structure 1 (length)	Structure 2 (length)	Sequence homology (%)	No. of positions aligned	R.m.s.d. (Å)	z-Score	Time (s)
1BPI:_ (58)	1BUN:B (61)	34	55	1.5	4.7	<1
1BPI:_ (58)	5EBX:_ (62)	2.5	40	5.3	2.3	<1
1WAJ:_ (903)	1NOY:A (388)	61	337	1.6	7.2	298
1WAJ:_ (903)	1BDP:_ (592)	7.0	143	3.2	4.4	860

Comparison with other Web accessible 3D structure comparison methods: Dali and VAST

CE was compared to results from Dali as found in FSSP (Holm and Sander, 1996) and VAST as found in MMDB (Hogue *et al.*, 1996), two Web-accessible resources for comparing the 3D structure of proteins.

The first experiment used a set of unique structures assigned by Dali (using FSSP results of August 27, 1997). Using CE we searched for similarities within this set and compared the search results to assignments made by VAST. Most similarities found by CE and not found by Dali and VAST are relatively short alignments (less than 100 residues) and often involve small proteins (Table V). There are a number of cases in Dali where very short proteins or protein fragments are interpreted as unique structures, but which can be easily aligned with an r.m.s.d. of ~1.0 Å and no gaps to other unique structures detected by Dali, for example:

1LYP:_ (32) versus 1OLG:A(42) – 32/0.7 Å
1HLE:B(31) versus 2ACH:B(40) – 24/0.7 Å
1BBT:4(85) versus 1TMF:4(31) – 24/1.0 Å
1AIE:_ (31) versus 2FUA:_ (215) – 24/0.6 Å
1CPT:_ (412) versus 1FCT:_ (32) – 24/1.3 Å
1LBD:_ (282) versus 1PSM:_ (38) – 32/1.6 Å
4ICB:_ (76) versus 1CTD:A(36) – 32/1.7 Å

There are also several apparent misinterpretations by Dali where polypeptide chains with identical sequences and an r.m.s.d. $<$2.0 Å are reported as unique structures for example:

2SEC:I(71) versus 1EGP:A(45) – 34/1.2 Å
1SCE:A(112) versus 1PUC:_ (105) – 93/1.3 Å

The second experiment involved 10 'difficult' similarities from a representative sample of known structurally related proteins (Fischer *et al.*, 1996; Table VI). Given a typical z-score cut-off value for each method (i.e., 3.5 for CE), nine out of the 10 similarities was found with CE, eight with Dali and VAST. The number of matched positions for CE differed on average by 14% from Dali and VAST. In eight of a total of 15 cases the number of aligned residues was larger than that reported by either Dali or VAST. Similarly, the r.m.s.d. was smaller or the same in 7 of the 15 cases. For all methods, the larger the number of matched positions, the larger the r.m.s.d.

Discussion

These results demonstrate that the combinatorial extension (CE) approach yields a good search resolution for finding structural similarity in proteins. Depending on the search criteria, the length of a structural match can be extended or shortened with a corresponding increase or decrease, respectively, in r.m.s.d. CE and the associated heuristics reported here provide a structural search capability comparable to other Web-accessible methods, namely Dali and VAST. Differences come between structures with low structure similarity, which, from a functional perspective, may be the most interesting. The task becomes one of elevating comparisons that currently reside in the noise, from pure geometric-based comparison methods, to the top of the list based on additional information. Until then users should work with all available geometric methods and carefully sift the twilight zone for interesting geometric comparisons.

Finding a geometric similarity in 3D structure does not necessarily imply functional similarity, it may only imply an energetically favorable conformation. In fact a dissimilarity in 3D structure may imply biological significance as was illustrated for the cyclin-dependent protein kinases (Figure 5). Part of the enzyme's active site had undergone a conformational change between inactive and active forms. Using the inactive form as the probe may not reveal the active form of the enzyme as a target based on pure geometric considerations unless care is taken with the search criteria, which implies some a priori knowledge about the system under study. To search for biologically meaningful alignments requires that other comparable properties be taken into account. A step in this direction has been taken (Ponomarenko *et al.*, manuscript submitted) where a total of 495 geometric and physicochemical properties have been considered, but using a different similarity criteria. Another approach is to apply the CE algorithm to this larger set of properties, including the appropriate heuristics. This is work in progress and some initial results are reported elsewhere (Shindyalov,I.N. and Bourne,P.E., manuscript submitted).

Some performance metrics using different alignment parameters are given in Table I. The time to align two structures is highly dependent on the length of the polypeptide chains and the length of the overlap established from the optimal path. Table VII provides timings for several test cases. A random sample of 100 structures returned an average alignment time of 20 seconds per structure using a single Sun Microsystems Inc. Ultra Sparc II processor (248 Mhz). These times are short enough for interactive Web access.

One-on-one structure alignment using CE is available via the Web at http://cl.sdsc.edu/ce.html. Users may select complete or partial polypeptide chains from the PDB, or upload their own coordinates in PDB format. Statistics for the alignment are returned along with the sequence alignment resulting from the structure alignment. Users may download atomic coordinates of the superimposed structures or view them using Rasmol (Sayle and Milner-White, 1995) as a helper application called from a Web browser. A Java applet (Compare3D) is also available for more detailed analysis.

An all-by-all comparison similar to that used in constructing the latest FSSP (Holm and Sander, 1998) is complete. Based on sequence identity and similarity the approximate 11 000 polypeptide chains in the current PDB were reduced to a set of approximately 1800 unique chains, for a total search time of approximately 209 days using an Ultra Sparc II processor. Since comparisons could be performed in parallel using a 256 processor Cray T3E results were produced in less than one day of Cray T3E time. The initial comparison is then updated on a daily basis using a desktop workstation, even given a near-exponential growth rate in the number of structures becoming available. Searches of the precomputed all-by-all comparison are available from the same Web site. Full details of the complete database will be published elsewhere.

Searches against the complete PDB database using a coordinate set not found in the PDB are possible at the same Web address with results being mailed to the submitter. Users wishing to run CE locally should contact one of the authors for the software. Binaries are available for several major UNIX platforms.

Acknowledgements

This work was supported by grant DBI 9630339 from the National Science Foundation.

References

Alexandrov,N.N. (1996) *Protein Engng*, **9**, 727–732.

Alexandrov,N.N. and Fischer,D. (1996) *Proteins*, **25**, 354–365.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,Jr.,E.F., Brice,M.D., Rogers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Chothia,C. (1992) *Nature*, **357**, 543–544.

Feng,Z.K. and Sippl,M.J. (1996) *Fold Des*., **1**, 123–132.

Fischer,D., Bachar,O., Nussinov,R. and Wolfson,H. (1992) *J. Biomol. Struct. Dyn*., **9**, 769–789.

Fischer,D., Elofsson,A., Rice,D.W. and Eisenberg,D. (1996) *Proc. 1st Pacific Symposium on Biocomputing*, 300–318.

Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Godzik,A. (1996) *Protein Sci.*, **5**, 1325–1338.

Hendrickson,W.A. (1979) *Acta Crystallogr.*, **A35**, 158–163.

Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.

Hogue,C.W.V., Ohkawa,H. and Bryant,S.H. (1996) *Trends Biol. Sci.*, **21**, 226–229.

Holm,L. and Sander,C. (1993) *J. Mol. Biol.*, **233**, 123–138.

Holm,L. and Sander,C. (1994) *Proteins*, **19**, 165–173.

Holm,L. and Sander,C. (1996) *Nucleic Acids Res.*, **24**, 206–210.

Holm,L. and Sander,C (1998) *Nucleic Acids Res*., **26**, 316–319.

Johnson,M.S. and Doolittle,R.F. (1986) *J. Mol. Evol.*, **23**, 267–278.

Madej,T., Gibrat,J.F., Bryant,S.H. (1995) *Proteins*, **23**, 356–369.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.

Needleman,SB. and Wunch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.

Orengo,C.A., Brown,N.P. and Taylor,W.T. (1992) *Proteins*, **14**, 139–167.

Orengo,C.A., Jones,D. and Thornton,J.M. (1994) *Nature*, **372**, 671–674.

Ponomarenko,M.P., Shindyalov,I.N., Kolchanov,N.A. and Bourne,P.E. (1998) submitted.

Russo,A.A, Jeffrey,P.D., Patten,A.K., Massague,J. and Pavletich,N.P. (1996) *Nature*, **382**, 6325–6331.

Sayle,R. and Milner-White,E. (1995) *Trends Biol. Sci.*, **20**, 374–376.

Smith,C.M., Gribskov,M., Shindyalov,I.N., Taylor,S.S., Ten Eyck,L.F., Veretnik,S. and Bourne,P.E. (1997) *Trends Biol. Sci.*, **22**, 444–446.

Vriend,G. and Sander,C. (1991) *Proteins*, **11**, 52–58.

Received December 2, 1997; revised February 25, 1998; accepted February 26, 1998