



A segment alignment approach to protein comparison

Yuzhen Ye, Lukasz Jaroszewski, Weizhong Li and Adam Godzik*

The Burnham Institute, La Jolla, CA 92037, USA

Received on July 13, 2002; revised on October 15, 2002; November 13, 2002; accepted on November 27, 2002

ABSTRACT

Motivation: Local structure segments (LSSs) are small structural units shared by unrelated proteins. They are extensively used in protein structure comparison, and predicted LSSs (PLSSs) are used very successfully in *ab initio* folding simulations. However, predicted or real LSSs are rarely exploited by protein sequence comparison programs that are based on position-by-position alignments.

Results: We developed a SEgment Alignment algorithm (SEA) to compare proteins described as a collection of predicted local structure segments (PLSSs), which is equivalent to an unweighted graph (network). Any specific structure, real or predicted corresponds to a specific path in this network. SEA then uses a network matching approach to find two most similar paths in networks representing two proteins. SEA explores the uncertainty and diversity of predicted local structure information to search for a globally optimal solution. It simultaneously solves two related problems: the alignment of two proteins and the local structure prediction for each of them. On a benchmark of protein pairs with low sequence similarity, we show that application of the SEA algorithm improves alignment quality as compared to FFAS profile-profile alignment, and in some cases SEA alignments can match the structural alignments, a feat previously impossible for any sequence based alignment methods.

Availability: SEA is freely available for academic users on a web server <http://ffas.ljcrf.edu/sea>.

Supplementary information: <http://ffas.ljcrf.edu/sea>

Contact: adam@burnham.org

INTRODUCTION

With increasing evolutionary distance the similarities between homologous proteins become less and less evident on the sequence level, until the only remaining relationship is a general fold similarity (Rost, 1997).

*To whom correspondence should be addressed. Program in Bioinformatics and Biological Complexity, The Burnham Institute, 10901 N. Torrey Pines Road, La Jolla, CA 92037, USA.

Proteins with similar folds can be described as having a similar spatial arrangement of small structural units, most conspicuous being alpha helices and beta strands. Such units are often shared by proteins with different folds. We define local structure segments (LSSs) as maximal structural units that are shared by proteins with different folds. Such segments can be predicted by nearest-neighbor methods which typically produce a list of *Predicted Local Structure Segments (PLSSs)* for a given protein (Fig. 1, Rychlewski and Godzik, 1997; Yi and Lander, 1993; Bystroff and Baker, 1998). Generally, such predictions are ambiguous—multiple and often contradictory PLSSs are predicted along a sequence. This ambiguity can be viewed either as a result of prediction defects, or as a fundamental feature of local structure preferences. In the latter interpretation, PLSSs can be viewed as a list of potential local segments, some of which are later eliminated by other factors, such as non-local interactions in the final structure.

Although nearest-neighbor algorithms initially consider the ambiguity in local structure, most do not carry these ideas further. Instead, they use only single position secondary structures averaged over the segments (Rychlewski and Godzik, 1997; Yi and Lander, 1993). The notable exception is Baker and colleagues (Bystroff and Baker, 1998) who further combined the predicted segments for a compact tertiary structure in their *de novo* protein structure prediction program ROSETTA (Simons *et al.*, 1999).

Meanwhile, most protein comparison methods are firmly based on the concept of residue-level alignments (Waterman, 1995), including programs that recognize distantly homologous or even non-homologous, but structurally similar proteins (Sternberg *et al.*, 1999). One can ask whether the language of *position-by-position* alignments adequately describes distant evolutionary relationships, or whether this is an oversimplifying assumption that discards otherwise important information. Analysis of structural similarities between distantly related proteins suggest the latter, but so far, lack of

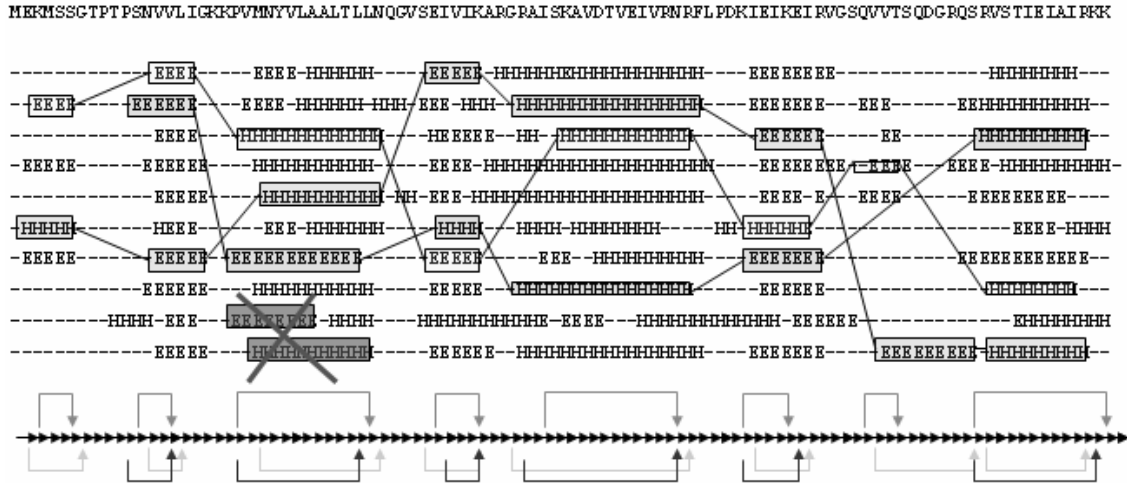


Fig. 1. Construction of the network of PLSSs. Each PLSS is represented as a string of local structure symbols, here a three state (helix, extended, no classification), in the actual calculations more detailed definitions are used (see text for details). Some PLSSs overlap so that their connections are forbidden (a pair of dark shaded segments). All compatible combinations of PLSSs correspond to the paths in the network (only three examples here are shown for clarity). As segments pairs are not always connected end-to-start, segments of amino acids are required to fill in the gaps between them (not shown).

adequate algorithms has made this a purely rhetorical question.

Here we present a *Segment Alignment (SEA)* algorithm, a segment-based protein comparison algorithm that operates in the entire space of predicted local structure segments. SEA finds the best match between all possible paths through PLSS networks representing two proteins, and therefore can use segments different from the locally best to contribute to the globally optimal alignment.

SEGMENT ALIGNMENT (SEA) FORMULATION

Given two protein sequences $S_1[1 \dots M]$ and $S_2[1 \dots N]$ whose local structure segments (LSSs) are predicted, SEA's goal is to find an optimal alignment between these two proteins by comparing all possible combination of segments. Even if some combinations can be excluded (see Fig. 1 for explanation), it would be prohibitively expensive to enumerate all of them. By representing the set of PLSSs for each protein as a network (unweighted graph), protein alignment is transformed into the problem of getting a path from source to sink vertex in each network with the optimal similarity score to a path in another network. This is a well-known *network matching problem* that can be solved by dynamic programming in polynomial time. A similar problem, called the spliced sequence alignment, has been proposed and solved for assembling genes from alternative exons (Gelfand *et al.*, 1996; Novichkov *et al.*, 2001).

Figure 1 shows how to construct a network of PLSSs for a given protein. First, each residue is described as a

vertex in the graph, and two artificial vertices are added to the very beginning (*source vertex*) and the end (*sink vertex*) of this protein. Then, for each PLSS α , we add an edge between the vertices of its first (denoted as $first(\alpha)$) and last (denoted as $last(\alpha)$) positions. We say that the segment α covers position i if $first(\alpha) \leq i$ and $last(\alpha) \geq i$, and specify i position as i_a . The set of PLSSs covering position i is denoted as $E(i)$. In practice, some parts of the protein may not be covered by any segments due to poor predictions. For the sake of continuity in any potential path, virtual edges (i.e. edges that do not correspond to a predicted segment) are added to all residues to form a complete network. An assembly of connected PLSSs corresponds to a path in this network.

The task of SEA is then to compare two networks of PLSSs by dynamic programming (Fig. 2). For any pair of positions, i and j , their covering segments are considered in a combinatorial way (total $|E(i)||E(j)|$ combinations) and are compared to get the optimal similarity score; it makes SEA different from sequence pair-wise alignments where only residues at positions i and j are compared. We define $V(i, j)$ as the maximum similarity score for transforming $S_1[1 \dots i]$ to $S_2[1 \dots j]$, calculated by

$$V(i, j) = \max_{\text{all}(\alpha, \beta) \text{ combinations}, \alpha \in E(i), \beta \in E(j)} V(i_\alpha, j_\beta) \quad (1)$$

where $V(i_\alpha, j_\beta)$ is the maximum similarity score for transforming $S_1[1 \dots i_\alpha]$ to $S_2[1 \dots j_\beta]$. It is computed recursively by the following formula, where $S(i_\alpha, j_\beta)$, $D(i_\alpha, j_\beta)$ and $I(i_\alpha, j_\beta)$ are the maximum similarity scores for transforming $S_1[1 \dots i_\alpha]$ to $S_2[1 \dots j_\beta]$ ending with

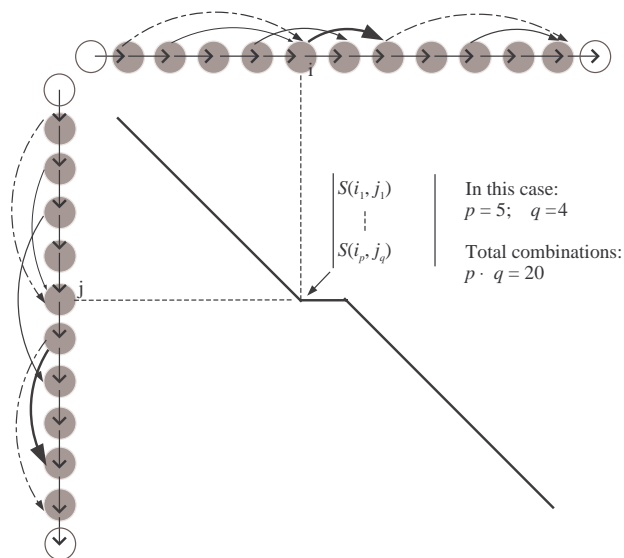


Fig. 2. Solving the network matching problem by dynamic programming. The two networks are shown at the top and the left of the graph. The optimal alignment path is shown as a dark line.

substitution, deletion and insertion at (i_α, j_β) , respectively.

$$\begin{cases} V(i_\alpha, j_\beta) = \max [S(i_\alpha, j_\beta), D(i_\alpha, j_\beta), I(i_\alpha, j_\beta), 0] \\ S(i_\alpha, j_\beta) = \max_{\gamma, \delta} [V((i-1)_\gamma, (j-1)_\delta) \\ \quad + \Delta(i_\alpha, j_\beta)] \\ D(i_\alpha, j_\beta) = \max_{\gamma} [\max[V((i-1)_\gamma, j_\beta) - g, \\ \quad D((i-1)_\gamma, j_\beta)] - h] \\ I(i_\alpha, j_\beta) = \max_{\delta} [\max[V(i_\alpha, (j-1)_\delta) - g, \\ \quad I(i_\alpha, (j-1)_\delta)] - h] \end{cases} \quad (2)$$

where

$$\begin{cases} \gamma = \alpha & \text{if } (first(\alpha) \neq i) \\ \gamma \in E(i-1) \& last(\gamma) = i-1 & \text{else} \end{cases} \quad (3)$$

$$\begin{cases} \delta = \beta & \text{if } (first(\beta) \neq j) \\ \delta \in E(j-1) \& last(\beta) = j-1 & \text{else} \end{cases} \quad (4)$$

Equations (3) and (4) define the important *compatibility* requirement for the continuation of segments. The affine gap function is applied, with g and h standing for the gap initiating penalty and gap extension penalty, respectively (Gusfield, 1999). The similarity score of aligned positions i and j is $\Delta(i_\alpha, j_\beta)$ (see Implementation), and in principle it could be any measure of similarity between segments.

As SEA considers all segment combinations at each pair of positions, its computational complexity is about $O(NMC_1C_2)$, where C_1 and C_2 are the average numbers of segments that cover a position in each protein (the *segment coverage*).

IMPLEMENTATION

We implemented the SEA algorithm in C++ on a Linux platform. The running time of SEA on a 1 GHz PIII with 1 Gb of RAM varies from several seconds to minutes, depending on the length of the query proteins and their PLSSs distributions. In this section, we address some of the practical issues in implementing the SEA approach. It is important to note that all the solutions discussed below are specific for this particular implementation of the SEA algorithm. In particular, non-local segment similarity measures can be used instead of a formula from Equation (5).

The prediction and representation of local structures

We used the HMMSTR/Rosetta server (<http://honduras.bio.rpi.edu/~isites/hmmstr/server.html>) with its default parameters to predict the local structure segments (PLSSs) and the one-dimensional local structures (1D). Correspondingly, we adopted its 11 symbols for local structures {HGEEBdbLlxc}, each having different backbone dihedral (ϕ and ψ) regions (Bystroff and Baker, 1998), and simply described each PLSS as a short string of local-structure symbols. Several variants of the SEA algorithm were introduced, including SEA_all (using the entire set of PLSSs), SEA_cn (n is the maximum segment coverage, e.g. SEA_c30 and SEA_c5), and two special cases in which the network representing a protein is simplified to a single path: SEA_1D using the 1D prediction and SEA_true using segments derived from the actual 3D structure.

Scoring scheme

The similarity score between two aligned positions in the current implementation of SEA is formulated as,

$$\Delta(i_\alpha, j_\beta) = W_a \times \Delta(Aa_i, Aa_j) + W_s \times \Delta(\alpha, \beta) \quad (5)$$

where W_a and W_s are the relative weights of sequence similarity and local structure similarity, satisfying $W_a + W_s = 1$. $\Delta(Aa_i, Aa_j)$ is the sequence similarity defined by Blosum62 similarity matrix (Henikoff and Henikoff, 1992), and $\Delta(\alpha, \beta)$ is the similarity between local structures defined in the next paragraph. W_s is set to 0.5, and the gap opening and elongation parameters are set to -5 and -1 . This set of parameters was partly optimized on a small set of SEA_true alignments.

We extracted a local structure similarity matrix from the subset of the HOMSTRAD database (SUB177) containing 706 proteins in 177 structural families (Shi *et al.*, 2001). The proteins' local structures were calculated from their 3D structures. The local structure similarities were then calculated from the log-odds of the probability of matching a pair of local structures in this database relative to a random one (Henikoff and Henikoff, 1992).

Table 1. General performance of SEA incorporating different local structure diversities

Subset	Measures	CE	SEA_true	SEA_c30	SEA_c10	SEA_c5	SEA_1D	BLAST	ALIGN	FFAS
Family (409 pairs)	Average-shift		0.61	0.56	0.56	0.54	0.49	0.44	0.48	0.49
	Shift > 0.9		73	69	63	56	47	51	60	43
	Shift > 0.7		207	199	192	183	152	146	165	161
	Shift > 0.5		282	260	259	251	215	197	228	227
	RMSD ≤ 3.0	257	95	82	82	76	63	77	54	40
	RMSD ≤ 5.0	397	237	184	171	177	147	157	138	118
	RMSD ≤ 8.0	408	294	248	249	249	231	196	206	194
	All	409	345	404	398	368	366	232	372	409
Superfamily (225 pairs)	Average-shift		0.27	0.12	0.12	0.12	0.08	0.09	0.06	0.07
	Shift > 0.9		3	3	3	2	0	1	2	1
	Shift > 0.7		17	8	9	7	4	10	9	7
	Shift > 0.5		54	26	23	21	17	18	18	17
	RMSD ≤ 3.0	55	12	6	6	7	6	8	3	1
	RMSD ≤ 5.0	160	44	16	18	18	11	18	11	1
	RMSD ≤ 8.0	163	69	37	34	41	28	23	22	15
	All	166	128	217	204	181	177	41	149	225

Two subsets of the benchmark (family-level subset and superfamily-level subset) are compared with SEA variants and other programs: CE, BLAST, ALIGN and FFAS (see text for the descriptions of these programs). The average-shift row lists the shift score averaged over all the alignments of each subset by different programs. The number of alignments with shift score >0.9, >0.7 and >0.5, RMSD ≤3.0, ≤5.0 and ≤8.0 in each subset are also listed in rows. For the alignment to be counted in RMSD evaluation, its length must be at least half of its corresponding structural alignment. As the CE structural alignments are used as reference alignments, its evaluation by shift score is meaningless and the corresponding numbers are blanked out.

(bottom of Figure 3, the local structures confirmed by SEA from a set of possible PLSSs are shown beside the bottom alignment) than the SEA_1D and pure sequence-based alignments (Table 2). More importantly, the segments contributing to the final alignment are not locally optimal, but are closer to the true structures.

Another example is the comparison of *E.coli* bacterioferritin (PDB code 1bcf, chain A) and amphibian red-cell L ferritin (PDB code 1rcd). These two proteins have very low sequence identity but high structural similarity (RMSD is 1.75). BLAST was not able to align these two proteins whereas the SEA_c30 alignment (RMSD is 2.77) is similar to the CE based 3D comparison with only minor differences in the gap regions. This is a big success compared to the results of SEA_1D and programs based purely on sequence comparisons.

SEA_30 also aligned mating-type protein-2 from *Saccharomyces cerevisiae* (PDB code 1ymn, chain B) and the Myb DNA-binding domain (PDB code 1mse, chain A) with an RMSD value similar to CE's, but the two alignments were different with a shift score less than 0. Actually, the Myb DNA-binding domain has two repeated domains: R2 and R3 (Ogata et al., 1994). CE matched the R2 domain to the mating-type protein-2, whereas SEA aligned R3 domain to mating-type protein-2. Despite SEA's alignment length being a little shorter than CE's, the sequence similarity is higher with fewer gaps. SEA therefore provided some interesting hints as to the evo-

lutionary relationship between the two repeated domains in Myb and mating-type protein-2: it shows that the evolutionary relationship between mating-type protein-2 with the R2 domain is closer than with the R3 domain.

Alignment quality versus local structure prediction ambiguity

Accounting for the ambiguity of the local structure prediction is necessary for a good comparison; however, using too diverse PLSS set will increase the probability of a random match and slow down the calculation. It is important to keep proper prediction diversity in SEA calculation.

Table 2 lists the alignment results for different segment coverage (all, 50, 30, 10 and 5). Although there is no simple relationship between segment coverage and the alignment quality, the results show that proper diversities (e.g. c30) are superior to low diversities (e.g. c5) and to single predicted local structure (1D). Interestingly, very high diversities do always produce optimal results, as in the alignments between staphylococcal enterotoxin A (PDB code 1esf, chain A) and toxic shock syndrome toxin-1 from *Staphylococcus aureus* (PDB code 2tss, chain A). Briefly, the middle regions of the two proteins are consistently aligned, while the less-conserved terminals are differently aligned by different programs and SEA incorporating different structure diversities. SEA_c50, SEA_30 and SEA_c10 achieved better alignments in the unstable regions and thus produced better complete

Table 2. A more detailed comparison of the performances of SEA and other alignment programs

Pro	Len	Ide		SEA_all	SEA_c50	SEA_c30	SEA_c10	SEA_c5	SEA_1D	SEA_true	FFAS	BLAST	CE
1lliA	89	33	RMSD	2.65	2.02	2.03	2.03	2.20	9.23	2.03	4.99	<i>1.74</i>	1.95
1r69	63		Shift	0.92	0.96	0.94	0.94	0.91	-0.10	0.94	0.78	<i>0.68</i>	1.00
			Ali	62	60	57	57	57	56	57	89	<i>31</i>	61
1bcfA	158	18	RMSD	3.98	2.99	2.77	2.50	12.14	7.88	2.35	13.61	-	1.75
1rcd	171		Shift	0.67	0.82	0.83	0.84	0.15	0.30	0.90	0.29	-	1.00
			Ali	164	160	159	161	155	101	159	184	14	158
1yrnB	78	12	RMSD	5.27	4.82	4.50	3.20	6.44	-	3.58	14.97	-	4.56
1mseC	105		Shift	-0.08	-0.08	-0.08	-0.07	-0.09	-	-0.07	0.04	-	1.00
			Ali	60	60	60	53	61	19	52	108	12	73
1esfA	229	18	RMSD	9.04	4.62	4.81	5.14	14.34	10.54	4.66	8.73	<i>3.46</i>	2.81
2tssA	194		Shift	0.68	0.77	0.77	0.72	0.26	0.46	0.75	0.57	<i>0.55</i>	1.00
			Ali	222	212	224	222	212	195	219	233	<i>93</i>	214
1ash	147	10	RMSD	3.80	4.36	4.38	3.54	3.70	<i>4.06</i>	3.02	4.16	-	2.31
1ithA	141		Shift	0.75	0.62	0.63	0.76	0.76	<i>0.37</i>	0.81	0.70	-	1.00
			Ali	146	147	146	143	142	68	124	149	10	146
1hbg	147	19	RMSD	2.97	4.71	2.82	5.74	3.11	6.22	3.08	5.53	-	2.11
1mbd	153		Shift	0.89	0.76	0.91	0.70	0.82	0.49	0.80	0.70	-	1.00
			Ali	153	150	153	150	153	126	155	158	0	152
1iyu	79	17	RMSD	2.85	2.85	2.85	2.84	2.85	7.55	2.24	12.04	<i>4.63</i>	2.57
1fyc	106		Shift	0.91	0.91	0.91	0.90	0.91	-0.09	-0.08	0.64	<i>-0.09</i>	1.00
			Ali	75	75	75	74	75	45	25	105	29	77
1prtF	98	14	RMSD	4.83	4.83	4.85	6.53	12.93	7.42	3.37	7.37	-	2.39
1ltsD	103		Shift	0.69	0.69	0.69	0.35	-0.11	-0.07	0.69	0.58	-	1.00
			Ali	90	90	89	94	57	40	94	113	0	96
1wiu	93	10	RMSD	3.41	3.41	3.41	3.45	3.45	15.44	3.54	4.08	12.24	2.15
1tiu	89		Shift	0.82	0.82	0.82	0.80	0.80	0.31	0.65	0.68	0.35	1.00
			Ali	89	89	89	89	89	85	67	95	78	90
1mjc	69	19	RMSD	3.78	6.36	3.14	6.90	6.74	<i>3.46</i>	-	7.00	4.59	2.44
1sro	76		Shift	0.65	0.56	0.72	0.49	0.56	<i>0.44</i>	-	0.57	0.33	1.00
			Ali	70	70	70	72	72	30	11	76	36	72

The Pro column lists the pdb code of each protein. The Len column lists the length of each protein. The Ide column lists the identity percentage of each protein pair calculated from the structural alignments by CE. The Ali row lists the alignment length of each pair. The low RMSD values caused by short alignments are shown in italic fonts.

alignments between these two proteins than SEA_c5, SEA_1D, FFAS and BLAST (see Table 2). This case shows that local structure information is crucial for improving alignments, especially in the less conserved regions.

CONCLUSION

In this paper, we introduced the *SEgment Alignment* (SEA) approach for comparing proteins described as collections of segments. SEA utilizes the full extent of predicted local structure information by exploring all possible PLSSs, some of which are correctly predicted while others are not. Such prediction uncertainty is unavoidable, due to limitations of individual prediction approaches but also due to the cooperative effect of the

whole structure, as evident in frequent cases of identical sequence segments adopting different local structure in different proteins (Mezei, 1998). SEA solves the chicken and egg dilemma of how to accurately predict protein local (secondary) structure which requires accurate protein alignment, while accurate alignment requires accurate structure prediction, by doing both simultaneously.

The preliminary version of SEA performed very well on a limited set of test cases. It is worth noting that many of the choices made in this exploratory effort were not fully optimized. Systematic study of the effects of LSS prediction in SEA is necessary. It would also be interesting and vital to compare the LSS prediction by the I-site method; (Bystrhoff and Baker, 1998) and other approaches (e.g. nearest-neighbor method Yi and Lander, 1993). We expect better results as we improve

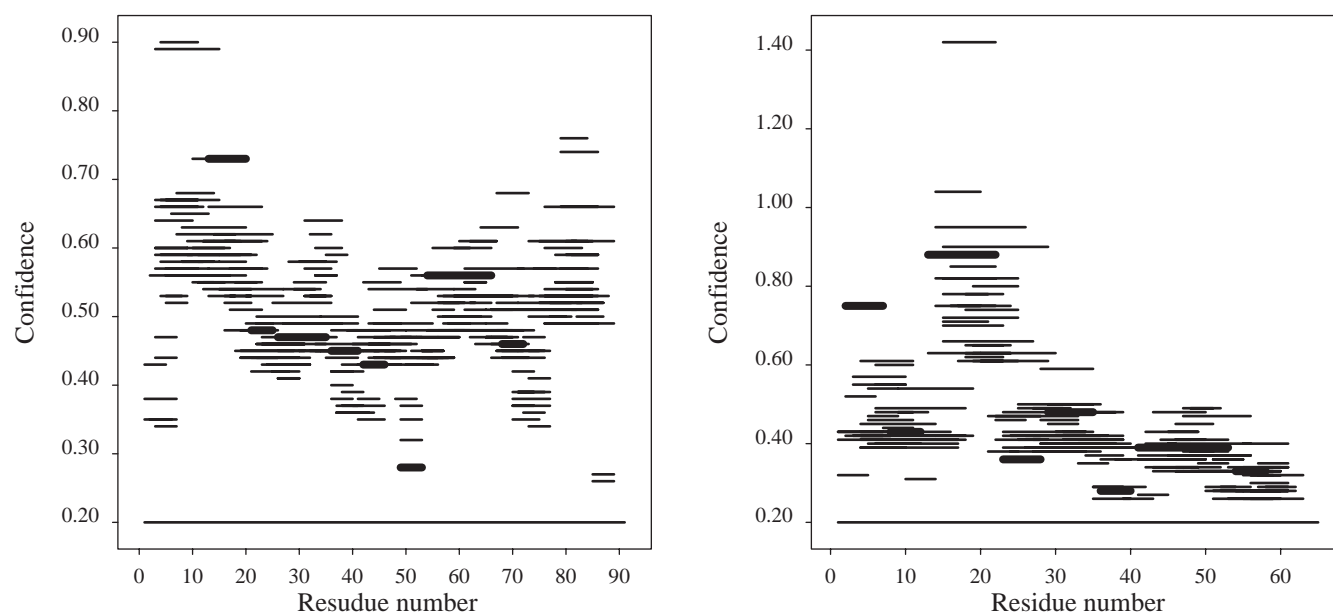


Fig. 4. The PLSSs distributions of λ -repressor from *E.coli* (pdb code 1lli) and 434 repressor from phage 434 (pdb code 1r69). The PLSSs derived from HMMSTR/Rosetta server are shown in lines in the graph, where the y-axis shows the confidence of PLSSs. The structure segments in the optimal alignment derived from SEA are highlighted in bold lines.

segment definitions, similarity measurements (especially those segment-based such as RMSD between segments), and significance evaluations. We are now exploring the local structure prediction using FFAS and testing the influence of length of segments on the alignment quality.

SEA has many potential applications, such as fold-recognition, distant homology detections and refining protein alignments for better structure modeling. As SEA also predicts the local structures, it can also be used for local structure predictions where many possibilities of local structures are given. We are currently developing and testing applications of SEA to protein modeling and fold recognition, which will be described in separate publications.

ACKNOWLEDGEMENTS

We appreciate valuable comments about the algorithm development from Dr Haixu Tang. We are indebted to Dr Chris Bystroff for his kind and immediate responses about the I-site related services and their prediction server. We thank Dr Melissa Cline for kindly providing us their shift score calculation program, Dr Ilya N. Shindyalov for kindly providing us his structural comparison program CE and Drs Jason Hoffman and Iddo Friedberg for help in editing. This research was supported by a grant NIH GM63208.

REFERENCES

- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Bystroff,C. and Baker,D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.*, **281**, 565–577.
- Cline,M., Hughey,R. and Karplus,K. (2002) Predicting reliable regions in protein sequence alignments. *Bioinformatics*, **18**, 306–314.
- Gelfand,M.S., Mironov,A.A. and Pevzner,P.A. (1996) Gene recognition via spliced sequence alignment. *Proc. Natl Acad. Sci. USA*, **93**, 9061–9066.
- Gusfield,D. (1999) *Algorithms on strings, trees and sequences: computer science and computational biology*, Second edn, CUP, New York.
- Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
- Jaroszewski,L., Li,W. and Godzik,A. (2001) Improving the quality of twilight-zone alignments. *Protein Sci.*, **9**, 1487–1496.
- Mezei,M. (1998) Chameleon sequences in the pdb. *Protein Eng.*, **11**, 411–414.
- Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
- Myers,E.W. and Miller,W. (1988) Optimal alignments in linear space. *CABIOS*, **4**, 11–17.

- Novichkov,P.S., Gelfand,M.S. and Mironov,A.A. (2001) Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics*, **17**, 1011–1018.
- Ogata,K., Morikawa,S., Nakamura,H., Sekikawa,A., Inoue,T., Kanai,H., Sarai,A., Ishii,S. and Nishimura,Y. (1994) Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell*, **79**, 639–648.
- Rost,B. (1997) Protein structures sustain evolutionary drift. *Fold Des.*, **2**, S19–24.
- Rychlewski,L. and Godzik,A. (1997) Secondary structure prediction using segment similarity. *Protein Eng.*, **10**, 1143–1153.
- Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
- Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.
- Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
- Simons,K.T., Bonneau,R., Ruczinski,I.I. and Baker,D. (1999) *Ab initio* protein structure prediction of CASP III targets using rosetta. *Proteins*, **37**, 171–176.
- Sternberg,M.J., Bates,P.A., Kelley,L.A. and MacCallum,R.M. (1999) Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.*, **9**, 368–373.
- Waterman,M.S. (1995) *Introduction to Computational Biology*, First edn, Chapman and Hall, London, pp. 183–228.
- Yi,T.M. and Lander,E.S. (1993) Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.*, **232**, 1117–1129.