Protein Science (1993), *2*, 884–899. Cambridge University Press. Printed in the USA.

Families and the structural relatedness among globular proteins

DAVID P. YEE AND KEN A. DILL

Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143-1204

(RECEIVED January 6, 1993; REVISED MANUSCRIPT RECEIVED February 18, 1993)

Abstract

Protein structures come in families. Are families "closely knit" or "loosely knit" entities? We describe a measure of relatedness among polymer conformations. Based on weighted distance maps, this measure differs from existing measures mainly in two respects: (1) it is computationally fast, and (2) it can compare any two proteins, regardless of their relative chain lengths or degree of similarity. It does not require finding relative alignments. The measure is used here to determine the dissimilarities between all 12,403 possible pairs of 158 diverse protein structures from the Brookhaven Protein Data Bank (PDB). Combined with minimal spanning trees and hierarchical clustering methods, this measure is used to define structural families. It is also useful for rapidly searching a dataset of protein structures for specific substructural motifs. By using an analogy to distributions of Euclidean distances, we find that protein families are not tightly knit entities.

Keywords: protein family; relatedness; structural comparison; substructure searches

Pioneering work over the past 20 years has shown that proteins fall into families of related structures (Levitt & Chothia, 1976; Richardson, 1981; Richardson & Richardson, 1989; Chothia & Finkelstein, 1990). How many families are there? Are the families "tightly knit" or "loosely knit"? That is, do two proteins within a family have much greater structural similarity than two proteins from different families? If so, they are tightly knit. What can we learn about the forces of protein folding and evolution from observing how proteins cluster into families?

In order to address these questions, it is necessary to have a suitable measure of the structural similarity between proteins, because a "family" relationship can only be defined in terms of some degree of similarity. Several measures of structural similarity have been developed (Remington & Matthews, 1978; Taylor & Orengo, 1989; Rackovsky, 1990; Sali & Blundell, 1990). There is no underlying fundamental principle dictating that one similarity measure is better than others. Ultimately, the concept of "similarity" is based upon some criterion arbitrarily chosen for a particular purpose (Maggiora & Johnson, 1990). For example, a common measure of structural similarity is the root mean square (RMS) deviation of atomic positions after superposition. RMS is a useful distance metric for comparing structures that are nearly identical: for example, when refining or comparing structures obtained from X-ray crystallography or NMR experiments. However, RMS is of limited value as a general measure of similarity because it is a "maximum likelihood estimator" of the standard deviation between two structures only if the individual errors are Gaussian-distributed with zero mean (Beers, 1957). The Gaussian distribution assumption can be reframed as an assumption that the differences between two compared structures arise from fluctuations that obey a square-law potential. A square-law potential is only a good approximation for small conformational deviations. If two structures are not in the same energy well, or if errors are large, RMS will lose its underlying justification. In addition, the use of an RMS distance criterion to compare two protein structures requires making assignments in which atom i of protein 1 "is equivalent to" atom j of protein 2. When comparing proteins with little sequence identity or unequal chain lengths, this requires making arbitrary decisions.

Some similarity measures require making structural alignments of one protein with the other (Taylor & Orengo, 1989; Sali & Blundell, 1990). When there is a biological or evolutionary basis for making these alignments, such methods have the advantage of allowing a high degree of structural discrimination among highly similar

Reprint requests to: Ken A. Dill, Department of Pharmaceutical Chemistry, Box 1204, University of California, San Francisco, California 94143-1204.

proteins. For proteins that are not highly similar, however, making alignments requires making certain arbitrary choices about the possible locations of insertions and deletions and the choices of gap penalties. These decisions can be computationally intensive.

Rackovsky (1990) has developed a similarity measure that compares distributions of conformations of chain segments up to four residues in length. Whereas it captures structural information of residues close together in sequence, our interest here is to capture information about contacting residues at all separations along the chain.

Our purpose here is better served by yet a different measure of structural relatedness. The following questions motivate the need for a different measure. What is the shape of protein conformational space? What is a useful "reaction coordinate" along which a protein folds to its native state? In models of proteins, such as those involving chains on lattices, how similar is a model conformation to the true native conformation? To address these questions, we need a similarity measure for which the two most important criteria are (1) that it must be able to compare any two conformations, no matter how different, and (2) that it must entail making the fewest possible arbitrary decisions. Furthermore, the measure must avoid comparing structures based on microscopic details such as hydrogen bond angles, since these are not appropriate for some low-resolution models. Many such problems do not involve insertions, deletions, or gaps, and therefore do not require that a similarity measure have sophisticated alignment machinery.

If an algorithm that measured structural relatedness were computationally efficient enough, it could also be put to other uses. For example, since the number of known protein structures is $P \approx 100$–$1{,}000$ (depending on whether we choose all known structures, or whether they are selected in some way to avoid repeats of nearly identical molecules), the number of pairwise comparisons involved is $[P \times (P-1)]/2 \approx 10^4$–$10^6$. If we could compute all these pairwise "distances," we could measure the interrelatedness among proteins to learn how they cluster into families. Different similarity measures make different trade-offs between speed, number of arbitrary decisions, and discrimination. By choosing a measure that is as simple, fast to compute, and nonarbitrary as possible, we trade off the degree of discrimination among highly similar proteins obtained by other measures, but the latter is less important for our purposes.

The outline of this paper is as follows. We first introduce the algorithm for measuring dissimilarity. (It measures "*dis*similarity" because it is 0 for identical structures, and increases as the structural similarity between two proteins diverges.) We call it CONGENEAL (CONformational GENEALogy) because it compares conformations and can generate family trees describing their relatedness. Much of this paper is devoted to showing that CONGENEAL is a reasonable measure of relatedness. For example, in one test we show that it is a useful tool for searching databases of protein structures to locate specified substructures within proteins. We then apply this measure to the pairwise comparison of 158 diverse protein structures and use clustering algorithms to identify families. Finally, we compare the dissimilarity distribution of protein structures to simulations of points distributed in d-dimensional Euclidean spaces to explore the tightness with which proteins cluster into families.

CONGENEAL: A dissimilarity measure

The CONGENEAL dissimilarity measure compares the weighted distance maps of two polymer conformations (see Fig. 1). The weighted distance map of a protein chain conformation that has N residues is an $N \times N$ matrix in which each matrix element (i,j) is a weight, w, equal to the distance, $d_{i,j}$, between the α-carbons of residues i and j, raised to a power $-p$ ($p > 0$):

$$w_{i,j} = d_{i,j}^{-p} \tag{1}$$

Two residues that are adjacent in space are assigned a large weight, whereas two residues that are far apart in space have a small weight. Because the matrix is symmetric, it is only necessary to compute the upper (or lower) triangle of the matrix. The difference between the weighted distance map and the contact map, first introduced by Liljas and Rossmann (1974), is that, in the former, the weights are assigned from a continuous range of values whereas, in the latter, only weights of 0 or 1 are used to indicate whether or not a pair of residues are adjacent.

The distance dependence in Equation 1 resembles that of intermolecular forces. For the purpose of dissimilarity measures, however, there are no underlying principles

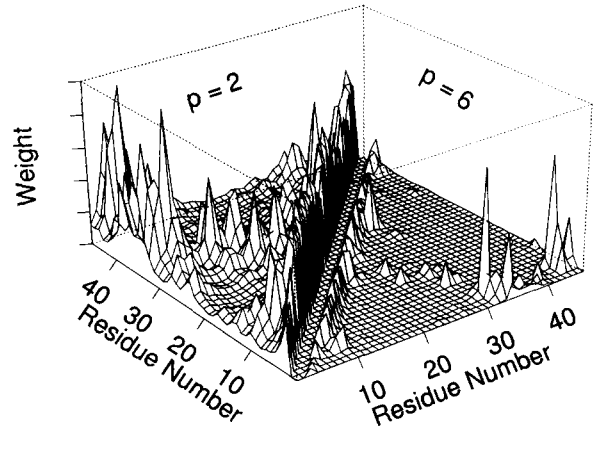

Fig. 1. Weighted distance maps of crambin with $p = 2$ on the left and $p = 6$ on the right. The height of the peaks corresponds to the magnitude of the weight, w.

that direct us to choose a particular value of p. We have investigated $p = 1$, 2, 4, and 6. When $p = 6$, only the closest neighbors contribute to the weighted distance map. When $p = 2$, pairs of residues separated in space by greater distances also contribute. We compare different values of p below, but the qualitative results are found to be sensibly independent of p. We mainly use $p = 2$.

We now describe how the dissimilarity score is obtained from the weighted distance maps for two conformations. Given two proteins, R and S, let r_{ij} be the distance between residues i and j in protein R and let s_{ij} be the distance between residues i and j in protein S. First, consider a simple case. When R and S have the same chain length, N, and have a direct residue-to-residue alignment, the dissimilarity between the two proteins is given by:

$$d(R,S) = \frac{\sum_{i=1}^{N}\sum_{j=i+2}^{N} |r_{ij}^{-p} - s_{ij}^{-p}|}{\frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=i+2}^{N} r_{ij}^{-p} + \sum_{i=1}^{N}\sum_{j=i+2}^{N} s_{ij}^{-p}\right)}. \quad (2)$$

If two proteins have identical weighted distance maps, then $d(R,S) = 0$.

Now, in order to compare proteins with different chain lengths and unknown alignments, we define a score based upon sliding one map across another, similar to a correlation function. That is, if two proteins, R and S, have chain lengths M and N, respectively, where $M \leq N$, then we calculate a series of dissimilarities as follows:

$$d'(R,S,\tau) = \frac{\sum_{i=1}^{M}\sum_{j=i+2}^{M} |r_{i,j}^{-p} - s_{i+\tau,j+\tau}^{-p}|}{\frac{1}{2}\left(\sum_{i=1}^{M}\sum_{j=i+2}^{M} r_{ij}^{-p} + \sum_{i=1}^{M}\sum_{j=i+2}^{M} s_{i+\tau,j+\tau}^{-p}\right)}, \quad (3)$$

where the range of "offsets," τ, of one weighted distance map relative to the other varies from $-M/2$ to $N - M/2$ for a total of N different alignments. The dissimilarity between the proteins R and S is then obtained by finding the offset for which the similarity is greatest:

$$d(R,S) = \min\{d'(R,S,\tau)\}. \quad (4)$$

This procedure alone, however, is not sufficient to specify a score, since the sliding of distance maps means that some (i,j) pairs of one conformation will sometimes go unpaired with (i',j') pairs in the other. For example, if $\tau = -5$, then the residue pair $(1,3)_R$ of protein R will be compared to a nonexistent pair $(-4,-2)_S$ of protein S. Therefore, there are two additional steps in the scoring method. First, the weighted distance maps are made periodic (i.e., "wrapped around") so that residue pairs are defined for all offsets. In this way, the number of compared residue pairs is the same for all alignments. Second, because a "wrapped-around" weighted distance map may imply some structural features that are not present in the actual conformation (e.g., a helix can move from the N-terminus end of a conformation to the C-terminus end), a randomization procedure is used to ensure that the dissimilarity score and alignment do not contain artifacts from using periodic weighted distance maps. The randomization procedure is as follows. For any comparison of residue pairs, $(i,j)_R$, $(i',j')_S$, involving a wrapped-around residue pair, a difference weight is not added directly to the dissimilarity score. Instead, these weights are collected in separate bins based on contact order (the contact order for residue pair (i,j) is defined as $|j-i|$, i.e., the separation of the residues along the chain). The binned distance weights from one conformation are then randomly matched with the binned distance weights from the other conformation and then added to the dissimilarity score. In practice, the offset that gives rise to the best alignment of two proteins contains few references to nonexistent (i.e., wrapped-around) residue pairs, and several different methods that we tried for treating them gave similar results.

When comparing a larger protein with a smaller one, CONGENEAL finds the part of the large protein that is most similar to the weighted distance map of the smaller protein. This feature makes CONGENEAL useful for rapidly finding specific substructures within different proteins in a structural database. To search a database for a specific substructure, CONGENEAL is used to generate scores between the substructure and each protein in the database: a small score for some alignment with a given protein locates that motif within the protein.

Validation of the dissimilarity measure

How can one validate a dissimilarity measure? For any two proteins, different measures can predict different degrees of relatedness. As noted before, there is no fundamentally correct measure of relatedness. Therefore, the validation of a dissimilarity measure ultimately depends on whether it seems sensible in light of other knowledge. In the section below, we characterize CONGENEAL in the following ways:

1. *Pairwise tests.* (a) *Finding sequence alignments.* When two different proteins contain the same substructure, a dissimilarity measure should find the sequence alignment for which the structures most closely superimpose. (b) *Using a probe protein structure to search the database.* A dissimilarity measure should find related proteins or substructures in a search of a structural database.
2. *Cluster analysis.* We compare 158 proteins pairwise and apply clustering algorithms to ask whether the dissimilarity measure finds sensible family relationships among them. We use two types of clustering

methods: minimal spanning trees and hierarchical trees based on agglomerative clustering.

All protein coordinates were obtained from the Brookhaven Protein Data Bank (PDB) (Bernstein et al., 1977; Abola et al., 1987). The set of 158 protein structures was derived from Appendix 3 of *Protein Architecture* (Lesk, 1991). Table 1 lists the proteins and their PDB filenames.

Pairwise tests

Finding sequence alignments

To first choose a few examples, it is reasonable to believe that sperm whale myoglobin (1mbd) and the A chain of human hemoglobin (1hho_a) are closely related proteins, that sperm whale myoglobin and the orange subunit of superoxide dismutase (2sod) are unrelated, and that the lysozymes from T4 bacteriophage (3lzm) and from hen egg white (1lyz) are only distantly related. Figure 2 shows the dissimilarity score as a function of alignment for these three comparisons using CONGENEAL with either $p = 2$ or $p = 6$. The point at which the score dips to a minimum (1) indicates the degree of similarity between the two proteins, and (2) gives the offset (i.e., shift) of one sequence starting position relative to the other sequence for which the structures bear closest resemblance.

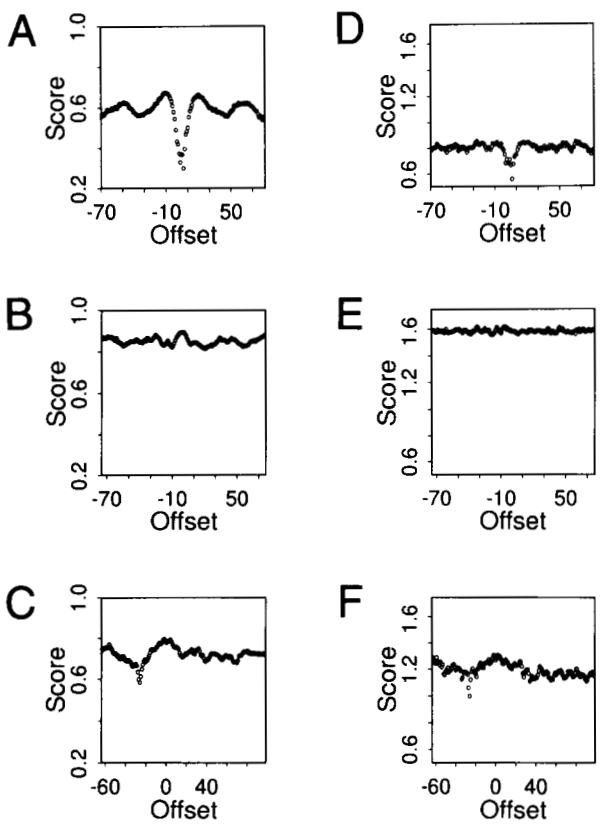

Fig. 2. Plots of dissimilarity score versus alignment. For A-C, $p = 2$. For D-F, $p = 6$. **A** and **D** show the comparison of sperm whale myoglobin with human hemoglobin. **B** and **E** show the comparison of two unrelated proteins: sperm whale myoglobin and superoxide dismutase. **C** and **F** show the comparison of two weakly similar proteins: T4 (bacteriophage) lysozyme and hen egg white lysozyme.

For the three pairwise protein comparisons mentioned above, CONGENEAL finds the expected relationships. Sperm whale myoglobin is found to be similar to the A chain of human hemoglobin with an offset of −6 residues. On the other hand, Figure 2B indicates that there is no similarity between sperm whale myoglobin and the orange subunit of superoxide dismutase. The dissimilarity measure finds hen egg white lysozyme and T4 bacteriophage lysozyme to have only a small degree of similarity. In this case, the best score is obtained at an offset of −26 residues, in agreement with the observations of Remington and Matthews (1978) and Rossmann and Argos (1976), who noted that when residues 1-80 of the phage lysozyme are aligned with residues 27-106 of the hen egg white lysozyme, there is overlap of the active sites.

Using a probe to search the database

When a probe protein or substructure is scanned through a protein database, a dissimilarity measure should properly rank order the proteins by their similarity to the probe. Below we show three examples – the helix-turn-helix DNA binding motif, the EF hand calcium binding motif, and the globin fold – for which the dissimilarity score identifies closely related protein conformations.

1. *DNA binding motif.* A number of proteins are known to have similar helix-turn-helix substructures that bind DNA. λ Cro, λ repressor, 434 Cro, 434 repressor, *trp* repressor, and catabolite gene activator protein (CAP) all have sequence similarity in a region of 22 amino acids, corresponding to the helix-turn-helix structural motif (Ohlendorf et al., 1983). How widely distributed is the helix-turn-helix motif throughout the protein database? We use CONGENEAL to search the dataset for the helix-turn-helix conformation. In our search, the helix-turn-helix substructure is defined as the 23-residue stretch from 434 Cro starting with methionine 15 and ending with glycine 37. As a simple test, Figure 3 shows the result of aligning the 434 Cro helix-turn-helix substructure with the full 434 Cro protein. The deepest minimum in Figure 3 correctly identifies the proper alignment with itself, and the score of 0 indicates that it is an exact match. The other three minima correspond to the three other turns between helices found in Cro.

The 434 Cro helix-turn-helix DNA binding substructure was then scanned across the dataset of 158 proteins. The distribution of dissimilarity scores is shown in Figure 4. Several proteins are found to have helix-turn-helix substructures similar to that of 434 Cro. Table 2 lists the 20 proteins with the greatest similarities to the target sub-

Table 1. *Key to 158-protein dataset*

Code	No. of residues	Protein name
451c	82	Cytochrome c_{551}
155c	134	Cytochrome c_{550}
256b	106	Cytochrome b_{562}
1aat	288	Aspartate aminotransferase
1abp	306	L-Arabinose binding protein
2abx	74	α-Bungarotoxin
2act	218	Actinidin
1acx	107	Actinoxanthin
6adh_a	374	Alcohol dehydrogenase
3adk	194	Adenylate kinase
2ait	74	Tendamistat
1alc	122	α-Lactalbumin
2alp	198	α-Lytic protease
4ape	330	Endothiapepsin
7api	339	α_1-Antitrypsin
3app	323	Penicillopepsin
2apr	325	Rhizopuspepsin
2atc_c	305	Aspartate transcarbamylase (regulatory subunit)
2atc_r	152	Aspartate transcarbamylase (catalytic subunit)
2aza_a	129	Azurin (*Alcaligenes denitrificans*)
3b5c	85	Cytochrome b_5
1bds	43	Sea anemone antiviral protein
3blm	257	β-Lactamase
1bp2	123	Phospholipase A_2
3c2c	112	Cytochrome c_2
2ca2	256	Carbonic anhydrase
8cat_a	498	Beef liver catalase
1cbp	86	Cucumber basic protein
1cc5	83	Cytochrome c_5
1ccr	111	Rice cytochrome *c*
2ccy_a	127	Cytochrome *c′*
2cdv	107	Cytochrome c_3
2ci2	65	Barley chymotrypsin inhibitor
3cln	143	Calmodulin
1cms	323	Chymosin B
2cna	237	Concanavalin A
5cpa	307	Carboxypeptidase A
2cpp	405	Cytochrome P450 CAM
5cpv	108	Carp parvalbumin
1crn	46	Crambin
1cro_o	66	λ Cro
2cro	65	434 Cro
1cse	274	Subtilisin carlsberg
1ctf	68	C-terminal domain of ribosomal protein L7/L12
1ctx	71	α-Cobratoxin
5cyt	103	Tuna cytochrome *c*
2cyp	293	Cytochrome *c* peroxidase
3dfr	162	Dihydrofolate reductase
5ebx	62	Erabutoxin A
1ecd	136	Erythrocruorin
1efm	130	Elongation factor TU
2enl	436	Enolase
2est	240	Porcine elastase
1etu	141	Elongation factor TU
2fb4_h	229	F_{ab} KOL (heavy chain)
2fb4_l	216	F_{ab} KOL (light chain)
1fc1	206	F_c fragment of immunoglobulin
4fd1	106	Ferredoxin
3fxn	138	Flavodoxin
3gap_c	208	Catabolite gene activator protein (closed form)
3gap_o	205	Catabolite gene activator protein (open form)
2gbp	309	Galactose-binding protein
1gcr	174	γ-Crystallin
1gd1_o	334	Glyceraldehyde-3-phosphate dehydrogenase
2gls_a	468	Glutamine synthetase
1gp1_a	184	Glutathione peroxidase
3grs	461	Glutathione reductase
1hho_a	141	Human hemoglobin
1hip	85	High potential iron protein
1hkg	457	Hexokinase
2hla_h	270	Human class 1 histocompatibility complex (heavy)
2hla_m	99	Human class 1 histocompatibility complex (β-2-microglobulin)
2hmg_1	328	Influenza hemagglutinin (HA1)
2hmg_2	175	Influenza hemagglutinin (HA2)
1hmq_a	113	Hemerythin
1hoe	74	α-Amylase inhibitor
3hvp	99	HIV protease
2il1b	153	Interleukin-1β
3icb	75	Intestinal calcium-binding protein
4ins_a	21	2Zn insulin
1kga	173	2-Keto-3-deoxy-6-phosphogluconate aldolase
2lbp	346	Leucine-binding protein
3ldh	329	Dogfish lactate dehydrogenase
2lh4	153	Lupin leghemoglobin
2liv	344	Leucine/isoleucine/valine-binding protein
1lrd	87	λ Repressor
1lyz	129	Hen egg white lysozyme
1lz1	130	Human lysozyme
3lzm	164	T4 lysozyme
1mbd	153	Sperm whale myoglobin
4mdh	334	Malate dehydrogenase
2mev_vp1	268	Mengo virus VP1
2mev_vp2	249	Mengo virus VP2
2mev_vp3	231	Mengo virus VP3
4mlt	26	Mellitin
1mon_a	44	Monellin (A chain)
1nxb	62	Neurotoxin B
2ovo	56	Ovomucoid, third domain
2pab	114	Prealbumin
9pap	212	Papain
2paz	123	Pseudoazurin
1pcy	99	Plastocyanin
4pep	326	Pepsin
1pfk_c	320	Phosphofructokinase (closed form)
1pfk_o	320	Phosphofructokinase (open form)
3pgk	416	Phosphoglycerate kinase
3pgm	230	Phosphoglycerate mutase
1phh	394	*p*-Hydroxybenzoate hydroxylase
1phy	126	Photoreactive yellow protein
2pka	232	Kallikrein A
2plv_vp1	283	Polio virus VP1
2plv_vp2	268	Polio virus VP2
2plv_vp3	235	Polio virus VP3
1pp2_r	122	Snake venom phospholipase

(*continued*)

Table 1. *Continued*

Code	No. of residues	Protein name	Code	No. of residues	Protein name
1ppt	36	Avian pancreatic polypeptide	7rsa	124	Ribonuclease A
1prc_c	332	Photosynthetic reaction center *Rhodopseudomonas viridis* C subunit	5rub_a	260	Rubisco
			5rxn	54	Rubredoxin
1prc_l	273	Photosynthetic reaction center *R. viridis* L subunit	4sbv_a	199	Southern bean mosaic virus
			2sga	181	*Streptomyces griseus* proteinase A
1prc_m	323	Photosynthetic reaction center *R. viridis* M subunit	3sgb	185	*S. griseus* proteinase B
			1sn3	65	Scorpion neurotoxin
1prc_h	258	Photosynthetic reaction center *R. viridis* H subunit	2sns	141	Staphylococcal nuclease
			2sod_o	151	Superoxide dismutase (orange subunit)
2prk	279	Proteinase K	1srx	108	Thioredoxin
1pte	348	Carboxypeptidase/transpeptidase	2ssi	107	*Streptomyces* subtilisin inhibitor
5pti	58	Bovine pancreatic trypsin inhibitor	2stv	184	Satellite tobacco necrosis virus
4ptp	223	Trypsin	2taa	478	Taka-amylase
1pyp	280	Pyrophosphatase	2tbv	286	Tomato bushy stunt virus
1r69	63	434 Repressor (N-terminal domain)	1tec	279	Thermitase
1rbb_a	124	Ribonuclease B	1thi	207	Thaumatin I
1rei	107	Immunoglobulin V_κ domain	1tim	247	Chicken triosephosphate isomerase
1rhd	293	Rhodanese	2tmv	154	Tobacco mosaic virus
2rhe	114	Immunoglobulin V_λ domain	4tnc	160	Troponin C
4rhv_vp1	273	Rhinovirus VP1	1tnf	152	Tumor necrosis factor
4rhv_vp2	255	Rhinovirus VP2	1ubq	76	Ubiquitin
4rhv_vp3	236	Rhinovirus VP3	1utg	70	Uteroglobin
3rn3	124	Ribonuclease A	9wga	171	Wheat germ agglutinin
1rns	72	Ribonuclease S	1wrp	102	*trp* repressor
2rnt	104	Ribonuclease T_1	4xia	393	Xylose isomerase
3rp2	224	Rat mast cell protease	2yhx	457	Hexokinase

structure. All seven proteins known to have the DNA binding helix-turn-helix substructure are in this group. In all seven cases, the predicted alignment of the substructure with the protein is identical to the alignment produced by sequence analysis (Ohlendorf et al., 1983). The protein with the most similar substructure (other than 434 Cro) is 434 repressor. This is consistent with the observation of Mondragon et al. (1989) that the amino terminal domain of 434 repressor is remarkably similar to 434 Cro and that the substructures are virtually identical. The DNA binding protein that is least similar to 434 Cro is *Escherichia coli trp* repressor. The latter differs from the other DNA binding proteins in two respects: (1) the end of the first helix is more open, and (2) the orientation of

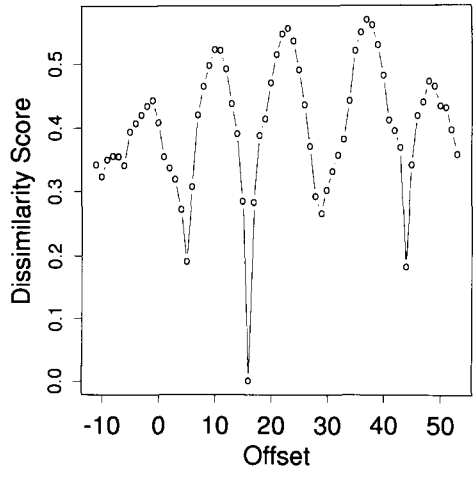

Fig. 3. Dissimilarity score versus alignment of the 434 Cro DNA binding substructure with the complete structure of 434 Cro.

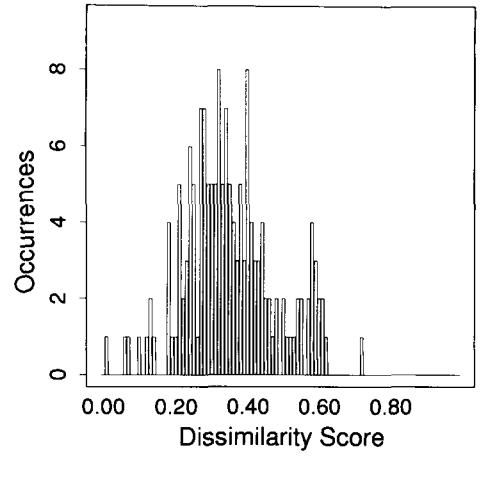

Fig. 4. Histogram showing the distribution of dissimilarity scores when the 434 Cro DNA binding substructure is compared with a dataset of 158 proteins.

Table 2. *Proteins with helix-turn-helix motif*

Protein name	PDB filename	CONGENEAL	RMS
434 Cro	2cro	0.000	0.000
434 Repressor (N-terminal domain)	1r69	0.052	0.380
λ Cro	1cro_o	0.068	0.585
λ Repressor	1lrd	0.099	0.830
Enolase	2enl	0.110	1.634
Catabolite gene activator protein (open form)	3gap_o	0.120	1.135
Catabolite gene activator protein (closed form)	3gap_c	0.126	1.068
Cytochrome P450 CAM	2cpp	0.135	1.742
C-terminal domain of ribosomal protein L7/L12	1ctf	0.170	1.938
Xylose isomerase	4xia	0.177	2.047
Cytochrome *c* peroxidase	2cyp	0.178	2.053
Photosynthetic reaction center *R. viridis* M subunit	1prc_m	0.178	2.963
Photosynthetic reaction center *R. viridis* L subunit	1prc_l	0.189	2.835
trp repressor	1wrp	0.197	1.729
Beef liver catalase	8cat_a	0.203	2.652
Erythrocruorin	1ecd	0.205	2.978
Proteinase K	2prk	0.205	2.543
Glutamine synthetase	2gls_a	0.207	3.694
Hemerythin	1hmq_a	0.209	4.206
Sperm whale myoglobin	1mbd	0.210	4.531

the second helix in the helix-turn-helix substructure is constrained by the binding of L-tryptophan (Schevitz et al., 1985).

The seven DNA binding proteins rank 1, 2, 3, 4, 6, 7, and 14 in similarity to the probe helix-turn-helix structural motif. Some non-DNA binding proteins also score well for the presence of the helix-turn-helix substructure (see Table 2). In many cases, the best match of the substructure to a protein occurs when the helix-turn-helix substructure is aligned with the last half of a long helix, a turn, and the first few residues of the following helix. Two of the proteins identified here as having a substructure similar to that of the probe substructure have been previously noted by Richardson and Richardson (1988). They found that cytochrome *c* peroxidase and ribosomal L7/L12 protein contain conformations similar to the DNA binding helix pairs in gene activator and repressor proteins. In the present analysis, the non-DNA binding protein that has a substructure most similar to the probe DNA binding substructure is yeast enolase. Enolase has two domains consisting of (1) a three-stranded β meander and four α-helices and (2) an eightfold α/β barrel (Lebioda et al., 1989). The helix-turn-helix substructure of 434 Cro aligns with enolase near the end of the N-terminal domain. Figure 5 shows the top eight alignments found by CONGENEAL for substructures from DNA binding proteins or non-DNA binding proteins with the DNA binding substructure of 434 Cro (see also Kinemage 1).

2. *Calcium binding: The EF hand.* Another well-characterized substructural motif is the EF hand calcium binding conformation, first described by Kretsinger and Nockolds (1973) from carp muscle calcium binding parvalbumin (5cpv). The EF hand is also found in several other proteins that bind calcium, including calmodulin (3cln), troponin C (4tnc), and intestinal calcium binding protein (3icb). Although CONGENEAL finds relatively few substructures identical to EF hands, many proteins contain substructures that are fairly similar.

We define the EF hand substructure in carp parvalbumin as the 29 residues from asparagine 79 to lysine 107. The E helix is 12 residues (79–90), the loop is 8 residues (91–98), and the F helix is 9 residues long (99–107). Figure 6 shows the result of aligning this substructure with the complete structure of carp parvalbumin. There are five minima, corresponding to the joining regions between the six helices of parvalbumin (labeled A–F): AB, BC, CD, DE, and EF. Strong matches are found at two positions, corresponding to C-loop-D and E-loop-F. Both of these substructures are in the EF hand conformation. Kretsinger and Nockolds (1973) suggested that A-loop-B is related to the EF hand, but our results do not show significant structural similarity between A-loop-B and the EF hand substructure. In fact, the A and B helices are oriented nearly parallel to one another, whereas the helices in an EF hand are nearly perpendicular.

We then use the EF hand substructure as a probe to search the dataset of 158 proteins. Figure 7 shows the distribution of dissimilarities. The four best scoring proteins are all calcium binding proteins (see Table 3). The EF hand in troponin C was found to be the most similar to the parvalbumin structure; the dissimilarity scores as a function of alignment are shown in Figure 8. Minima identify the four EF hand substructures in troponin C. The two minima on the right in Figure 8 indicate two substructures that are the most similar to the parvalbumin substructure and correspond to the EF hands nearest the C-terminal end of the protein. The two minima on the left identify two N-terminal EF hands that are less similar to the parvalbumin EF hand. Interestingly, the N-terminal EF hands do not bind calcium (Satyshur et al., 1988).

The method correctly identifies bovine intestinal calcium binding protein as being similar to the EF hand. On the other hand, given that carp parvalbumin and bovine intestinal calcium binding protein are related, the RMS deviation of C_α positions between their EF hands seems to be surprisingly large. The large RMS deviation arises because the residues at both ends of the substructure have different conformations in the two proteins. The next most similar substructure to the EF hand is in T4 lysozyme (see Kinemage 2); this similarity was first noted by Tufty and Kretsinger (1975).

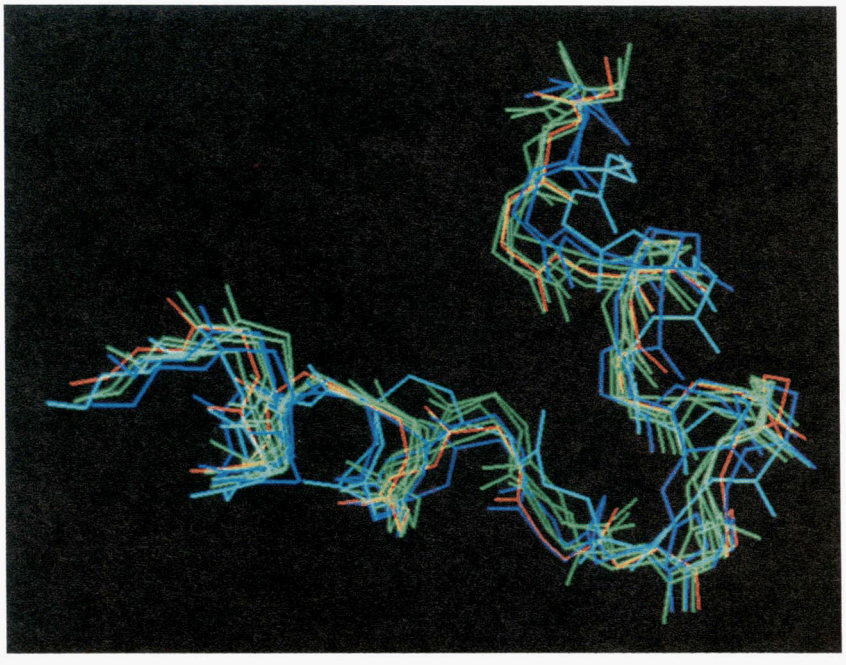

Fig. 5. Structural alignment of 434 Cro DNA binding motif to top 8 matches identified by CONGENEAL and *trp* repressor, which scored 14th. 434 Cro DNA binding substructure is shown in green, DNA binding proteins are shown in red, and non-DNA binding proteins are shown in blue. *trp* repressor, the DNA binding protein least similar to the target 434 Cro substructure, is shown in cyan.

3. *Searching protein structures for functional subunits.* The ability to perform fast searches for substructures within proteins allows for searching a database of protein structures for specific functional subunits. We show an example of using CONGENEAL to find possible calcium binding proteins in the dataset. In the EF hand substructure, the calcium is bound to residues within the loop region. Therefore, we search the dataset for the presence of the E helix, the F helix, and the loop. E and F are α-helices, so all proteins with α-helices score well for the presence of the E and F helix substructures (data not shown). Figure 9 shows the distribution of dissimilarities with the calcium binding loop. Although most proteins are predicted to have at least one short loop conformation similar to the calcium binding loop, five proteins are clearly distinct as being more similar than the other proteins. They include the four calcium binding proteins identified above. Hence, the weighted distance map for this loop region is a good identifier of the calcium binding motif. In addition, galactose binding protein scores well for the presence of a calcium binding loop. Consistent with this finding, galactose binding protein was reported to have a calcium binding site (Vyas et al., 1987) that resembles the EF hand without the helices. T4 lysozyme, which scored 5th for the presence of an EF hand, scores 69th for the presence of the calcium binding loop. Although T4 lysozyme has two helices similar in orientation to the EF hand helices in parvalbumin, the intervening loop is clearly not in the calcium binding conformation.

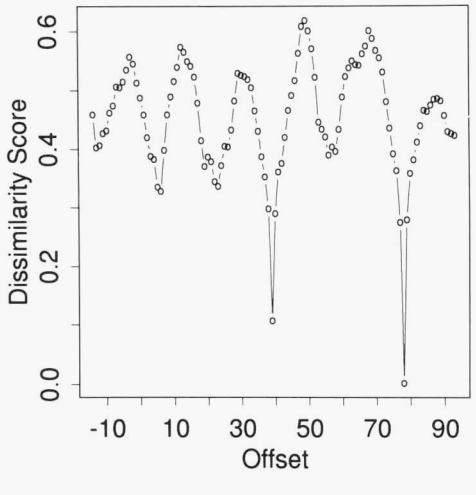

Fig. 6. Dissimilarity score versus alignment of the carp parvalbumin EF hand substructure with the complete structure of carp parvalbumin.

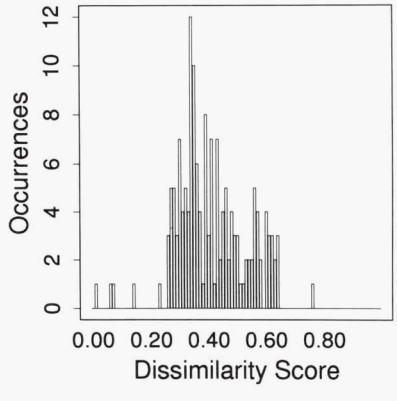

Fig. 7. Histogram showing the distribution of dissimilarity scores when the EF hand substructure is compared with a dataset of 158 proteins.

Table 3. *Proteins with EF hand*

Protein name	PDB code	CONGENEAL	RMS
Carp parvalbumin	5cpv	0.002	0.000
Troponin C	4tnc	0.054	0.644
Calmodulin	3cln	0.064	0.987
Intestinal calcium-binding protein	3icb	0.135	2.868
T4 lysozyme	3lzm	0.227	2.994
Sperm whale myoglobin	1mbd	0.251	5.199
Human hemoglobin	1hho_a	0.253	5.143
Subtilisin carlsberg	1cse	0.255	4.107
Cytochrome P450 CAM	2cpp	0.261	6.142
Erythrocruorin	1ecd	0.264	4.991
Lupin leghemoglobin	2lh4	0.265	4.612
Hemerythin	1hmq_a	0.266	5.050
Enolase	2enl	0.269	4.524
Thermitase	1tec	0.270	4.159
Cytochrome *c* peroxidase	2cyp	0.275	4.565
trp repressor	1wrp	0.277	4.272
Cytochrome b_{562}	256b	0.277	4.474
Proteinase K	2prk	0.280	5.069
Leucine-binding protein	2lbp	0.282	3.518
Malate dehydrogenase	4mdh	0.282	5.295

4. *Globins.* Figure 10 shows the dissimilarities of sperm whale myoglobin (1mbd) to the set of 158 protein structures. The four most similar structures are all globins: sperm whale myoglobin (1mbd), erythrocruorin (1ecd), human hemoglobin (1hho_a), and leghemoglobin (2lh4). The next most similar proteins are all dominated by α-helices. They include uteroglobin (1utg), *trp* repressor (1wrp), calcium binding protein (3icb), and cytochrome b_{562} (256b). The proteins least similar to myoglobin are all β-sheet proteins: they include immunoglobulin fragments (2fb4_h, 2fb4_l, 1fc1), tumor necrosis factor (1tnf), monellin (1mon_a), and Cu,Zn superoxide dismutase (2sod). The dissimilarity distribution from CONGENEAL resembles an earlier comparison made by Bowie et al. (1991) of sperm whale myoglobin versus a protein dataset based on their three-dimensional (3D) profiling method. Their method shows the degree to which the *sequences* of other globins are compatible with the *structure* of sperm whale myoglobin. Our method shows the degree to which the *structures* of other globins are similar to the *structure* of sperm whale myoglobin. At least for sperm whale myoglobin, the distributions from 3D profiling and CONGENEAL bear considerable resemblance.

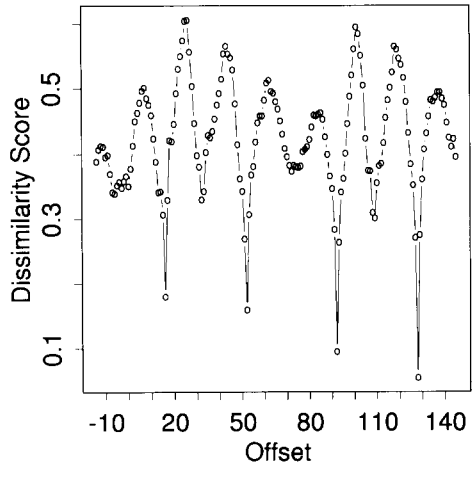

Fig. 8. Dissimilarity score versus alignment of the EF hand substructure with troponin C. The four minima correspond to the four EF hand substructures in troponin C.

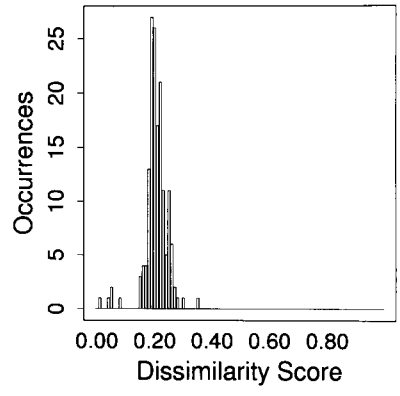

Fig. 9. Histogram showing the distribution of dissimilarity scores when the calcium binding loop from carp parvalbumin is compared with a dataset of 158 proteins.

Most closely related proteins

As another test, we compare the 158 proteins pairwise ($[P \times (P-1)]/2 = 12{,}403$ tests) and ask which pairs are the most closely related. Of course, because this is not a "selected" set of unrelated conformations, some of these

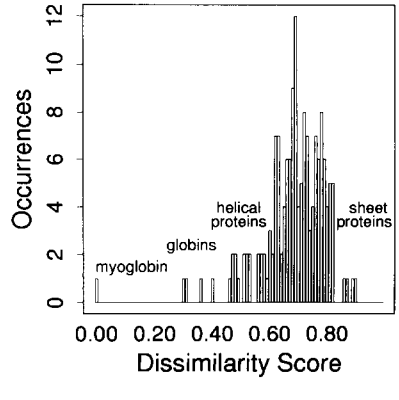

Fig. 10. Histogram showing the distribution of dissimilarity scores when sperm whale myoglobin is compared with a dataset of 158 proteins.

proteins are quite similar; these highly similar pairs are controls that we study here. Table 4 lists the 20 most similar protein pairs. Not surprisingly, several of these pairs represent the same protein in two different conformations. For example, six of the closest structural similarities are pairs of ribonucleases. The dataset contains four ribonuclease structures: two independently determined structures of ribonuclease A (3rn3, 7rsa), ribonuclease B (1rbb_a), and ribonuclease S (1rns). Each pair of structures involves only a small structural variation. For example, ribonucleases A and B have identical amino acid sequences, but differ by a polysaccharide moiety that is attached to asparagine 34 of ribonuclease B.

Table 4. *Most closely related proteins in dataset*

Pair number	Protein name	PDB code	Score
1	Ribonuclease A Ribonuclease A	7rsa 3rn3	0.014
2	Phosphofructokinase (open form) Phosphofructokinase (closed form)	1pfk_o 1pfk_c	0.035
3	Rice cytochrome *c* Tuna cytochrome *c*	1ccr 5cyt	0.052
4	Ribonuclease A Ribonuclease B	3rn3 1rbb_a	0.056
5	Ribonuclease A Ribonuclease B	7rsa 1rbb_a	0.057
6	434 Cro 434 Repressor (N-terminal domain)	2cro 1r69	0.072
7	Erabutoxin A Neurotoxin B	5ebx 1nxb	0.076
8	F_{ab} KOL (light chain) Immunoglobulin V_λ domain	2fb4_l 2rhe	0.079
9	Ribonuclease A Ribonuclease S	3rn3 1rns	0.103
10	Ribonuclease A Ribonuclease S	7rsa 1rns	0.103
11	Hexokinase Hexokinase	2yhx 1hkg	0.114
12	Ribonuclease B Ribonuclease S	1rbb_a 1rns	0.116
13	CAP (closed form) CAP (open form)	3gap_c 3gap_o	0.125
14	Tendamistat α-Amylase inhibitor	2ait 1hoe	0.136
15	Leucine-binding protein Leu/Ile/Val-binding protein	2lbp 2liv	0.140
16	Human lysozyme Hen egg white lysozyme	1lz1 1lyz	0.222
17	Rhinovirus VP3 Polio virus VP3	4rhv_vp3 2plv_vp3	0.231
18	λ Repressor 434 Repressor (N-terminal domain)	1lrd 1r69	0.237
19	λ Repressor 434 Cro	1lrd 2cro	0.289
20	Elongation factor TU Elongation factor TU	1etu 1efm	0.277

Ribonuclease B is about as similar to ribonuclease A as the two ribonuclease A are to each other. This result is consistent with the conclusion of Williams et al. (1987) that the conformation of ribonuclease B is not significantly different from that of ribonuclease A. The small variability occurs mostly in the β-sheet regions.

CONGENEAL also finds the correct alignment of ribonuclease S with the other ribonucleases (i.e., at an offset of 21 residues). All the ribonuclease structures are quite similar to each other, but ribonuclease S is the least similar among them. The deviations of ribonuclease S relative to the other ribonuclease structures are attributable to the contacts formed by residues 21–23 with the rest of the protein. Because ribonuclease S is formed by cleavage of ribonuclease A between alanine 20 and serine 21, the conformations of residues 21–23 presumably readjust in response to the cleavage of the peptide bond between residues 20 and 21.

CONGENEAL finds other highly similar pairs. It finds rice cytochrome *c* to be very similar to tuna cytochrome *c*. The rice structure has 8 additional residues at the N-terminus, and there are 43 substitutions in the other 103 residues. Despite these sequence differences, the structures are found to be nearly identical (Matthews, 1985). Other sets of proteins that are found to be highly similar by CONGENEAL include DNA binding proteins (434 Cro, 434 repressor, λ repressor), neurotoxins (erabutoxin A, neurotoxin B), immunoglobulins (F_{ab} KOL, immunoglobulin V_λ domain), and viral VP3 domains (rhinovirus VP3, polio virus VP3).

CONGENEAL has limitations. First, it does not treat insertions, deletions, or gaps. An example of this limitation is in the comparison of α/β barrels such as triose phosphate isomerase (TIM). It has been suggested that all the known α/β barrels may have diverged from a common ancestor (Farber & Petsko, 1990). If so, and if the process of evolutionary divergence involves changing loop lengths while retaining secondary structural domains, then evolutionary "distance" requires a similarity measure that carries only weak penalties for changing lengths of loops between domains. Although some similarity methods do this (Taylor & Orengo, 1989), CONGENEAL does not, and therefore would not be useful as a measure of evolutionary divergence by this mechanism. Hence, again we caution that different similarity measures will find different degrees of relatedness among proteins and will find different family clusters, but there is no unique right way to do this. And we note that the approach taken in CONGENEAL, while it is disadvantageous for measuring evolutionary divergence by this mechanism, is advantageous

for other purposes, because it is based on making no assumptions about mechanisms of how one conformation is caused to differ from another. Such a need arises in the comparison of conformations of a given sequence, in which case there are no gaps, insertions, or deletions, or in the comparison of very different conformations that may not be related by a known evolutionary mechanism, in which case we believe it may often be preferable to measure similarity with an algorithm having a minimum number of degrees of freedom.

Second, when comparing sets of proteins with different alignments and chain lengths, the dissimilarity measure is not a true distance metric. That is, as with many other similarity measures, the triangle inequality law,

$$d(a,b) + d(b,c) \geq d(a,c),$$

can be violated. For example, in the pairwise comparison of a sheet (S), a helix (H), and a protein consisting of both a sheet and a helix (P):

$$d(S,P) = 0 \quad \text{and} \quad d(P,H) = 0,$$

but

$$d(S,H) > 0.$$

Third, as with other contact-map based approaches, CONGENEAL does not distinguish structures by their chiralities. A molecule is indistinguishable from its mirror image. For comparing molecules with consistent chiralities, such as two real proteins, this is not a limitation. For comparing a lattice model and a real protein, however, chiral errors will not be detected. In a most general way, CONGENEAL only attempts to characterize distances pertinent to nonlocal interactions. In this sense, right-handed and left-handed helices are similar. When it is important to distinguish them, CONGENEAL is not appropriate.

Protein clustering into families

CONGENEAL is a measure that computes the structural similarity between any two compact polymer conformations. We have shown a few tests indicating where it is sensibly consistent with other knowledge. We now use this measure to study how it divides proteins into families. We define a family as a set of structures that collectively share a high degree of similarity to one another. The concept of family carries the implication that there are relatively sharp boundaries between families. Given a measure of similarity, there are several different methods for identifying clustering. As with similarity measures, there are no right or wrong clustering methods. In order to determine whether the families obtained are sensitive to the choice of clustering method, we study the clustering of protein structures by two different methods: a minimal spanning tree and a hierarchical method. Different similarity measures and clustering methods can lead to different, but equally valid, divisions of proteins into families.

Clustering by minimal spanning trees

First, we construct a minimal spanning tree, which is a graph that provides one way to describe relatedness among proteins. Consider a graph in which each one of the P protein structures is represented by a node. Every possible pair of nodes is connected by an edge. Each edge is weighted by the dissimilarity score relating the two proteins. Hence, there are $[P \times (P-1)]/2$ edges. A spanning tree is a subgraph in which there are only $P-1$ edges connecting the P vertices (proteins). A minimal spanning tree is a spanning tree in which the sum of the weights of the edges is as small as possible. Thus the only connectivity is among the most similar proteins. We construct a minimal spanning tree using Kruskal's algorithm (Horowitz & Sahni, 1978), as follows. First, the pairwise scores are sorted from most similar to least similar. The tree is then constructed edge by edge. The first edge is defined as the protein pair with the lowest score (highest similarity). The second edge is chosen to be the next lowest score that does not lead to a cycle in the graph. If a cycle were formed, then there would be more than one path between two vertices, and at least P edges for P vertices, thus violating the criterion of a spanning tree. The process continues until there are $P-1$ edges.

Figure 11 shows the minimal spanning tree for the set of 158 protein structures based on the CONGENEAL dissimilarity measure. A tree is unique provided that no two edges have the same weight. Note that no meaning should be attributed to the edge lengths shown in the figure, because they are not drawn in proportion to their respective dissimilarity weights. An edge connecting two proteins implies structural similarity between the two proteins.

By this clustering method, proteins are found to collect around hubs, which may be thought of as consensus family structures or structural paradigms. For example, monellin (1mon_a), uteroglobin (1utg), crambin (1crn), and λ repressor (1lrd) are all hubs. Many other proteins are connected to each hub. Each hub represents some characteristic topological feature (i.e., some specific protein fold). For example, the A chain of monellin forms three strands of an antiparallel β-sheet. Any protein in the dataset that has three strands in a similar conformation will score well when compared to monellin, and may be connected to the monellin hub. Similarly, uteroglobin, a progesterone binding protein consisting of four α-helices, is a hub for structures with similar helical and turn features. Crambin has both β-sheet and α-helix and serves as a hub for proteins with similar secondary structural features.

Proteins are found to cluster into families, often around hubs. For example, the globins cluster together. Lyso-

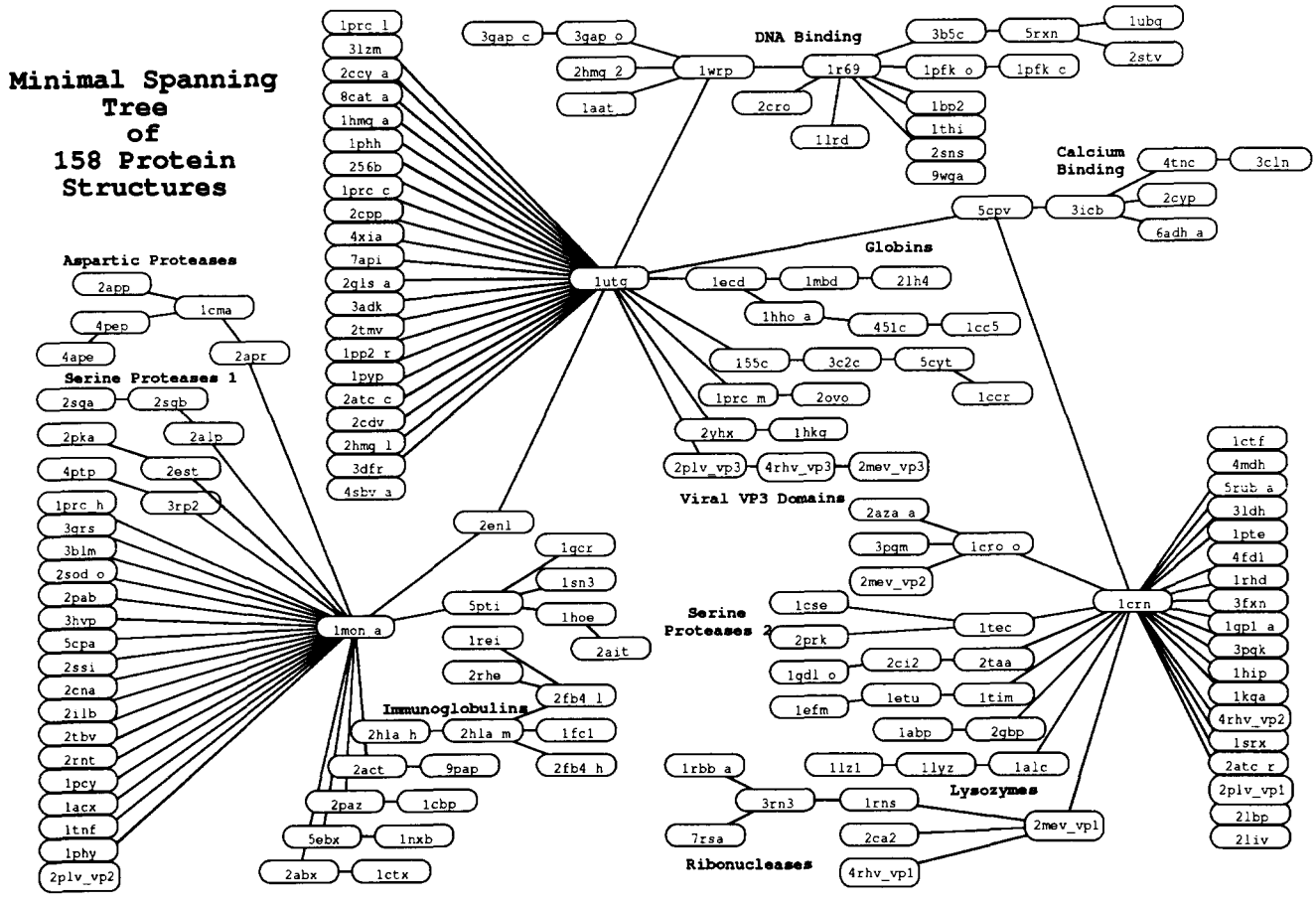

Fig. 11. Minimal spanning tree of 158-protein dataset. Proteins are referenced by the codes listed in Table 1. An edge connecting two proteins implies structural similarity between them. Edge lengths are not proportional to the dissimilarity between proteins. As a guide, the general locations of some major family relationships are indicated in bold.

zymes from hen egg white and humans cluster with α-lactalbumin. Other protein clusters include (1) viral VP3 domains, (2) cytochrome *c* structures, (3) immunoglobulin domains, (4) aspartic proteases, (5) trypsin-like serine proteases, and (6) subtilisin-like serine proteases. Interestingly, T4 bacteriophage lysozyme is separated by four nodes from the other lysozymes. In this case, despite the structural similarity of the active sites, the remainder of the structure of T4 phage lysozyme is different from that of the other lysozymes.

Hierarchical clustering

In order to learn whether the family partitions found by CONGENEAL depend on the clustering method, we now consider a different clustering algorithm for collecting proteins into families. Here we use hierarchical clustering, which successively groups proteins into increasingly larger sets. At first, there are P proteins in P groups. Step 1 is to combine the two most similar proteins to form the first group; there are now $(P-2)$ single-protein groups and one two-protein group. This is recorded as the first decision. Step 2 is to combine the two groups that now have the greatest similarity. To determine group similarities, all pairwise dissimilarities between groups are calculated; this generates $[M \times (M-1)]/2$ average dissimilarities for M groups. The group dissimilarity is the average of the dissimilarities between the members of one group with respect to the members of another group. The merging process is repeated until all groups are combined into a single group. The merging process is a sequence of decisions that can be represented as a tree (see Fig. 12). Nodes at a given level in this tree represent a given degree of dissimilarity.

As with the spanning tree, the hierarchical method finds that immunoglobulins, serine proteases, ribonucleases, globins, aspartic proteases, and viral VP domains form families. Hence, the general division into these families appears to be relatively independent of the clustering method, although the details differ.

One interesting consequence of the hierarchical clustering is evident from Figure 12. It leads to a partitioning of families in which sheet structures are concentrated at the top of Figure 12 and helix structures are concentrated at the bottom. According to this partitioning, sheet struc-

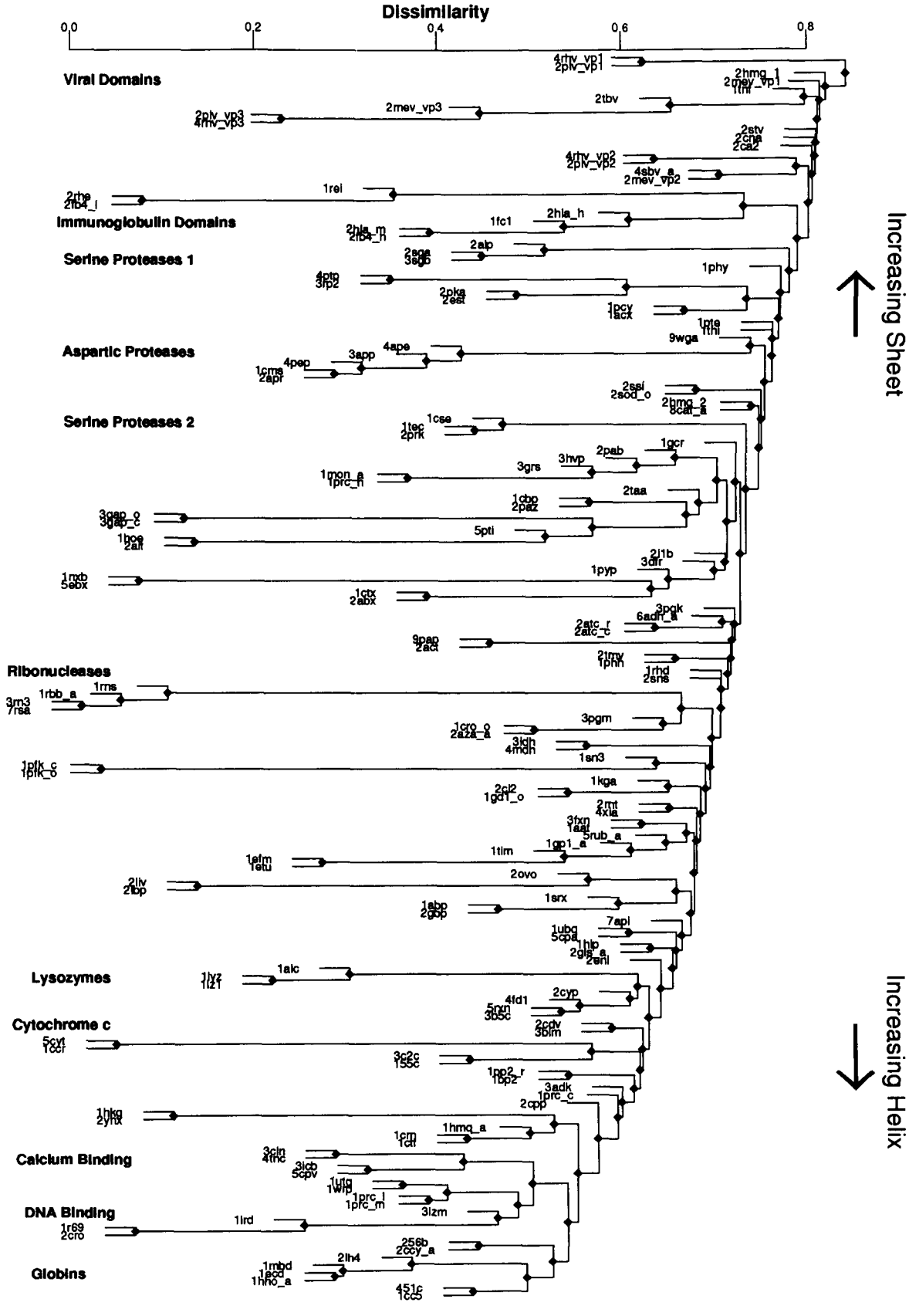

Fig. 12. Hierarchical clustering of 158 proteins into a relatedness tree. The mean dissimilarities between the group members are indicated by diamonds at the branch points of the tree. Proteins are referenced by the codes listed in Table 1. As a guide, some major family relationships are indicated.

tures are less related to one another than are helical structures. Helical structures are more related to one another because of the regular pattern of close contacts formed by residues in the helical conformation.

Are proteins tightly clustered?

Are protein families tight or loose forms of organization? "Tight" organization means that any two proteins within a family are much more similar than any two proteins from different families. There would be a sharp boundary between protein families. "Loose" organization means that two proteins within a family may be only slightly more similar than two proteins taken from different families. The boundaries between protein families would not be sharp and might even overlap.

We assess tightness of protein families by studying the shape of the histogram of pairwise dissimilarities (Fig. 13). Qualitatively, if proteins are tightly clustered, the histogram in Figure 13 would have two peaks: one representing the high similarities within families and the other representing the low similarities between families. The distribution of dissimilarities for the 158 protein dataset structures, however, shows mainly a single broad peak, which indicates a wide range of relatedness among proteins. There is only a very small peak indicating high similarities and tight families. The mean dissimilarity is 0.737, indicating that two arbitrary proteins are relatively unrelated. By this qualitative criterion, protein structural families are only loose entities.

A second way to assess the tightness of clustering draws on the analogy between (1) P proteins as points separated by their pairwise dissimilarities and (2) a set of P points distributed in a d-dimensional space separated by their Euclidean distances. To pursue this analogy, we generate points in Euclidean spaces with varying degrees of clustering in several different dimensionalities, d. The distribution of distances between the points is compared to the distributions of pairwise protein dissimilarities shown in Figure 13. We create varying degrees of clustering as follows. First, we assume there are f families of points in a d-dimensional space. We randomly generate f points that represent the family centers. Within each family, we then generate P/f points which are Gaussian-distributed around each family center. That is, the probability distribution for a point x within a family is given by:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma k} \exp\left[\frac{-d^2(x,c)}{2\sigma^2 k^2}\right],$$

where c is a family center, $d(x,c)$ is the distance between x and c, k is the average distance between any two family centers, and σ is the parameter that controls the degree of clustering. When σ equals 1, the standard deviation of points around a family center is equal to the average distance between the family centers. As σ decreases, the tightness of the clustering increases. As an example, Figure 14 shows scatter plots of points distributed in two dimensions around 15 "families" with three different values of σ.

After randomly generating points as described above, we calculate all the pairwise distances between the points within each set. Figure 15 shows histograms for $N = 200$, $d = 7$, $f = 25$, and varying σ. When σ is small (tight clustering) the histograms have two peaks, as expected. The

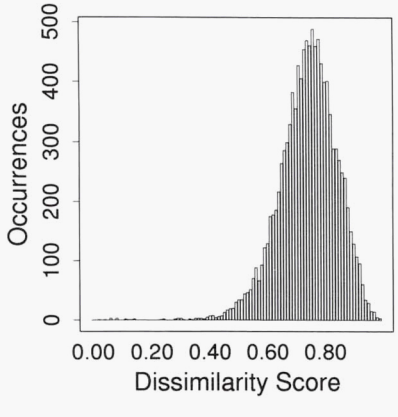

Fig. 13. Histogram showing distribution of 12,403 pairwise comparisons of the 158-protein dataset.

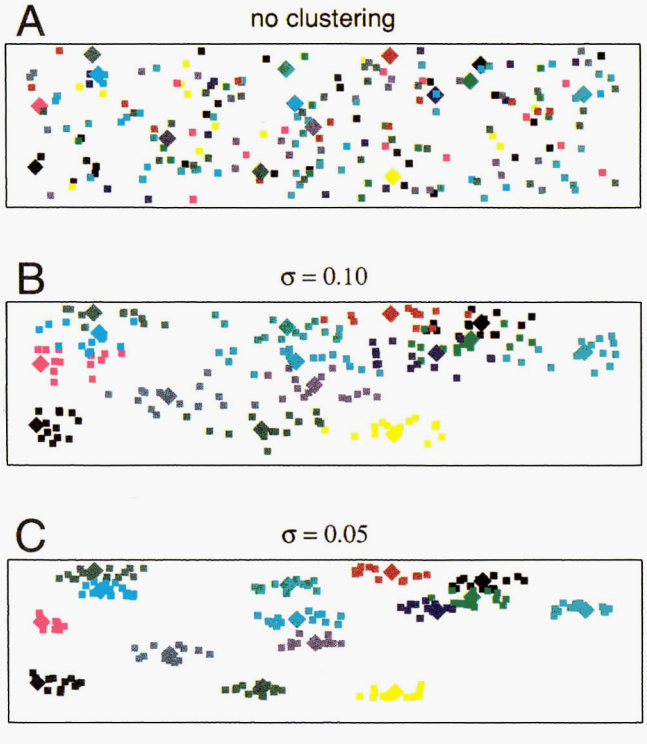

Fig. 14. Degrees of clustering using a Euclidean distance analogy: increasing the tightness of clustering from A to B to C. Scatter plots of points randomly distributed around family centers. Family centers are represented by large diamonds. Points associated with a family center are represented as small squares. In **A–C**, the number of family centers is 15 and the total number of points is 200.

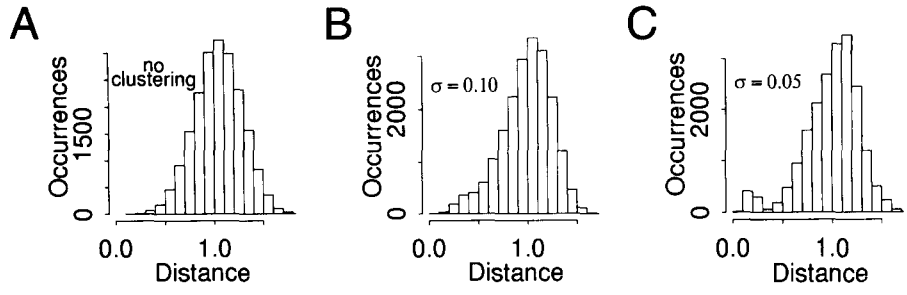

Fig. 15. Distribution of pairwise distances between points randomly distributed between 25 families within a seven-dimensional sphere.

leftmost peak is due to intrafamily distances and the rightmost peak is due to interfamily distances.

One method to compare the shapes of two distribution functions is the quantile-quantile (Q-Q) plot (Chambers et al., 1983). A Q-Q curve plots the sorted values of one distribution against the sorted values of a second distribution. If two distributions have the same shape, and differ only by a multiplicative factor that scales the width, or if they differ by a constant factor that shifts the mean values, then a Q-Q plot gives a straight line. Figure 16A shows a Q-Q plot of the histogram of Figure 13 versus a nonclustered distribution of points in Euclidean space. The deviations at both extremes of the plot indicate that protein structures are more broadly distributed than would be predicted by the completely nonclustered distribution. Hence, a nonclustered uniform distribution of points is not a good model for the distribution protein dissimilarities.

On the other hand, Figure 16C shows that proteins are also not well represented as being very tightly clustered ($\sigma \leq 0.05$). When the distributions are tightly clustered, the Euclidean distribution underestimates the number of similar protein pairs in the range of dissimilarities between 0.40 and 0.60, and overestimates them for distances less than 0.40. The closest correspondence between the Euclidean distances and protein similarities is obtained for values around $\sigma = 0.10$. This is the case for which the Q-Q plot is most linear (see Fig. 16B). While this value of σ implies that family members are considerably closer together than are interfamily centers, protein families are not sufficiently tightly knit to avoid considerable overlap between families, and individuals cannot be unambiguously assigned to families. This degree of clustering is indicated schematically in Figure 14B for a two-dimensional Euclidean space. By systematically varying the clustering parameter, σ, and the dimensionality, d, we find the optimal dimensionality to be about $d = 7$. It is not clear to us if this dimensionality in the Euclidean space analogy has any physical meaning since our dissimilarity measure is not a true metric. This comparison should be viewed simply as an analogy.

Our results are consistent with those of Rackovsky (1990). His similarity measure, which is based on local conformational preferences, also orders proteins from helices to sheets and finds families to be only loosely knit entities.

Conclusions

We have described a simple quantity for characterizing the structural similarity between any two compact polymer or protein conformations. Based on differences between weighted distance maps, it requires no alignments or gap penalties and makes few assumptions or arbitrary decisions about polymer structure. It is computationally fast. The only parameter is the exponent p in the distance de-

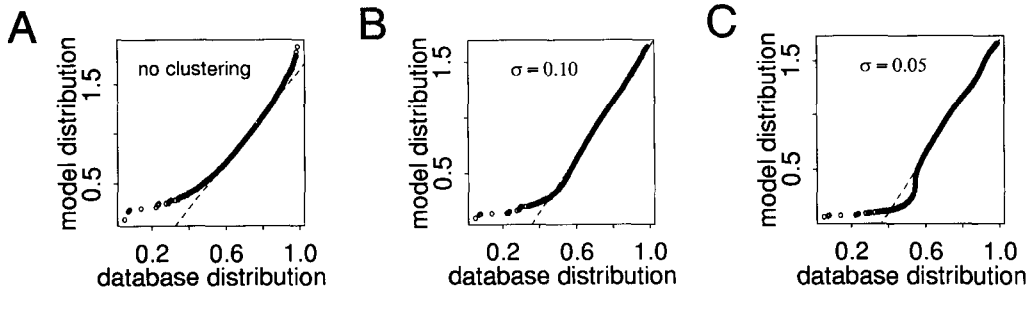

Fig. 16. Q-Q plots comparing protein dataset pairwise dissimilarity distribution with distributions of random point pairwise distances. For the random point distributions, $f = 25$ and $d = 7$.

pendence of the weights. The results are not very sensitive to this parameter.

The method can compare any two conformations, no matter how similar or different, and does not require identical chain lengths. It is intended for the purpose of comparing diverse polymer conformations. Several tests show that the relatedness among proteins reported by this measure is sensibly consistent with existing knowledge. This method can be used to rapidly search through a database of protein structures to find specified substructures and to seek given functional components in other proteins. For example, it searches the protein dataset in minutes to find possible calcium binding motifs similar to the EF hand.

Such a similarity measure can be used to test algorithms of protein folding for which generated conformations may be distant from the native structure. Hence, the CONGENEAL measure can serve as a sort of "reaction coordinate" for nativeness. For such problems, gaps are unimportant.

We combine this measure with two different clustering methods to identify protein families. We then ask how tightly clustered are families by drawing an analogy with points distributed in Euclidean space. The analogy indicates that protein families are only loosely knit entities, and that individual proteins may often not be unambiguously assignable to a unique family.

Supplementary material

This work would not have been possible without the enormous amount of effort required to experimentally determine the protein structures used in this analysis. References for all the structures obtained from the PDB are available on the Diskette Appendix and by request from the authors.

Acknowledgments

We thank Sarina Bromberg, Fred Cohen, Kai Yue, and Chris Carreras for many helpful discussions, and the DARPA University Research Initiative program and the NIH for financial support. Molecular graphics images were produced using the MidasPlus software system from the Computer Graphics Laboratory, University of California, San Francisco.

References

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., & Weng, J. (1987). Protein Data Bank. In *Crystallographic Databases—Information Content, Software Systems, Scientific Applications* (Allen, F.H., Bergerhoff, G., & Sievers, R., Eds.), pp. 107–132. Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Beers, Y. (1957). *Introduction to the Theory of Error.* Addison-Wesley, Reading, Massachusetts.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535–542.

Bowie, J.U., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*, 164–170.

Chambers, J.M., Cleveland, W.S., Kleiner, B., & Tukey, P.A. (1983). *Graphical Methods for Data Analysis.* Wadsworth International Group, Duxbury Press, Boston.

Chothia, C. & Finkelstein, A.V. (1990). The classification and origins of protein folding patterns. *Annu. Rev. Biochem. 59*, 1007–1039.

Farber, G.K. & Petsko, G.A. (1990). The evolution of α/β barrel enzymes. *Trends Biochem. Sci. 15*, 228–234.

Horowitz, E. & Sahni, S. (1978). *Fundamentals of Computer Algorithms.* Computer Science Press, New York.

Kretsinger, R.H. & Nockolds, C.E. (1973). Carp muscle calcium binding protein II. Structure determination and general description. *J. Biol. Chem. 248*, 3313–3326.

Lebioda, L., Stec, B., & Brewer, J.M. (1989). The structure of yeast enolase at 2.25-Å resolution. *J. Biol. Chem. 264*, 3685–3693.

Lesk, A.M. (1991). *Protein Architecture [Practical Approach Series].* IRL Press, New York.

Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature 261*, 552–558.

Liljas, A. & Rossmann, M.G. (1974). Recognition of structural domains in globular proteins. *J. Mol. Biol. 85*, 177–181.

Maggiora, G.M. & Johnson, M.A. (1990). Introduction of similarity in chemistry. In *Concepts and Applications of Molecular Similarity* (Maggiora, G.M. & Johnson, M.A., Eds.), pp. 1–13. Wiley, New York.

Matthews, F.S. (1985). The structure, function, and evolution of cytochromes. *Prog. Biophys. Mol. Biol. 45*, 1–56.

Mondragon, A., Subbiah, S., Almo, S.C., Drottar, M., & Harrison, S.C. (1989). Structure of the amino-terminal domain of phage 434 repressor at 2.0 Å resolution. *J. Mol. Biol. 205*, 189–200.

Ohlendorf, D.H., Anderson, W.F., & Matthews, B.W. (1983). Many gene-regulatory proteins appear to have a similar α-helical fold that binds DNA and evolved from a common precursor. *J. Mol. Evol. 19*, 109–114.

Rackovsky, S. (1990). Quantitative organization of the known protein X-ray structures. I. Methods and short-length scale results. *Proteins Struct. Funct. Genet. 7*, 378–402.

Remington, S.J. & Matthews, B.W. (1978). A general method to assess similarity of protein structures, with applications to T4 bacteriophage lysozyme. *Proc. Natl. Acad. Sci. USA 75*, 2180–2184.

Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Adv. Protein Chem. 34*, 167–339.

Richardson, J.S. & Richardson, D.C. (1988). Helix lap-joints as ion-binding sites: DNA-binding motifs and Ca-binding "EF hands" are related by charge and sequence reversal. *Proteins Struct. Funct. Genet. 4*, 229–239.

Richardson, J.S. & Richardson, D.C. (1989). Principles and patterns of protein conformation. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., Ed.), pp. 1–98. Plenum, New York.

Rossmann, M.G. & Argos, P. (1976). Exploring structural homology of proteins. *J. Mol. Biol. 105*, 75–95.

Sali, A. & Blundell, T.L. (1990). Definition of general topological equivalence in protein structures. *J. Mol. Biol. 212*, 403–428.

Satyshur, K.A., Rao, S.T., Pyzalska, D., Drendel, W., Greaser, M., & Sundaralingam, M. (1988). Refined structure of chicken skeletal muscle troponin C in the two-calcium start at 2-Å resolution. *J. Biol. Chem. 263*, 1628–1647.

Schevitz, R.W., Otwinowski, Z., Joachimiak, A., Lawson, C.L., & Sigler, P.B. (1985). The three-dimensional structure of *trp* repressor. *Nature 317*, 782–786.

Taylor, W.R. & Orengo, C.A. (1989). Protein structure alignment. *J. Mol. Biol. 208*, 1–22.

Tufty, R.M. & Kretsinger, R.H. (1975). Troponin and parvalbumin calcium binding regions predicted in myosin light chain and T4 lysozyme. *Science 187*, 167–169.

Vyas, N.K., Vyas, M.N., & Quiocho, F.A. (1987). A novel calcium binding site in the galactose-binding protein of bacterial transport and chemotaxis. *Nature 327*, 635–638.

Williams, R.L., Greene, S.M., & McPherson, A. (1987). The crystal structure of ribonuclease B at 2.5 Å resolution. *J. Biol. Chem. 262*, 16020–16031.