J Mol Evol (1997) 44(Suppl 1):S174–S180

JOURNAL OF MOLECULAR EVOLUTION

Random Structural Models for Double Dynamic Programming Score Evaluation

William R. Taylor

Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, United Kingdom

Received: 23 July 1996 / Accepted: 28 August 1996

Introduction

The comparison of protein sequences and structures has provided great insight into biological problems. This is especially true when the comparison method has revealed an obscure similarity that may connect two otherwise unrelated groups of proteins—often resulting in some novel hypothesis concerning their origin and function. Since protein structures are better conserved through evolution than the sequences that determine them, the comparison of protein structure provides our best probe into the deep evolutionary past (Rossmann et al. 1974). An inherent problem with distantly related proteins, however, is that the similarity score calculated by the comparison program is difficult to interpret in terms of a measure of probability.

This problem has been addressed in the comparison of sequences by generating numerous randomized sequences that to varying degrees retain the overall properties of the original sequence. Typically, the length is retained and the amino acid composition of the protein, and occasionally, higher-order sequence correlations. If the comparison method involves a protein structure, as in sequence/structure comparison (threading) or structure/structure comparison, such randomized models are not so easily generated. Solutions to the problem in structure comparison have involved empirical normalization for protein length (Orengo et al. 1993) or in sequence/structure comparison have been based on the population of scores obtained when a sequence is compared to a large number of proteins or protein fragments (Jones et al. 1984; Sippl and Weitckus 1992).

In the current work, a method is developed that provides a normalization for scores calculated by the double dynamic programming method, used both in sequence/structure comparison (Jones et al. 1984) and structure comparison (Taylor and Orengo 1989). The approach is based on a simplified form of the double dynamic programming method that uses a structural model based only on α-carbons, so allowing greater freedom in the construction of randomized control models. The current work will concentrate on structure comparison.

Materials and Methods

Double Dynamic Programming/Comparison Measure/Direction Component. The comparison method is based on the comparison of geometric environments of residues. These are expressed as the set of vectors between α-carbons from a given position (say, i) to all other residues in the protein, where the vectors are resident in the coordinate frame defined by residues $i - 1$, i, and $i + 1$. The correspondence between two vectors was measured as their difference, inverted by a Gaussian function to produce as score, as follows:

$$^{ij}v_{mn} = \exp(-(^{i}V_m - {}^{j}V_n)^2/^{v}\sigma^2) \quad (1)$$

where $^{i}V_m$ is the interatomic vector in protein A from residue i to m, and $^{j}V_n$ the corresponding vector in protein B from residue j to n and $^{v}\sigma$ controls the decay of the score with increasing difference (equivalent to the standard deviation of the normal distribution).

Orientation Component. While the relative juxtaposition (direction) of residues is important, if the two pairs of residues are not in similar relative orientations, then they may not be equivalent. This can be quantified (retaining the above notation) as the relation of the coordinate frame of residue m in the coordinate frame of residue i (in protein A) to the coordinate frame of residue n in the coordinate frame of

residue j (in protein B). For simplicity, this will be referred to as relating m-in-i to n-in-j.

This relationship can be captured as three angles between equivalent axes of the transformed coordinate frames; however, to use this full relationship would either require excessive storage (of the order N^4 for proteins of equal length N) or multiple matrix rotation for every comparison. As a compromise between these extremes of storage and calculation, only the local orientations (all m-in-i, in A and all n-in-j in B) were stored—requiring storage of the order N^2 for both proteins. For computational simplicity, the cosines were stored and a difference of equivalent terms was taken as a measure of similarity. This measure still retains the desired property that each component is zero when the axes are coincident.

The difference of each component was converted to a score (as above with the vector difference) as follows:

$$\begin{aligned} {}^{ij}p_{mn} &= \exp[-({}^{i}x_m - {}^{j}x_n)^2/^{c}\sigma^2] \\ {}^{ij}q_{mn} &= \exp[-({}^{i}y_m - {}^{j}y_n)^2/^{c}\sigma^2] \\ {}^{ij}r_{mn} &= \exp[-({}^{i}z_m - {}^{j}z_n)^2/^{c}\sigma^2] \end{aligned} \tag{2}$$

where, for example, ${}^{i}x_m$ is the cosine of the angle between the X-axes of the coordinate-frame of residue i and the frame of residue m (when both are at a common origin) and similarly for the Y and Z axes. As the combined score should be sensitive to deviation in the direction of any of the components, their product was taken, giving:

$${}^{ij}c_{mn} = {}^{ij}p_{mn} \cdot {}^{ij}q_{mn} \cdot {}^{ij}r_{mn} \tag{3}$$

Weighted Combination. The direction and the orientation components were combined in a weighted sum giving:

$${}^{ij}r_{mn} = ({}^{v}w \cdot {}^{ij}v_{mn}) + ({}^{c}w \cdot {}^{ij}c_{mn}) \tag{4}$$

This score (r) over each m, n forms a matrix, ${}^{ij}\mathbf{R}$, that was used for the low-level dynamic programming calculation of the best path.

The double dynamic programming algorithm calculates the best path through a matrix formed by summing the best paths through each matrix $\mathbf{R}$ for all i,j. This can be summarized by defining a function Z that sets all elements of a matrix to zero, other than those that lie on the optimal path. Using this, the full double dynamic programming calculation can be repeated as:

$$\mathbf{S} = Z\left[\sum_{\forall i}\sum_{\forall j} Z({}^{ij}\mathbf{R})\right] \tag{5}$$

The remaining nonzero elements in the matrix $\mathbf{S}$ specify the alignment between the two proteins.

Iterative Algorithm. An iterative approach was used as described previously, with ten cycles of iteration beginning from a selection of pairings based on the criteria of similar solvent exposure, secondary structure and sequence. The paths calculated from these selected pairing were then reincorporated (summed) into the original matrix, referred to as the *bias* matrix $(\mathbf{Q})$, to improve the selection of pairs for the next circle. This approach incorporates a positive-feedback component on the selection and by the final cycle most (commonly all) selected pairs lie on the alignment.

If ${}^{t}\mathbf{Q}$ is the bias matrix on cycle t, then the next revision is calculated as:

$${}^{t+1}\mathbf{Q} = {}^{t}\mathbf{Q}/2 + \log(1 + \mathbf{S}/10) \tag{6}$$

"Random" Models/Constrained Random Walk. One of the simplest "random" models for a protein is a constrained random walk (Thornton and Sibanda 1983; Cohen and Sternberg 1980). This model preserves primarily the protein length and also, to varying degrees, compactness. A similar model (provided by T. Flores) is used below in which an α-carbon chain is "grown" by the addition of residues to the terminus such that steric clashing is avoided and local packing is favored.

Random Models From Distance Geometry. The distance geometry program DRAGON (Aszódi and Taylor 1994a) can take a matrix of random interatomic (α-carbon) distances and generate compact models with bulk properties as expected for of globular proteins, either by hydrogen bonds (Aszódi and Taylor 1994b) or without (Aszódi and Taylor 1994a).

Combinatorial Models. For some limited classes of protein architecture, simple frameworks can be constructed that specify potential locations for secondary structure elements (secondary structure lattices) (Murzin and Finkelstein 1988; Taylor 1993). Combinatorial enumeration of the connectivity over these lattices produces a variety of models that can be made to have the same length as the original protein and the same ordering of secondary structure elements along the sequence (Taylor 1991).

Combinatorial Reconnection. Following a similar approach, the native protein itself (as distinct from an idealized lattice) can be reconnected. "Switch" points can be identified where two pairs of adjacent residues pack in an approximate tetrahedral arrangement. Designating the termini of the first pair as N_1 and C_1, and the second pair as N_2 and C_2, making the connection $N_1 \rightarrow N_2$ and $C_1 \rightarrow C_2$, preserves the integrity of the chain. With more than one reconnection, it cannot easily be predetermined whether all the residues will be linked in a single chain. This difficulty was circumvented by generating all switch-point combinations ($N_i \leftrightarrow N_j$ with $C_i \leftrightarrow C_j$ and $N_i \leftrightarrow C_j$ with $C_i \leftrightarrow N_j$) and testing each result for chain integrity. All such valid reconnections produce a variety of pseudo-random structures that preserve the character of the internal packing of the native structure.

Chain Reversal. It will be apparent that the previous reconnection operations reverse the chain direction of some segments. However, as only α-carbons are used, this does not disrupt the geometry of the alternative protein to any significant extent. Indeed, as used previously, the complete sequence of the protein can be reversed to produce a useful control model that preserves all the local geometry and sequence patterns associated with secondary structure (Taylor 1986).

This operation can be applied either alone or in combination with any of the previous models.

Chain Reflection. Similarly, the α-carbon trace can also be reflected to produce a "random" model that preserves not only length but also all interatomic distance exactly. This model, however, inverts the chiral aspects of secondary structure and as these are more apparent in the α-helix, would produce distinctly different scores when used with the all-α class of protein compared to proteins of the all-β class.

Randomized Alignments. Two approaches can be made to test the statistical significance of a score value. Typically, the best score obtained for the correct alignment (the comparison of the two native structures) is normalized by the standard deviation obtained from a population of random models such as those described in the previous section. Some of the suggested random models described above, however (such as the single chain reversal), produce a population of limited size. In general, the closer the random model is to preserving the

properties of the native proteins, the more difficult it becomes to generate plausible alternatives.

This fundamental difficulty has been overcome in the current work by expanding the population of random models at the level of the alignment—generating from each individual model structure a family of near-optimal subalignments. This can be achieved very easily in the current iterative double dynamic programming method by introducing a random element into the initial bias matrix (**Q**) as:

$$Q_{ij} = (1 - f)s_{ij} + fr \tag{7}$$

where Q_{ij} is an element of the bias matrix, s_{ij} is the similarity score for residues i and j, r is a random number (between 1 and 0), and f is a parameter to control the degree of randomization (with a value also between 1 and 0). In the current work the generality of this form was not exploited and a purely random bias matrix ($f = 1$) was used.

If the double dynamic programming algorithm was perfect and given enough iteration cycles, it would achieve the global minimum alignment. However, with a randomized starting selection and a limited number of cycles there is little chance that the global minimum will be found. However, the resulting alignments will not be random as the algorithm will have refined any starting position toward a high score. Depending on the complexity of the landscape that it must traverse, the solution will become trapped in a local minimum. The distribution of the scores of these minima will depend on the initial degree of randomization (equivalent to temperature) and the number of refinement cycles (equivalent to annealing time).

An advantage of this approach is that it can be applied not only to structures belonging to the set of randomized models but also to the native structure itself. This, in effect, tests the stability (or uniqueness) of the global minimum. In this approach, the scores of both the native and the random distributions can be compared and tested for significance in terms of their joint distributions.

Results

The results presented below will concentrate on the situation where a limited population of random models is available. Specifically, those generated by reconnection of the native structure and the special case of this type in which the intact native structure has been reversed. However, as background controls, the results obtained with multiple models and no alignment perturbations will be considered.

Random Models

Twenty random models were generated for a chain length of 125 residues using the constrained random walk method and the distance geometry method both with and without hydrogen bonding. The models from each method were compared pairwise, giving 190 scores and the frequency distribution of these scores have been plotted in Fig. 1 (A–C). For comparison, the scores obtained from the pairwise comparison of the reversed structures were also plotted.

With a simple random walk, the scores show little variation and have a distribution that is almost indistinguishable from the reversed structures. As more structure is imposed on the random models using the distance

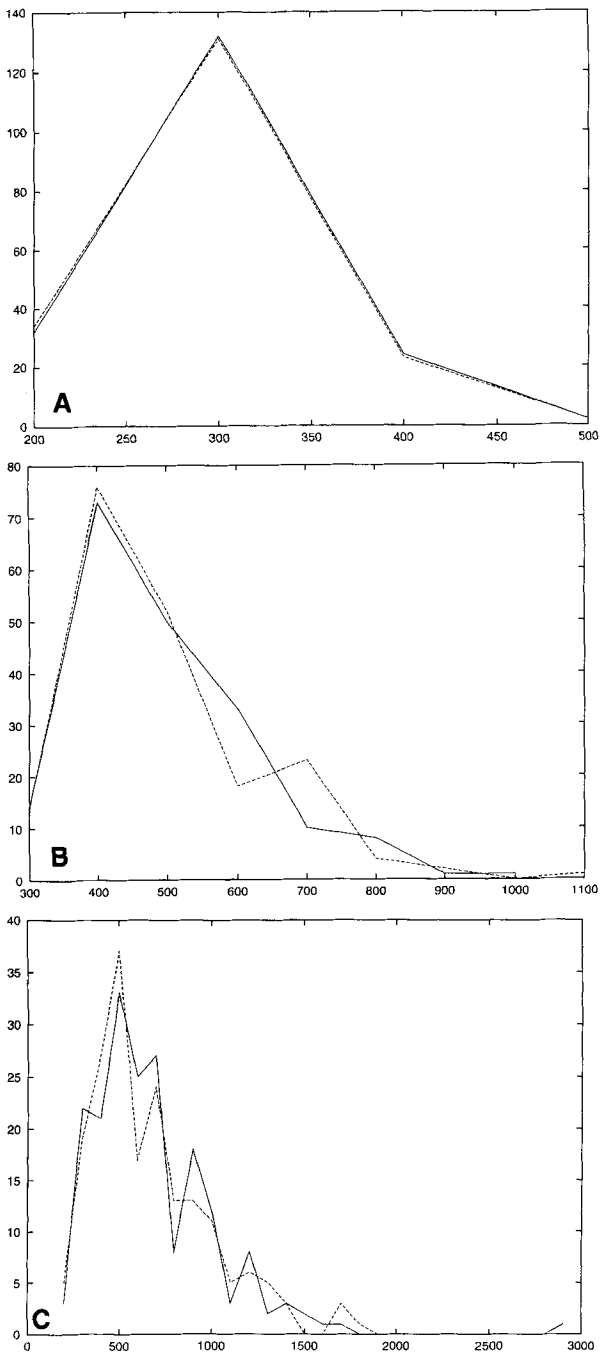

Fig. 1. Score distributions for random models **A** Constrained random walk model. **B** Distance geometry (DRAGON) models with no secondary structure. **C** Distance geometry (DRAGON) models with secondary structure. Each plot uses the same class interval of 100 score units to calculate the frequency. The frequency of the scores for the forward chains are plotted with a *line* and those for the reversed-chain models with a *dashed line*.

geometry method, the peak of the distribution increases slightly from 300 to 400 and then 500 (with hydrogen bonding); however, the spread of the distribution shifts markedly toward higher scores, especially when artificial secondary structures are included. Again, the distribution of the reversed structures matched the distribution of the forward structures closely—indicating that the distance

Table 1a. Scores for randomized alignments with reconnected structures[a]

349	287	383	313	189	386	261	373	313	284	+	1	2	3	−4
195	135	126	197	674	72	174	55	682	132	−	1	2	3	−4
679	338	463	174	553	644	683	813	505	798	+	1	2	3	−4
121	649	882	704	842	933	631	213	280	679	−	1	2	3	
402	292	453	380	147	438	323	323	407	409	+	1	2	−4	
79	483	597	218	682	355	103	502	81	138	−	1	2	−4	
451	881	979	346	832	964	817	245	810	834	+	1	2		
837	1028	727	471	737	1023	112	212	732	1032	−	1	2		
237	402	302	44	297	401	306	313	242	334	+	1	3	4	
132	240	483	87	258	276	129	674	158	68	−	1	3	4	
559	371	1270	1277	430	542	1165	1285	1183	266	+	1	3		
912	927	1016	345	226	1044	787	663	717	923	−	1	3		
626	567	378	92	339	39	574	336	351	582	+	1	4		
129	254	672	417	68	452	580	137	317	402	−	1	4		
639	115	477	694	619	637	651	603	670	760	+	1	−2	3	4
672	285	172	372	663	658	85	602	461	401	−	1	−2	3	4
513	834	589	615	585	325	459	677	120	625	+	1	−2	3	−4
735	371	347	347	188	649	173	789	656	206	−	1	−2	3	−4
564	934	789	668	370	789	872	666	931	463	+	1	−2	4	
517	648	974	309	594	304	403	528	407	996	−	1	−2	4	
965	678	777	906	795	196	287	224	645	802	+	1	−2	−4	
682	340	967	546	398	176	773	957	337	605	−	1	−2	−4	
1553	1520	674	103	603	689	1197	600	346	1354	+	1			
1076	677	1068	1075	1087	400	837	1084	1000	1090	−	1			
431	313	409	1310	972	729	110	976	376	355	+	2	3	−4	
514	235	170	398	275	156	256	244	611	703	−	2	3	−4	
2263	2262	2379	2263	1170	250	2265	2263	2265	623	+	2	3		
683	647	312	775	256	857	889	251	690	217	−	2	3		
507	299	339	853	1401	1432	1201	508	686	1449	+	2	−4		
40	315	271	290	376	549	547	368	404	347	−	2	−4		
2586	395	2456	379	356	2528	2588	1155	2569	1444	+	2			
694	335	655	1037	667	338	1017	1020	703	799	−	2			
829	1379	1061	1339	881	369	593	447	113	609	+	3	4		
498	245	467	128	362	356	404	145	211	289	−	3	4		
620	771	1879	2533	1967	2533	1841	1708	2517	1818	+	3			
928	951	789	925	1029	741	959	312	742	489	−	3			
1391	1195	1164	1177	224	1347	1170	1412	613	1187	+	4			
563	534	625	565	72	365	122	695	321	438	−	4			
2392	1016	199	2392	1149	253	341	2392	2392	2394	+	−2	3	4	
471	120	175	534	793	766	324	396	655	580	−	−2	3	4	
2377	1905	1225	2345	1163	600	2328	1196	610	23278	+	−2	3	−4	
644	234	210	421	336	695	670	304	543	489	−	−2	3	−4	
2536	1563	2541	731	2393	304	2564	487	1543	2565	+	−2	4		
697	532	191	962	584	466	689	629	776	684	−	−2	4		
1473	1281	1021	1565	2559	2566	957	2472	728	2472	+	−2	−4		
406	621	939	829	728	650	941	346	687	506	−	−2	−4		

[a] Reconnected chemotaxis-Y protein with flavodoxin

geometry modeling program (DRAGON) has no directional bias.

Reconnected Models

Two small proteins, the chemotaxis Y protein (3CHY) and a flavodoxin (4FXN), were chosen as example test data. Both have a mixture of β and α structure and share a common fold but have no appreciable sequence similarity. The overall score for the comparison of these two structures was 2,861 and the rigid-body superposition based on the alignment gave a weighted RMSd of 2.471 (over 106 atoms) and an unweighted RMSd of 3.791 over all matched atoms (106).

Four switch-points were automatically identified in the Che-Y protein, at positions 11/36, 87/107, 56/63, and 72/100 (N_i/N_j) and at 78/107, 88/118, 56/63, and 26/132 in flavodoxin. All combinations were generated, firstly for the Che-Y structure, with those that maintained an intact chain being compared to the flavodoxin, and similarly for the flavodoxin against Che-Y. For each reconnection, ten random starting configurations were generated for the bias matrix, giving the results tabulated below for both the reconnected Che-Y (Table 1a) and reconnected flavodoxin (Table 1b) structures.

Table 1b. Scores for randomized alignments with reconnected structures[a]

275	71	314	534	252	238	515	568	241	355	+	1	3	4	
80	320	52	194	306	153	220	239	125	169	−	1	3	4	
1201	1267	1177	1169	1158	1521	1457	1512	185	1165	+	1	3		
335	383	251	226	277	300	166	199	477	324	−	1	3		
176	214	60	513	64	528	334	534	115	506	+	1	4		
408	303	526	542	232	190	551	291	633	540	−	1	4		
515	179	176	158	238	509	394	151	198	475	+	1	−2	3	4
769	189	846	715	590	278	775	119	162	744	−	1	−2	3	4
1382	2260	2260	255	1426	1363	1575	1517	611	1481	+	1	−2	3	
436	867	628	959	580	440	250	636	88	543	−	1	−2	3	
212	273	188	143	48	369	24	163	160	123	+	1	−2	4	
327	77	287	431	144	584	522	277	87	623	−	1	−2	4	
1452	2261	1451	1458	2376	2261	1465	1485	1488	2264	+	1	−2		
804	286	665	296	457	216	706	614	463	998	−	1	−2		
1492	1455	1329	145	1196	1508	1154	1560	423	584	+	1			
126	313	620	153	200	357	96	463	328	197	−	1			
251	302	44	578	116	295	228	309	127	213	+	2	3	4	
32	439	536	344	339	480	486	928	482	653	−	2	3	4	
297	2655	485	1429	488	1475	1431	436	1806	2583	+	2	3		
655	692	530	705	815	712	155	747	785	686	−	2	3		
370	203	143	513	59	200	145	185	304	291	+	2	4		
383	719	18	206	485	290	653	317	246	376	−	2	4		
1487	2610	1498	493	321	2585	2590	285	2525	449	+	2			
734	131	193	902	786	774	717	585	523	893	−	2			
221	351	546	329	715	614	300	714	118	681	+	3	4		
1108	868	1078	1340	1251	465	515	638	507	1469	−	3	4		
2291	2759	2717	2125	2800	2326	2252	2883	2279	2759	+	3			
764	133	768	1026	330	1024	1023	614	353	1127	−	3			
290	705	711	331	77	348	230	315	331	656	+	4			
810	1261	802	789	236	556	185	582	1295	191	−	4			
230	570	75	591	188	145	92	245	389	146	+	−1	2	3	4
332	99	97	73	121	177	291	186	40	393	−	−1	2	3	4
1547	1456	1552	1168	1442	1311	1120	1112	1118	923	+	−1	2	3	
332	442	595	151	738	299	487	243	91	363	−	−1	2	3	
108	38	407	620	113	312	367	48	53	119	+	−	2	4	
386	433	91	194	122	169	552	480	165	428	−	−1	2	4	
568	350	1209	198	193	1091	315	1161	335	622	+	−1	2		
415	355	152	281	595	625	308	142	497	282	−	−1	2		
206	360	453	364	384	389	354	452	50	476	+	−1	−2	3	4
584	191	165	673	168	686	665	665	593	142	−	−1	−2	3	4
743	999	435	2358	488	269	418	416	1651	713	+	−1	−2	3	
837	806	688	694	589	713	44	318	590	568	−	−1	−2	3	
425	416	129	103	203	94	220	383	361	288	+	−1	−2	4	
211	237	702	558	166	688	245	641	508	323	−	−1	−2	4	
2478	2400	2201	2386	1610	1663	439	1604	1603	2386	+	−1	−2		
472	877	610	726	422	611	516	183	555	622	−	−1	−2		

[a] Reconnected flavodoxin protein with the chemotaxis-Y protein. In both tables, each reconnection model is specified by the switch-points used (1–4) and whether the connection was amino–amino (positive number) or amino-carboxy (negative number). The initial sign indicates whether the whole chain was matched in the (native) forward (+) or reversed (−) direction. For each model, ten scores are given which derive from alignments calculated from a random bias matrix

The scores for both structures were pooled and split into those using the reversed model (minus in the tables) and the forward models (plus in the tables). A frequency histogram of the scores was then generated for each group (Fig. 2). This revealed that the forward reconnections (beginning at the true amino-terminus) had a much greater range of variation toward higher scores.

Reversed Structure

The two proteins used in the previous section were retained to generate comparisons between the native structure and reversed structure. As only one comparison was possible without perturbing the alignment, 100 random starting configurations of the bias matrix were used to generate a population of alignment scores for both the reversed flavodoxin and reversed Che-Y structure (Fig. 3).

Discussion

Assessment of the Results

The simplest random model used above (constrained random walk) generated a distribution of low score (mean

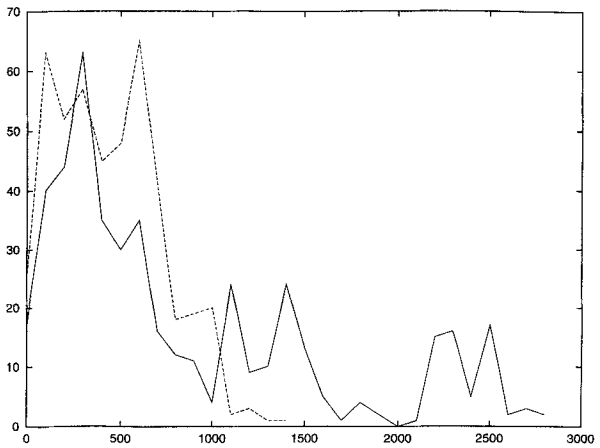

Fig. 2. Score distribution for reconnected models The frequency for the scores in Table 1 are plotted for the forward chain direction (plus in the table) (*line*) and for the reversed chain direction (minus in the table) (*dashed line*). The clear difference in the distributions is discussed in the text.

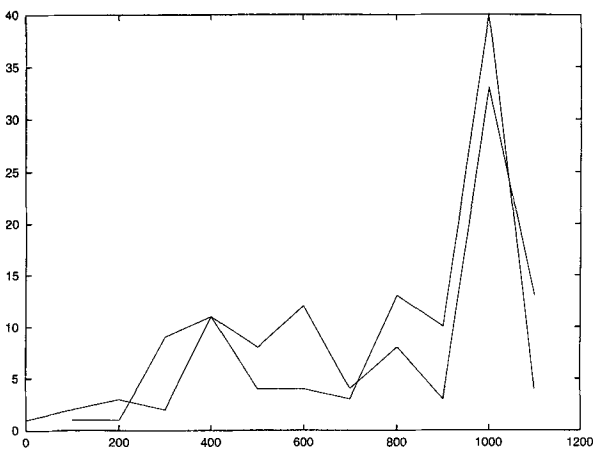

Fig. 3. Score distributions for full-reversed model: 100 alignments were calculated from random starting positions for the reversed chemotaxis-Y protein compared with flavodoxin and the reversed flavodoxin compared with the chemotaxis-Y protein. The dominant "spike" at 1,000 probably results from the internal symmetry of the super-secondary structures in these proteins.

300). Relative to this the score for the comparison of the Che-Y protein and flavodoxin was an order of magnitude higher. As would be expected, the imposition of protein-like bulk properties (density, axial ratios) using the distance geometry program DRAGON resulted in an overall potential to attain greater scores in the pairwise comparisons. This trend became more exaggerated with the imposition of (effectively random) secondary structures. This allowed some good scores to be obtained, even as high as that obtained between the two native proteins. Visual examination of the superposition of the two random structures that gave the best score revealed, as would be expected, that some secondary structure had been matched.

The reconnected structures compared to a native structure gave a similar distribution to the random structures with secondary structure but with a much greater density toward higher native/native-like scores. This would be expected as the reconnections, to varying extents, all contain pieces of intact native structure. In addition, the terminus is never altered, so all the reconnected structure begin and end in the same place and in the same secondary structural state. For example, in the two proteins considered above, the amino-terminus is a β-strand and the carboxy-terminus is an α-helix, whereas in the reversed reconnected variants, the amino-terminus is always an α-helix and the carboxy-terminus is always a β-strand. This bias to be unlike the native at the termini probably explains the failure of the reversed-reconnected distribution to extend toward high scores—even though the structures must contain some pieces of native-like structure. (The reconnections will create segments of re-reversed chain.)

The completely reversed structures (which rely on alignment randomization to generate a population of scores) would be expected to exhibit a roughly normal distribution with a mean roughly equivalent to the random structure with secondary structure. (The high-scoring tail of the latter distribution would be expected to be damped by the bias in the reversed structure to be un-native-like.) Unexpectedly, a markedly skewed distribution was observed with a distinct peak (or spike) around 1,000 (1/3 native score). This was found both with the reversed Che-Y structure (against flavodoxin) and the reversed flavodoxin (against Che-Y). Examination of a few superposed pairs with a scores just over 1,000 (at the spike) revealed that this feature resulted from the matching of super-secondary structures (three β-α units) as a consequence of the repetitive alternating arrangement of β-α structure in this class of protein.

On the Choice of Random Model

The choice of the best random model against which native/native comparison scores should be compared is not simple and depends on the degree to which the inherent nonrandom features of protein structure in general should be considered significant. Comparison of any two proteins containing similar proportions of secondary structure would give a significant match when compared against any of the two random models that do not model secondary structure. The most suitable of the random models would therefore be those generated with secondary structure. Ideally, these models should be calculated for each comparison to match the length of the native comparison and the secondary structure composition. However, these models are complex to generate and cannot be "tailor-made" for each individual comparison without excessive computation.

The reconnected structures were designed with the original intent of generating (at little computational expense) "random" protein structures with native length,

secondary structure content, and near-native internal packing—but with a non-native fold. However, the retention of sometimes large native-like segments makes these structures unattractive as a random model. This problem could be overcome through the use of more switch-points, both reducing the size of each native segment and the probability that many native segments will be retained in each model. However, if the limits are stretched on the criteria for acceptable switch-points, more distortions will be admitted into each model creating increased deviation from native-like packing.

The reconnected reversals go some way toward avoiding the retention of large native segments but perhaps overcorrect by their implicit bias to be persistently non-native like at the termini. This latter problem could be overcome if the termini themselves participated in switch-points—thus allowing any residue participating in a switch-point to be taken as the amino terminus. This situation closely resembles the combinatorial secondary structure lattice models (described above) and will be investigated elsewhere in that context.

The alignment-randomized reversed model is the simplest to generate computationally as it does not involve the construction of any explicit coordinate sets. However, besides the problem of the non-native-like bias, the full reversal also has the complicating feature that any internal symmetry (duplication or repeated super-secondary structure) will serve as a source for good local matching. This property, however, is not necessarily undesirable as proteins that have high internal symmetry should probably be judged by a more stringent criterion. This feature would be lost (or diluted) in the reconnected models; for example, in the alternating β/α proteins (considered above), the regular alternation of secondary structure type could be reconnected (in the extreme) into a sequence of nonalternating secondary structures (sequentially segregated β and α regions). The full reversal of the structure retains this secondary structure ordering and until reconnected structures can be generated to also preserve this feature, the full reversal (with alignment randomization) is currently the preferred model.

Acknowledgments. Dr. András Aszódi is thanked for generating the DRAGON models as is Dr. Tomas Flores for providing the program for generating the random-walk models.

References

Aszódi A, Taylor WR (1994a) Folding polypeptide α-carbon backbones by distance geometry methods. Biopolymers 34:489–506

Aszódi A, Taylor WR (1994b) Secondary structure formation in model polypeptide chains. Protein Eng 7:633–644

Cohen FE, Sternberg MJE (1980) On the prediction of protein structure: the significance of the root-mean-square deviation. J Mol Biol 138:321–333

Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358:86–89

Murzin AG, Finkelstein AV (1988) General architecture of the α-helical globule. J Mol Biol 204:749–769

Orengo CA, Flores TP, Taylor WR, Thornton JM (1993) Identification and classification of protein fold families. Protein Eng 6:485–500

Rossmann MG, Moras D, Olsen KW (1974) Chemical and biological evolution of nucleotide binding proteins. Nature 250:194-199

Sippl MJ, Weitckus S (1992) Detection of native-like models for amino-acid-sequences of unknown 3-dimensional structure in a data-base of known protein conformations. Proteins-struct Function Genet 13:258–271

Taylor WR (1986) Identification of protein sequence homology by consensus template alignment. J Mol Biol 188:233–258

Taylor WR (1991) Toward protein tertiary fold prediction using distance and motif constraints. Protein Eng 4:853–870

Taylor WR (1993) Protein structure prediction from sequence. Comput Chem 17:117–122

Taylor WR, Orengo CA (1989) A holistic approach to protein structure comparison. Protein Eng 2:505–519

Thornton JM, Sibanda BL (1983) Amino and carboxy-terminal regions in globular proteins. J Mol Biol 167:443–460