PROTEINS: Structure, Function, and Genetics 9:56–68 (1991)

Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment

Chris Sander and Reinhard Schneider
European Molecular Biology Laboratory, D-6900 Heidelberg, Federal Republic of Germany

ABSTRACT **The database of known protein three-dimensional structures can be significantly increased by the use of sequence homology, based on the following observations. (1) The database of known sequences, currently at more than 12,000 proteins, is two orders of magnitude larger than the database of known structures. (2) The currently most powerful method of predicting protein structures is model building by homology. (3) Structural homology can be inferred from the level of sequence similarity. (4) The threshold of sequence similarity sufficient for structural homology depends strongly on the length of the alignment. Here, we first quantify the relation between sequence similarity, structure similarity, and alignment length by an exhaustive survey of alignments between proteins of known structure and report a homology threshold curve as a function of alignment length. We then produce a database of homology-derived secondary structure of proteins (HSSP) by aligning to each protein of known structure all sequences deemed homologous on the basis of the threshold curve. For each known protein structure, the derived database contains the aligned sequences, secondary structure, sequence variability, and sequence profile. Tertiary structures of the aligned sequences are implied, but not modeled explicitly. The database effectively increases the number of known protein structures by a factor of five to more than 1800. The results may be useful in assessing the structural significance of matches in sequence database searches, in deriving preferences and patterns for structure prediction, in elucidating the structural role of conserved residues, and in modeling three-dimensional detail by homology.**

Key words: **secondary structure, tertiary structure, residue conservation, sequence variability, sequence profile, folding units**

INTRODUCTION

Database limitations to structure prediction

Given a newly sequenced gene, there are two main approaches to the prediction of structure and function from the amino acid sequence. Homology methods are the most powerful and are based on the detection of significant extended sequence similarity to a protein of known structure or of a sequence pattern characteristic of a protein family. Statistical methods are less successful but more general and are based on the derivation of structural preference values for single residues, pairs of residues, short oligopeptides, or short sequence patterns. Both kinds of methods are severely limited by the size of the database in which one performs searches or from which one derives structural preferences. For example, in order to have on the average 5 occurrences of all possible 8000 amino acid triplets in each of the 3 secondary structure states (helix, extended strand, loop) one would need a database of about 120,000 residues in different known structures; 20 times more for all quadruplets and so on.

Current size of databases

In comparison with these requirements, the present size of the database of known protein three-dimensional (3-D) structures is too small; it is also small compared to the database of known primary sequences and very small compared to the many thousands of different proteins estimated to exist in living cells. At the end of 1989 the Protein Data Bank[1] had about 100 different protein 3-D structures with about 20,000 residues; the current database of known sequences[2,3] was about 100 times larger: 12,000 proteins with more than 3 million residues; and the total number of distinctly different proteins in nature was estimated in the hundreds of thousands.

Link between structure and sequence databases

Fortunately, many proteins in the database of known sequences are similar in sequence to a protein of known structure and this fact can be exploited to close the gap in the size of the two data-

Received January 29, 1990; revision accepted June 8, 1990.
Address reprint requests to Chris Sander, European Molecular Biology Laboratory, Postfach 10-2209, Meyerhofstrasse, 1, D-6900 Heidelberg, Federal Republic of Germany.

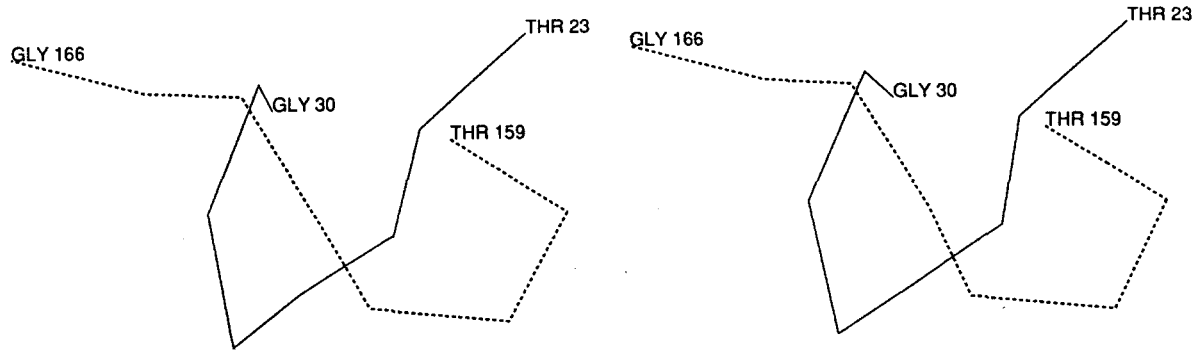

Fig. 1. The structural meaning of sequence similarity depends strongly on the length of alignments. In this extreme example, two short peptides have sequence similarity normally sufficient for structural homology (75% identical residues), yet their structures are very different. Residues 159–166 of a subtilisin protease (dashed line, data set 2SBT[7]) and residues 23–30 of an immunoglobulin (solid line, data set 3FAB[8]) have 6 out of 8 identical residues (TGSSSTVG/TGSSSNIG), but differ by 4.7 Å rms deviation in C(α) positions. Secondary structures are also very different (LTTSLLLL/ELLTTSST where T, H-bonded turn; S, geometrical turn; E, part of beta strand; L, extended loop). Protein fragments as stereo C(α) traces.

bases. For example, there are currently more than 100 known sequences of proteins homologous to the known structure of the GTPase domain or G-domain of elongation factor TU[4] and *ras* p21 oncogene protein.[5]

Key technical problem

The transfer of structure information to a potentially homologous protein is straightforward when the sequence similarity is high and extended in length, but the assessment of the structural significance of sequence similarity can be difficult when sequence similarity is weak or restricted to a short region. Note two extreme examples. (1) Extended weak sequence similarity yet very similar structures: *ras* p21 protein[5] and elongation factor TU,[4] after optimal superposition, are identical in the topology of the chain fold and similar in overall structure (with only 2.4 Å rms deviation in C(α) positions of 138 out of 166 residues), yet the two proteins are dissimilar in sequence with less than 20% identical residues. (2) Short strong sequence similarity yet very different structure[6]: octapeptides from subtilisin (2SBT)[7] and an immunoglobulin (3FAB)[8] are dissimilar in structure with as much 4.7 Å rms C(α) deviation, yet 75% identical in sequences (Fig. 1). These examples illustrate one of the two key problems (the other problem being refined measures of sequence similarity): *the shorter the length of the alignment, the higher the level of similarity required for structural significance.*

Homology threshold as a function of length

To solve this problem, we need to calibrate the length dependence of structural significance of sequence similarity. Empirically, this can be done by deriving from the database of known structures a quantitative description of the relationship between sequence similarity, structural similarity and alignment length. The resulting definition of a length-dependent homology threshold can provide the basis for reliably deducing the likely structure of globular proteins down to the size of domains and fragments. Previously, Chothia and Lesk[9] have quantified the relation between the similarity in sequence and three-dimensional structure for the cores of entire globular proteins.

Extending the database of known structures

Having solved the problem of length dependence, we can begin to merge the information in the sequence database with that in the structure database by exhaustive searches for sequence-similar fragments. Selection of sequence alignments with significant sequence similarity (homology) to proteins of known structure then leads to a database of homology-derived protein structures several times larger than the Protein Data Bank.

METHODS

Sequence and Structure Databases

The process of producing the database of homology-derived structures is effectively a partial merger of the database of known three-dimensional structures, here the PDB Protein Data Bank (fall 1989 release[1]) with the database of known protein sequences, here the EMBL/Swissprot database (release Nov. 12, 1989, 12,305 sequences).[3] The merger is partial in that only structurally homologous information is merged, where homology is based on currently available alignment methods with an empirically determined homology threshold.

Calibration of Structural Significance of Sequence Alignments

Alignment method

We perform an empirical determination of homology thresholds by studying thousands of sequence

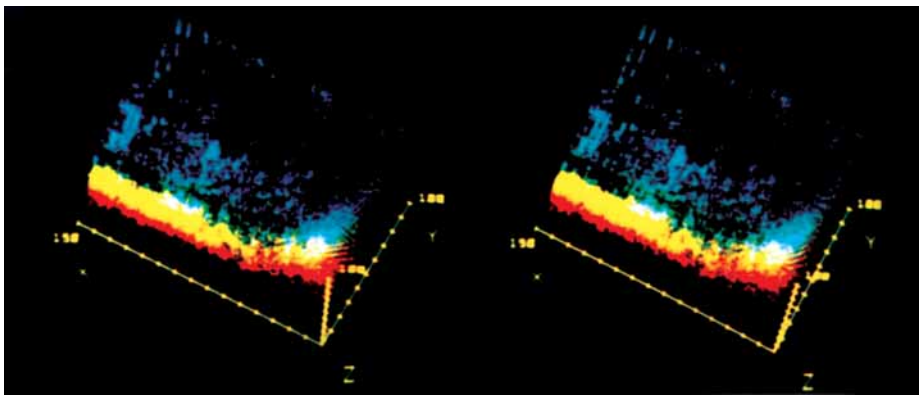

Fig. 2. Color stereo view. Calibration of the homology threshold is based on this 3-D scatter plot of sequence similarity (Y, range 0–100%), structure similarity (Z, range 0–100%) and alignment length (X, range 0–150 residues) for pairwise protein sequence alignments. Each point represents the alignment of two protein fragments, one each from a protein of known 3-D structure produced by a standard sequence alignment method. Red points are pairs dissimilar in structure (bad pairs), blue points are pairs similar in structure (good pairs), with other colors interpolating intermediate values of structural agreement. The rectangular blue slice (back) represents good pairs; they occur at almost all values of sequence similarity and all lengths—the thin population in the top half of the blue slice being an artifact of the database (very few protein pairs with 50–90% sequence similarity solved by crystallography). The absence of (yellow and red) points in the top left and front shows that no pairs with sufficiently high sequence similarity have low structure similarity: above the homology threshold, all points are "pushed" into the blue region of good structural agreement. Sequence-identical oligopeptides (5–10 residues long) with dissimilar local structure[6] are red points at the front top right. Homologous protein pairs of length about 150 residues are blue points at the back top left. Sequence similarity of an alignment is defined as the percentage of identical amino acids in an alignment, range 0–100%. Structure similarity is defined as the percentage of identical secondary structure symbols in DSSP notation, range 0–100%. Alignment lengths (number of residues) are as produced by the recursive (dynamic programming) sequence alignment algorithm[10] at a given value of smin (adjusts zero level of sequence similarity), with gaps allowed up to a total length of 10 residues (MaxDel = 10), a gap opening penalty of 3.0 units, and a gap elongation penalty of 0.1 units per residue. Similarity units are defined in the text.

alignments within the PDB database. Each protein from a selected set of high and low resolution protein structures is compared with all others from the set. We use the dynamic sequence alignment algorithm of Smith and Waterman[10] as implemented in MaxHom (C. Sander and R. Schneider, unpublished); local sequence similarity is given by the 20 by 20 matrix of amino acid similarities of McLachlan,[11] scaled to a minimum value of *smin* (usually negative) and a maximum value of 1.0 (for the top single residue identity). Maximum length of a deletion is *maxdel* (e.g., 10) residues, opening an insertion gap costs *gap opening penalty* (e.g., 3.0 similarity units) and gap elongation per residue costs *gap elongation penalty* (e.g., 0.1 units). Alignments terminate if the cumulative similarity value becomes negative. For each protein pair comparison, we use several values of *smin* in order to obtain different lengths of alignments and keep the optimal and several suboptimal alignments from each pair comparison.

Similarity measures

For each alignment, sequence similarity, structural similarity, and alignment length are noted. Sequence similarity as percent identity of amino acids; structural similarity of two aligned protein fragments either as the rms difference of equivalent C(α) positions in 3-D space after optimal superposition (tertiary structure similarity)[12] or as the percent identity of secondary structure symbols according to DSSP (secondary structure similarity)[13]; and alignment length as the number of amino acids, excluding gaps. Note that while it is important to use a more sophisticated measure of sequence similarity in producing the alignments, the simpler measure of percent identity is useful when comparing widely different alignment methods.

Homology threshold

By analyzing the distribution of points in a resulting three-dimensional scatter plot of thousands of alignments of protein fragments of known structure in terms of sequence similarity, structural similarity and alignment length (Fig. 2), we determine a safe threshold for each alignment length such that any alignment with a similarity value above this threshold represents structural homology.

Search for Homologies Between Structure and Sequence Databases

Given a safe structural homology threshold, we proceed to produce the database of homology derived protein structures. For each protein of known structure in PDB we perform a search in the sequence database for structurally significant alignments. Each sequence alignment is the result of a pairwise comparison. The end result is a multiple sequence alignment. For technical reasons, the search is performed in several steps.

1. Rapid scan of the database using FASTA[14] with sufficiently low similarity score cutoff yields a list containing all proteins potentially homologous to the reference PDB protein.

2. A more refined comparison with the proteins in this list, using MaxHom and retaining the five best distinctly different alignments for each protein pair comparison, yields an improved list of candidate alignments. MaxHom parameters used are *smin* = −0.7, *maxdel* = 10, *gap opening penalty* = 3.0, *gap elongation penalty* = 0.1. This choice of parameters is based on experience, but is not unique. Good alignments with rather long gaps may appear as two separate alignments.

3. Only alignments with similarity scores above the significance threshold (determined below, Table I) are retained.

4. All alignments are reported in register relative to a single instance of the sequence of the PDB reference protein: special notation is used to indicate insertions and deletions.

Measures of Sequence Variation

Given a multiple sequence alignment, sequence variation is measured in two ways: (1) based on the Dayhoff exchange matrix[15] (variability) and (2) based on entropy (variation entropy).

Sequence variability

This definition makes use of the Dayhoff exchange matrix and quantifies the extent to which exchanges at one position are more or less conservative. Since no sequence can be singled out as a root or master sequence, all $Npairs = N(N-1)/2$ pairs of sequences are considered. Conservation $cons(i)$ at position i is defined as the weighted sums of residue similarities over all sequence pairs ($k \neq l$)

$$cons(i) = \frac{\sum_{k,l}^{Npairs} w_{kl} sim(R_{ik}R_{il})}{\sum_{k,l}^{Npairs} w_{kl}}$$

and variability $var(i)$ as its complement relative to the maximum value $sim(max)$ of residue similarity (usually $sim(max) = 1.0$)

$$var(i) = sim(max) - cons(i)$$

where $sim(R_{ik},R_{il})$ is the similarity of residue R_{ik} in sequence k and residue R_{il} in sequence l, both at sequence position i.

Weights for sequence pairs. In defining variability, each sequence pair is weighted with its mutual distance in sequence space, defined here as the fraction of amino acid mismatches over the alignment length L.

$$w_{kl} = 1 - \frac{1}{L}\sum_{i}^{L} \delta(R_{ik},R_{il})$$

TABLE I. Homology Threshold for Different Alignment Lengths*

Alignment length L (number of residues)	Homology threshold t (% residue identity)
<10	–
10	79.6
12	71.9
14	65.9
16	61.2
18	57.2
20	53.9
22	51.1
24	48.7
26	46.6
28	44.7
30	43.0
35	39.4
40	36.6
45	34.2
50	32.3
55	30.6
60	29.1
65	27.8
70	26.7
80	24.8
>80	24.8

*A sequence alignment between two proteins is considered to imply structure homology if the sequence similarity is equal to or above the homology threshold t in a sequence region of a given length L. For example, an alignment with 30% sequence similarity over a length of 60 residues implies homology while one with 30% sequence similarity over a length of 40 residues does not. The threshold values $t(L)$ are derived from an analysis of thousands of aligned fragment pairs from the Protein Data Bank[1] and can be represented by the formula

$$t(L) = 290.15L^{-0.562}$$

where L is in the range 10–80 residues. For alignments shorter than 10 residues any value of sequence similarity appears to be consistent with any degree of structure similarity. Alignments longer than 80 residues have the asymptotic threshold of about 25% identical residues. The precise numerical values depend on the measure of sequence similarity used. Here, for simplicity, we use percent identical residues.

The weights are a way of correcting for the uneven representation of amino acid sequences in the current database. The more similar the sequences k and l are, the lower the influence of the pair kl on the family average. Very dissimilar pairs have a large weight. The underlying model of sequence variation in evolution assumes that the number of mutation events connecting two sequences is proportional to the distance between them as measured by the number of accepted point mutations and ignoring back mutations. One drawback of this particular form of pair weights is that a cluster of very similar sequences may effectively become the master sequence dominating all pair comparisons with members external to the cluster, although intracluster comparisons are appropriately weighted down.

Weights for single sequences. For some averaging purposes, e.g., deriving weighted structural preference parameters, one needs weights attached to single sequences, not sequence pairs. In such cases, we propose giving each sequence a weight related to the local density in sequence space. The more close neighbors a sequence has, the larger the local density, and hence the lower the weight of this sequence. Using any distance measure d_{kl} in sequence space ($d_{kl} = d_{lk}$ and $d_{kk} = 0.0$) we define the weight w_k for sequence k as the average distance to all other sequences l

$$w_k = \frac{1}{N} \sum_{l}^{N} d_{kl}$$

where d_{kl} is, e.g., the number of mismatches between sequences k and l, as above.[26] If the weights w_k themselves are used as weights in taking the average, we have an equation for a self-consistent set of weights. Because all elements d_{kl} of the distance matrix are positive, Perron's theorem[16] guarantees that the equivalent eigenvalue problem

$$\lambda\, w_k = \sum_{l}^{N} w_l d_{kl}$$

has a solution in terms of a unique eigenvector associated with the largest eigenvalue, which is real, and that iterative application of the right-hand side to a starting estimate for the w_k will converge to this solution. The components of this eigenvector constitute a set of unique sequence weights when normalized to unity:

$$\sum_{k}^{N} w_k = 1.0$$

A more detailed discussion of sequence weights will be presented elsewhere.

Variation entropy

The second way of defining sequence variation is based on the concept of entropy or information. Given the frequency of occurrence f_{R_i} of amino acid of type R at position i in the alignment, the entropy

$$S(i) = -\sum_{R}^{20} f_{R_i} \ln f_{R_i}$$

expresses the extent to which the distribution f_{R_i} is uniform. The sum is over all 20 possible amino acid residue types R. If all amino acids are equally frequent at position i, then

$$S(i) = -\sum_{R}^{20} 1/20 \ln 1/20 = \ln 20$$

so the range of values is $0 \leq$ entropy(i) $\leq \ln 20$. Smaller entropy(i) values represent strong conservation, larger values mean large variability. A relative measure, normalized to 1.0, is

$$relent(i) = S(i)/\ln 20$$

Both variability $var(i)$ and relative variation entropy $relent(i)$ are reported in the current version (v0.9) of the HSSP database.

RESULTS

3-D Scatter Plot of Sequence Similarity, Structure Similarity, and Alignment Length

Comparison of database structures

The systematic survey of a large number of alignments within the PDB database of known protein structures for the first time reveals the average relationship between sequence similarity (SeqSim) and structure similarity (StrSim) at various alignment lengths (AliLen) (Fig. 2). By observing the regularities in this plot we are able to quantify the notions of *"two protein structures are homologous"* and of *"two protein sequences are sufficiently similar to be considered structurally homologous."*

Variations in structure

One of the remarkable features of the 3-D scatter plot (Fig. 2) is the saturation behavior of StrSim with increasing SeqSim at a given AliLen: the wide scatter of StrSim at low SeqSim gradually narrows to a band of asymptotically constant width; and, the asymptotic width is approximately the same for various alignments lengths. In other words, for large SeqSim, secondary structure similarity is well within 30 percentage points of perfect (100%) and tertiary structure similarity within 2.5 Å of perfect (0.0 Å) (see Methods for definitions). However, perfect sequence similarity does not always imply perfect structural agreement: a protein crystal structure may vary, typically in loop regions or in domain orientation, as a result of different crystal packing, different substrate/cofactor interaction, or complex formation.

Definition of structural homology

In view of the inherent plasticity of globular protein structure reflected in the observed asymptotic width, it is reasonable to think of the structure of two aligned segments as *structurally essentially identical* or structurally *homologous* whenever the observed structures differ by not more than 30 percentage points in secondary structure (identity of DSSP 'summary' symbols) or not more than 2.5 Å in tertiary structure [rms deviation of C(α) positions].

Homology Threshold as a Function of Alignment Length

Definition of homology threshold

With a clear definition of structural homology we are now in a position to define a cutoff in sequence

similarity above which homology of structure can be inferred. For each alignment length, the cutoff is determined by inspection of the StrSim/SeqSim scatter plot (Fig. 2) or histograms (Fig. 3) as that value (arrows in Fig. 3) of sequence similarity above which almost all alignments are structurally homologous, i.e., fall within the plasticity margin of perfect structural identity.

The resulting homology cutoff (Fig. 4, Table I) is a strongly varying function of alignment length up to a length of about 70–80 residues. For example, for alignment length 30, sequence similarity has to be at least 43% (gaps allowed with a gap opening penalty of three residue identities) to infer structural homology. For very long alignment lengths 25% sequence identity is sufficient. Note that below these values of sequence similarity structural homology cannot be asserted nor excluded—the region of weaker sequence similarity is a "don't know" region (mixture of squares and crosses in Fig. 4).

Sharpness of homology threshold

There is a residual margin of error in applying the threshold to infer structural homology (Fig. 3). This is because the transitions in the scatter plot (Fig. 2) are not infinitely sharp and because the present database is a limited subset of all possible protein structures. In the absence of a correct physical theory of sequence–structure relation, inferences based on empirical relationships are subject to at least a small margin of error. The margin of error is larger for shorter alignments, for which statistical noise appears to be stronger. Visual inspection of HSSP files confirms the suspicion that a few (1–2% of total) short alignments of dubious structural significance lie above the chosen threshold. Raising the threshold by 3 percentage points relative to the values in Table I eliminates most of these, but decreases the sensitivity of the procedure.

Threshold as a tool for error detection

Violation of the threshold in a few cases actually indicated problems with PDB datasets. For example, there are a number of fragment pairs from datasets 4ATC/7ATC or 2ATC/7ATC (aspartate transcarbamoylase) with 100% sequence similarity but very different structure—up to 4.96 Å C(α) rms for 146 residues. The differences turned out to be due to a correction in the chain tracing of the regulatory chain between 2ATC/4ATC (unliganded form[17]) and 7ATC (CTP liganded form[18]). The apparently incorrect data set 4ATC was never corrected in the Protein Data Bank nor is it flagged there as incorrect.

In a similar but as yet unresolved case, fragment pairs involving the multiheme cytochromes 1CY3[19] and 2CDV[20] show surprisingly large and potentially very interesting structural differences, in spite of above-threshold sequence similarity (37% identity over 59 residues). As data set 1CY3 is known to be a preliminary structure [authors' remarks in PDB data set and Holm and Sander, J. Mol. Biol., submitted], these differences may be due to possible inaccuracies in the 1CY3 dataset.

Database of Homology-Derived Structure (HSSP)

Content of database

Notation and layout are described in Fig. 6. More than 300 files were produced, one for each PDB protein from the fall 1989 release of PDB with release 12 of EMBL/Swissprot (12305 sequences). This corresponds to derived structures for 3512 proteins or protein fragments; 1854 of these are homologous over a length of at least 80 residues. Some of these proteins are very similar to their PDB cousin, differing by as little as one residue out of several hundred. Others are more distant relatives and their homology-derived structure represents a nontrivial addition to the corpus of known structures. Some PDB proteins have several hundred known homologous sequences, others have none; e.g., human hemoglobin (4HHB)[21] has 466 aligned sequences, with from 25 to 100% identical residues, of which almost all are globins (except the last 16 which are unexpected—interesting or false—positives). Crambin (1CRN)[22] has 12 aligned sequences with 38–53% sequence identity (one unexpected positive). Rhodanese (1RHD)[23] had no homologous partner (yet).

Size of database

The increase in total information content in HSSP over PDB is as difficult to quantify as the increase in information when a homologous protein is solved by crystallography. A rough conservative estimate can be made as follows. The average number of aligned sequences is 103 per PDB entry. Of the 3512 aligned sequences (counting each protein exactly once) 1831 are more than 50% different (sequence identity) from any PDB cousin; after filtering out short fragments and potential unexpected positives by requiring an alignment length of at least 80 residues, 775 of these remain. As some investigators have chosen a cutoff of 50%[13] sequence identity for a nonredundant PDB database, one may say that the HSSP database has increased the number of nonredundant (relative to PDB) datasets by about 700–800 proteins. Allowing for close homologies within this set of additional proteins (factor of up to 0.5), the increase in nonredundant information corresponds to a factor of about 3 to 6 (387/120 to 775/120).

It is important to note that the derived structures are three-dimensional, although only secondary structure information is given in HSSP files. Based on the alignments, a rough three-dimensional model of each of the aligned proteins can be produced with relative ease.

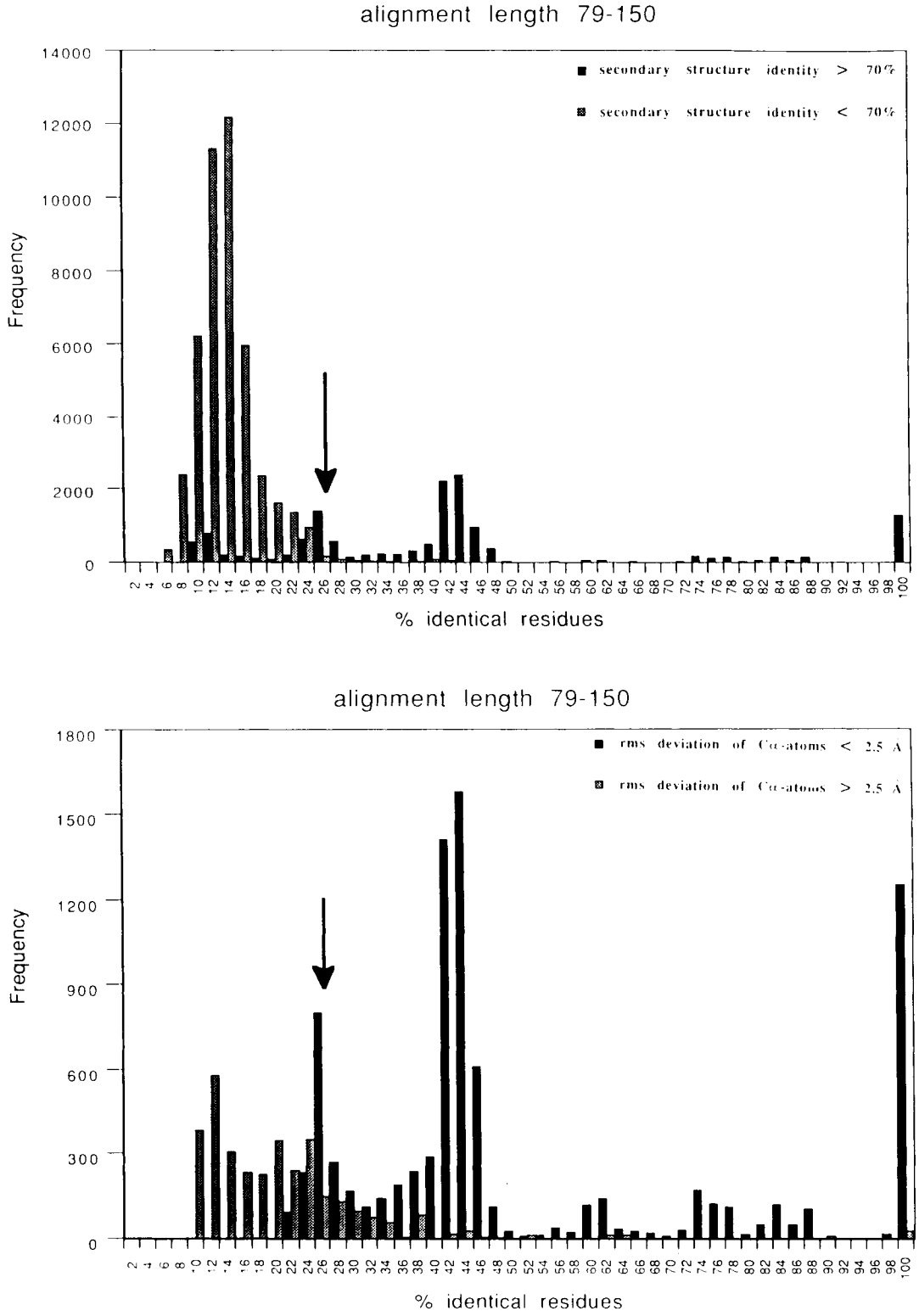

Fig. 3. Detailed justification for the particular values of the homology threshold (arrow) is provided by histogram projections of the data in Figure 2: frequency of structurally similar/dissimilar alignments as a function of the percentage of identical residues in the alignment, for alignments of length 79–150 residues. The threshold is perfect, if all fragment pairs to the right of the threshold arrow are similar in structure (black bars), without intrusion by structurally dissimilar pairs (gray bars). The strong mixture of black and gray bars to the left of the arrows indicates that below the threshold one cannot use percent sequence identity as indicator of structure similarity. The particular choice of threshold represents an attempt to divide the range of sequence identity values into a "do not know" region (left) and a 'sequence similarity implies structure similarity' region (right). **(Top)** Structure similarity assessed by identity of secondary structure, with the similar (black)/dissimilar (gray) dividing line at 70% identity of secondary structure symbols (H,E,T etc.). **(Bottom)** Structure similarity assessed by rms deviation of C(α) atom positions after optimal superposition of the two fragments, with the similar (black)/dissimilar (gray) dividing line defined to be at 2.5 Å rms C(α) deviation per residue. Apparently, comparing **(Top)** and **(Bottom),** the dividing line is less clear cut in terms of C(α) deviation than it is in terms of identity of secondary structure. Therefore identity of secondary structure was used for calibration of the homology threshold. In part, possibly inaccurate structures like 2ACT and 1CY3 contribute to the excess of gray bars to the right of the arrow in **(Bottom).**

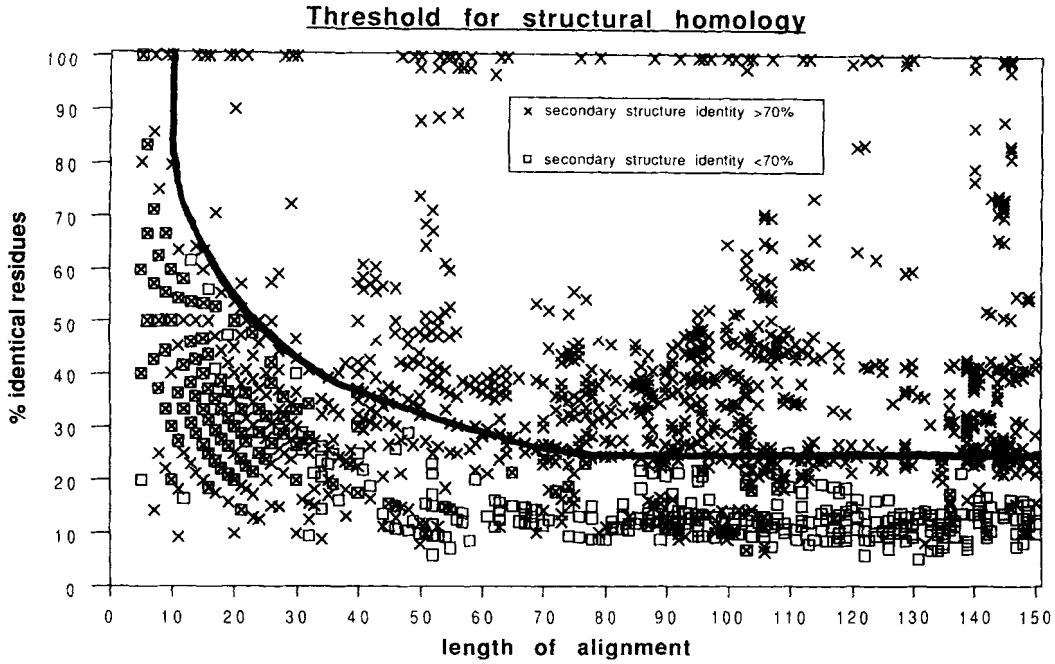

Fig. 4. Homology threshold for structurally reliable alignments as a function of alignment length, a principal result (numerical values in Table I). Each data point represents an alignment between two fragments from proteins of known structure. The graph is a two dimensional projection of Figure 2 onto the plane of sequence similarity/alignment length, with structural similarity collapsed to a one bit yes/no description (crosses/squares). The data points are a subset of the data in Figure 2. The homology threshold (curved line) divides the graph into a region of *safe structural homology* (upper right) where essentially all fragment pairs are observed to have good structural similarity (crosses, secondary structure identity above 70%) and a region of *homology unknown or unlikely* (lower left) where some fragment pairs are structurally similar (crosses) and some are not (squares, secondary structure identity below 70%). The histogram of Figure 3a corresponds to a vertical slice of this graph in the length range 79–150 residues, summing all available data points in that length range.

Sequence Variation

Role of conserved residues

Sequence elements conserved in evolution are taken as evidence of selective pressure resisting mutational events. Much can be learned from studying sequence conservation in a three-dimensional protein structure, especially about the possible contribution of individual residues to the architecture of the protein fold and to protein function.[24] An example is shown in Figure 5.

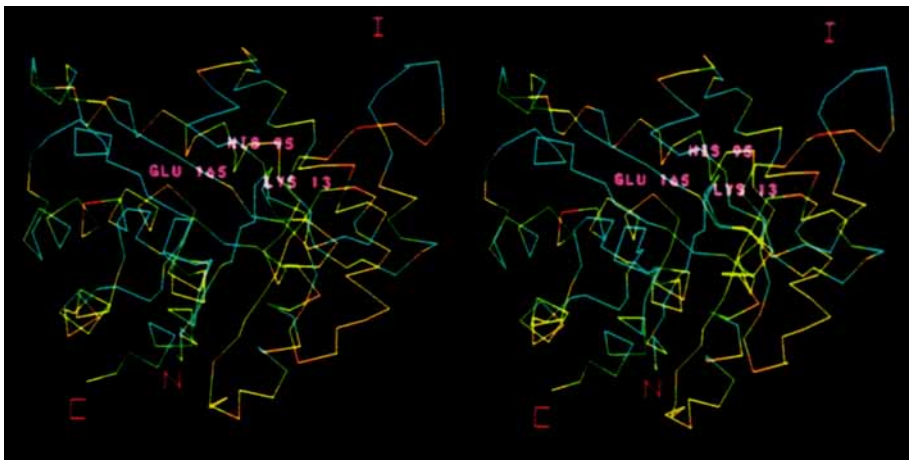

Fig. 5. Sequence variability mapped into the tertiary structure of chicken triose phosphate isomerase, 1TIM.[25] The three-dimensional color stereo view of the TIM monomer shows placement of most conserved (blue) and most variable (red) residues in the crystal structure. Residues are colored according to variability (12 homologous sequences, see Methods) on a sliding scale: blue-green-yellow-red. The most conserved residues are near the active site (Glu-165, His-96, Lys-13) at one end of the parallel β barrel (front) and in a loop (I, top right front) that makes important contacts in the dimer interface. N is N-terminus, C is C-terminus.

Computer graphics display

For example, residues conserved in the 3-D structure of the enzyme triose phosphate isomerase[25] (color Fig. 5) appear to be located in (1) the ends of β strands of the parallel barrel, (2) near the active site, and (3) in a loop at the dimer interface. Coloring residues in the 3-D structure according to sequence conservation—which can be done routinely and quickly from the PROTID.HSSP output file—provides a visual way of developing intuition about the importance of residues in protein function and folding. Even a linear graph of sequence variability against the protein sequence can be instructive.

DISCUSSION

Current Limitations

The main results, the homology threshold curve, the reported alignments and the implied secondary

Fig. 6. Description of HSSP files: One HSSP file contains a structural protein family: one test protein of known structure and all its structurally homologous (as judged by our homology threshold Table I, Fig. 3) relatives from the database of known sequences. The file is divided into four blocks, HEADERS, PROTEINS, ALIGNMENTS and SEQUENCE PROFILE. The HEADERS block is mandatory. The other three blocks are present only if at least one homologous alignment is found; each of the additional blocks begins with the string "##". File organization is line-oriented. Lines have a maximum length of 132 bytes. Some of the line types are self-explanatory.

(a) HEADERS block: the first four bytes in the file, 'HSSP', can be used for file type detection. The first line also has the version number of the HSSP software. The PDBID (protein data bank identifier) line identifies the test protein of known structure (e.g. 1PPT), the SEQBASE-line specifies the source of the aligned sequences (e.g. EMBL/Swissprot or PIR/NBRF). The PARAMETER line specifies alignment parameters used in MaxHom (smin = lowest similarity, smax = highest similarity between amino acids, maxdel = maximum length of deletion; maxdel restriction will be removed in future releases). The THRESHOLD line refers to the homology threshold curve used. Information about the test protein as copied from PDB (name, source, author) and as derived (length of the sequence SEQLENGTH, number of distinct chains NCHAIN, and the number of aligned sequences NALIGN).

(b) PROTEINS block: pair alignment data for each of the proteins deemed structurally homologous to the test protein, where the word pair alignment refers to the alignment of the test protein with the single homologous protein

ID	EMBL/SWISSPROT identifier of the aligned (homologous) protein
STRID	if the 3-D structure of this protein is known, then STRID (structure ID) is the Protein Data Bank identifier as taken from the database reference line or DR-line (latest date) of the EMBL/SWISSPROT entry
%IDE	percentage of residue identity of the alignment.
IFIR/ILAS	first and last residue position of the alignment in the test protein
JFIR/JLAS	first and last residue position of the alignment in the aligned protein.
LALI	length of the alignment excluding insertions and deletions.
NGAP	number of insertions and deletions in the alignment.
LGAP	total length of all insertions and deletions
LSEQ2	length of the entire sequence of the aligned protein
PROTEIN	one-line description of aligned protein.

(c) ALIGNMENTS block: residue-by-residue details of the family alignment. From left to right in one line: sequence and structure information for one position in the test protein taken from the corresponding DSSP file [13]; sequence variability for this position followed by the aligned sequences in the same order as in the PROTEINS-block; equivalent (aligned) residue in each of the homologous database proteins. The sequences of the test protein and the aligned database proteins run vertically.

SeqNo	sequential residue number of test protein as in DSSP file.
PDBNo	residue number/name as in PDB file.
AA	amino acid type in one letter code
STRUCTURE	secondary structure summary, hydrogen bonding patterns for turns and helices, geometrical bend, chirality, one character name of β-ladder and of β-sheet
BP1, BP2	β-bridge partners.
ACC	solvated residue surface area in $Å^2$ (number of contacting water molecules *10)
VAR	sequence variability (see text) as derived from the NALIGN alignments
....:....1	ruler to identify alignments by their number in the PROTEINS block.

NOTE that lower case characters in the sequence of the test protein (AA-column) indicate cysteines in SS-bridges. Insertions and deletions in either sequence are indicated by special characters in the sequence of the aligned protein:

dots (...)	indicate a deletion in the aligned sequence
lower case characters	bracket an insertion point in the aligned sequence, e.g. AkeV means AK[insertion]EV

There are residues from up to 70 database proteins in one line. If the number of alignments (NALIGN) is greater than 70, the alignments block is repeated (1..70, 71–140 etc) until the total number of alignments is reached.

(d) SEQUENCE PROFILE block: relative frequency for each of the 20 amino acid residue in a given sequence position, from counting the residue at that position in each of the aligned sequences including the test sequence. A value of 100 means that at this position only one type of amino acid is found. Asx and Glx are counted in their acid/amide form in proportion to their database frequencies (Asx to Asp: 0.521, Asx to Asn: 0.439, Glx to Glu: 0.623, Glx to Gln: 0.410 as in EMBL/Swissprot release 12, November 1989). For each line, corresponding to a particular sequence position:

NOCC	number of aligned sequences spanning this position (including the test sequence).
NDEL	number of sequences with a deletion in the test protein at this position
NINS	number of sequences with an insertion in the test protein at this position
ENTROPY	entropy measure of sequence variability at this position (see Methods)
RELENT	relative entropy, i.e. entropy normalized to the range 0–100.

Headers **a**

```
HSSP        HOMOLOGY DERIVED SECONDARY STRUCTURE OF PROTEINS , VERSION 0.9 1989
PDBID       1PPT
DATE        file generated on 18-Mar-90
SEQBASE     RELEASE 13.0 OF EMBL/SWISS-PROT WITH  13837 SEQUENCES
PARAMETER   SMIN: -0.7  SMAX: 1.0  MAXDEL: 10
THRESHOLD   according to t(L)=290.15 * L ** -0.562
REFERENCE   Sander C., Schneider R. : Database of homology-derived protein structures. Proteins,1990.
CONTACT     e-mail Schneider@EMBL.bitnet or Sander@EMBL.bitnet / phone +49-6221-387361 / fax +49-6221-387306
AVAILABLE   Free academic use. Commercial users must apply for license.
HEADER      PANCREATIC HORMONE
COMPND      AVIAN PANCREATIC POLYPEPTIDE
SOURCE      TURKEY (MELEAGRIS GALLOPAVO) PANCREAS
AUTHOR      T.L.BLUNDELL,J.E.PITTS,I.J.TICKLE,S.P.WOOD
SEQLENGTH     36
NCHAIN         1
NALIGN        23
```

Proteins **b**

PROTEINS : EMBL/SWISSPROT identifier and alignment statistics

NR.	ID	STRID	%IDE	IFIR	ILAS	JFIR	JLAS	LALI	NGAP	LGAP	LSEQ2	PROTEIN
1 :	PAHO$CHICK	1PPT	1.00	1	36	1	36	36	0	0	36	PANCREATIC HORMONE.
2 :	PAHO$STRCA		0.94	1	36	1	36	36	0	0	36	PANCREATIC HORMONE.
3 :	PAHO$ALLMI		0.80	2	36	2	36	35	0	0	36	PANCREATIC HORMONE.
4 :	PAHO$ANSAN		0.78	1	36	1	36	36	0	0	36	PANCREATIC HORMONE.
5 :	NEUY$PIG		0.57	2	36	2	36	35	0	0	36	NEUROPEPTIDE Y (NPY).

Alignments **c**

ALIGNMENTS 1 - 23

SeqNo	PDBNo	AA	STRUCTURE	BP1	BP2	ACC	VAR	....:....1....:....2....:
1	1	G		0	0	101	0	GG G
2	2	P	-	0	0	60	0	PPPPPPPPPPPPPPPPP PPPP
3	3	S	-	0	0	106	50	SALSSSSSAAIPPLLLLL LLPL
4	4	Q	-	0	0	139	33	QQQQKKKKKKKKKEEEEEEEEQE
5	5	P	-	0	0	26	0	PPPPPPPPPPPPPPPPPPPPPP
6	6	T	-	0	0	126	46	TTKTDDDDEEEEEVVEVVEVMEM
7	7	Y	-	0	0	121	43	YYYYNNNNAAANNYYYYYYYYSY
8	8	P	-	0	0	55	0	PPPPPPPPPPPPPPPPPPPPPP
9	9	G	> -	0	0	27	0	GGGGGGGGGGGGGGGGGGGGGG
10	10	D	T 3 S+	0	0	128	16	DDDNEEEEEEEEEDDDDDDDDGD

Sequence profile **d**

SEQUENCE PROFILE AND ENTROPY

SeqNo	PDBNo	V	L	I	M	F	W	Y	G	A	P	S	T	C	H	R	K	Q	E	N	D	NOCC	NDEL	NINS	ENTROPY	RELENT
1	1	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0.000	0
2	2	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	23	0	0	0.000	0
3	3	0	39	4	0	0	0	0	0	13	13	30	0	0	0	0	0	0	0	0	0	23	0	0	1.397	47
4	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	38	25	38	0	0	24	0	0	1.082	36
5	5	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	24	0	0	0.000	0
6	6	21	0	0	8	0	0	0	0	0	0	0	17	0	0	0	4	0	33	0	17	24	0	0	1.630	54
7	7	0	0	0	0	0	0	58	0	13	0	4	0	0	0	0	0	0	0	25	0	24	0	0	1.053	35
8	8	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	24	0	0	0.000	0
9	9	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0.000	0
10	10	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	38	4	54	24	0	0	0.965	32

//

and tertiary structures are subject to a number of limitations.

Validity of homology threshold

The principal limitation in the calibration of the homology threshold as reported is in the measure of sequence similarity used. Given two aligned protein sequences, we have used the simplest possible measure, the percent identity of amino acids, which is reported by most available alignment procedures. A more refined *local* measure (actually used here in producing the alignments) uses a mutational 20 by 20 frequency or similarity table.[11] In addition, a more refined *global* measure is a weighted sum over local similarity, in which more conserved positions are given a higher weight, as in multiple alignment methods.[26] The advantage of using the simplest measure is its immediate usefulness for other workers. The disadvantage is that the threshold transition is rougher and the reported alignments include more possible false positives than presumably would result with a threshold in terms of a more refined similarity measure. Plans for a future version include use of a more refined and weighted similarity measure (higher weights for conserved regions) both for threshold definition and for alignment production.

Accuracy of reported alignments

Considerable effort is being expended to improve the accuracy of sequence alignments relative to structural alignments.[27,28] In general, alignments may be inaccurate in local detail (trailing ends incorrectly aligned, incorrectly shifted gaps etc.). An example is given by loop 41–49 (DLKVAGGAS) in subtilisin[7] aligned to 47–57 (DLAGKVVGGWD) in thermitase,[29,30] where the alignment procedure incorrectly shifts a gap by two residues, i.e., KVAGG does not match KVVGG in 3-D. In such cases, the sequence alignment may correctly represent conservation in the evolutionary chain of events connecting the two sequences while the structural alignment reflects a local structural rearrangement as a result of mutations in sequence positions spatially near the conserved residues. We therefore see no obvious remedy for locally incorrect alignments in loop regions.

We plan to improve the accuracy of the alignments by going from the current independent pairwise method to a growing cluster alignment method. In a growing cluster alignment, each new sequence is brought into the cluster by alignment against the sequence profile of the existing cluster; in addition, each position i can be weighted with a conservation weight $c(i)$ derived from the existing cluster. Accuracy may be further increased by the use of newly derived exchange matrices, e.g. exchange matrices that depend on the structural state of the residue position at which the exchange takes place (work in progress).

Accuracy of homology-derived structure

Each alignment implies a homology-derived 3-D structure for a sequence of unknown 3-D structure. Even if the inference of homology is correct (true positives), the expected accuracy of the derived structures is not 100%. For example, trypsin/elastase (3PTN/3EST),[31,32] known to be homologous, have a sequence alignment with 35% sequence identity and 6 gaps for 240 residue positions. However, their secondary structures symbols (DSSP states) are only 80% identical and the aligned C(α) positions differ by 1.4 Å for 180 residues after optimal superposition. An extreme example is provided by the 15-residue long sequence similarity in a loop region of bovine and porcine phospholipase (BP2/P2P).[33,34] In spite of safe overall homology of 88% sequence identity over 122 residues, a 15-residue stretch with 80% sequence similarity (KLDSCKVLVDNPYTN/NLDSCKFLVDNPYTE; res 57–71), above our homology threshold for this length, has significantly different structure in the two crystals: 17% identity in secondary structure and 3.3 Å C(α) rms in tertiary structure. In response to mutations, loop regions are simply more plastic than secondary structure segments or core regions.

In general, the accuracy of derived structure is limited to the plasticity margin (Fig. 2), i.e., homology-derived structure can be expected to be occasionally wrong in local detail, e.g., in the conformation of some loop regions and in the precise delineation of the ends of some secondary structure segments. The average accuracy given by the middle of the plasticity margin is 85% identity of secondary structure (states H,G,E,B,T,S,blank) and 1–2 Å rms deviation of C(α) positions in tertiary structure. Cases of incorrectly inferred existence of α-helices and β-strands, however, are very rare—we are not aware of a single example in the database of deposited 3-D structures.

In addition, the inference of homology for alignments above the homology threshold may simply be incorrect (false positives). In the current implementation and with the current threshold values (Table I), we estimate, by visual inspection, the level of possible false positive alignments at roughly 1–2%, most of them short alignments which are subject to more statistical noise.

Limited database

Any empirical investigation is limited by the size of the database. Deviations from the principles observed here are possible as more and perhaps new classes of protein structures became known.

Recommendation for Use of HSSP Database

The current version (v0.9; January 1990) contains 1–2% unexpected positives. Which of these are false positives and which represent discovery of structural relationship is in principle unknown—the threshold was deliberately set to include a small fraction of these. The user of the database should allow for a small margin of error and, to be on the safe side, choose to delete a certain fraction of the lowest-scoring alignment upon reading in the HSSP files, using her/his own judgment and criteria, at the risk of deleting correct positives as well. A program to filter the files according to a higher, user-defined threshold is available. In the current version (v0.9) of the files being distributed via network, a filter corresponding to raising the threshold values by 3 percentage point for all lengths has been applied.

In using variability scores, the user should be aware that low occupancy positions (few alignments span that position) have poorly determined variability values—in the limit of zero occupancy the variability is undefined and set to zero. The user may choose to use only positions with occupancy larger than, say, five proteins.

Availability of database

The HSSP database (one file per PDB protein with a full coordinate set named PDBID.HSSP) will be distributed freely to end users (unlimited academic use; resale excluded; added value information copyrighted). Files will be mounted on the file server NETSERV@EMBL.bitnet (send e-mail message 'help' for more information). Source code to read and filter HSSP files is also available and serves as an operational definition of file format.

Future Extensions and Applications

The threshold for structural homology can be used to improve the evaluation of matches in sequence database searches. Although the length dependence of statistical significance of sequence matches is well known mathematically,[35] the most popular database alignment search programs (e.g., FASTA, Wordsearch[14,36]) sort the best hits on total similarity, without reference to length. We suggest that a homology threshold curve like the one presented here can be used to order the database matches by the extent to which their score exceeds the threshold, in appropriate units.

Unexpected positives in HSSP files, e.g., sequences aligned to a globin structure with sequence similarity above the threshold curve which are so far not known to be related to globins, are tantalizing candidates for possible discovery of structural homologies.

The significantly increased database with its structural family alignments, sequence profiles, sequence variability (or variation entropy) can be used

1. to study the evolution of protein sequence and structure; an example is the correlation between residue side chain contacts and sequence variation in the TNC family[24];
2. to derive statistically more reliable preference parameters or sequence patterns for structure prediction[37–39];
3. to extract weighted sequence profiles for database searches. For example, sequence positions along the profile can be given a weight corresponding to the degree of conservation, such that strongly varying positions are effectively ignored in a profile sequence comparison[40–43];
4. to define the core region of a structural family for model building by homology, even when only one structure is known. Strongly varying positions are considered not be part of the invariant core;
5. to derive structure-dependent similarity tables for amino acid types (or tuples of types) for use in aligning sequences to proteins of known 3-D structure and for use in planning point mutations based on known or predicted protein structures.

Updates of the database are planned for each new release of the protein structure and sequence databases.

ACKNOWLEDGMENTS

We thank all crystallography and NMR groups who made their three-dimensional structures available; Peter Rice for the installation of databases; Roy Omond for providing the network file server; Gerrit Vriend for special extensions to his program WHATIF; and Andrew McLachlan, Peter Sibbald, Martin Vingron, and Andreas Dress for interesting discussions regarding measures of sequence variation.

REFERENCES

1. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
2. Protein Identification Resource, National Biomedical Research Foundation. Georgetown University Medical Center, 3900 Reservoir Road, N.W. Washington D.C.
3. SWISS-PROT Protein Sequence Database. EMBL Data Library, D-6900 Heidelberg, FRG and Amos Bairoch, Departement de Biochimie Medicale, Centre Medical Universitaire, 1211 Geneva 4, Switzerland.
4. LaCour, T.F.M., Nyborg, J., Thirup, S., Clark, B.F.C. Structural details of the binding of guanosine diphosphate to elongation factor TU from E. coli as studies by x-ray crystallography. EMBO J. 4:2385–2388, 1985.
5. Pai, E.F., Kabsch, W., Krengel, U., Holmes, K.C., John J., Wittinghofer, A. Structure of the guanine-nucleotide-binding domain of the Ha-ras oncogene product p21 in the triphosphate conformation. Nature (London) 341:209–214, 1989.
6. Kabsch, W., Sander, C. On the use of sequence homologies to predict protein structure. Identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci. 81:1075–1078, 1984.
7. Drenth, J., Hol, W.G.J., Jansonius, J.N., Koekoek, R. A comparison of the three-dimensional structures of subtili-

sin and subtilisin novo. Cold Spring Harbor Symp. Quant. Biol. 36:107, 1972.
8. Saul, F.A., Amzel, L.M., Poljak, R.J. Preliminary refinement and structural analysis of the fab fragment from human immunoglobulin new at 2.0 Ångstroms resolution. J. Biol. Chem. 253:585–597, 1978.
9. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5: 823–826, 1986.
10. Smith, T.F., Waterman, M.S. Identification of common molecular subsequences. J. Mol. Biol. 147:195–197, 1981.
11. McLachlan, A.D. Tests for comparing related amino acid sequences. J. Mol. Biol. 61:409–424, 1971.
12. Kabsch, W. Discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr. A34:827–828, 1978.
13. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
14. Pearson, W.R., Lipman, D.J. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85:2444–2448, 1988.
15. Dayhoff, M.O. (ed.) "Atlas of Protein Sequence and Structure," Vol 5, Suppl. 3. Washington DC: National Biomedical Research Foundation, 1978.
16. Horn, R.A., Johnson, C.R. "Matrix Analysis," Cambridge: Cambridge University Press, 1985: 497.
17. Ke, H.M., Honzatko, R.B., Lipscomb, W.N. Structure of unligated aspartate carbamoyltransferase of Escherichia coli at 2.6 Å resolution. Proc. Natl. Acad. Sci. U.S.A. 81: 4037–4040, 1984.
18. Kim, K.H., Pan, Z.X., Honzatko, R.B., Ke, H.M., Lipscomb, W.N. Structural asymmetry in the CTP-liganded form of aspartate carbamoyltransferase from Escherichia coli. J. Mol. Biol. 196:853–875, 1987.
19. Pierrot, M., Haser, R., Frey, M., Payan, F., Astier, J.P. Crystal structure and electron transfer properties of cytochrome C3. J. Biol. Chem. 257:14341–14348, 1982.
20. Higuchi, Y., Kusunoki, M., Matsuura, Y., Yasuoka, N., Kakudo, M. Refined structure of cytochrome C_3 at 1.8 Ångstroms resolution. J. Mol. Biol. 172:109–139, 1984.
21. Fermi, G., Perutz, M.F., Shaanan, B., Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 Ångstroms resolution. J. Mol. Biol. 175:159–174, 1984.
22. Teeter, M.M. Water structure of a hydrophobic protein at atomic resolution. Pentagon rings of water molecules in crystals of crambin. Proc. Natl. Acad. Sci. U.S.A. 81:6014–6018, 1984.
23. Ploegman, J.H., Drent, G., Kalk, K.H., Hol, W.G.J. Structure of bovine liver rhodanese. I. structure determination at 2.5 Ångstrom resolution and a comparison of the conformation and sequence of its two domains. J. Mol. Biol. 123:557–594, 1978.
24. Godzik, A., Sander, C. Conservation of residue interactions in a family of Ca-binding proteins. Protein Eng. 2: 589–596, 1989.
25. Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Wilson, I.A. Atomic coordinates for triose phosphate isomerase from chicken muscle. Biochem. Biophys. Res. Commun. 72:146–155, 1976.
26. Vingron, M., Argos, P. A fast and sensitive multiple sequence alignment algorithm. CABIOS 5:115–121, 1989.
27. Argos, P. A sensitive procedure to compare amino acid sequences. J. Mol. Biol. 193:385–396, 1987.
28. Risler, J.L., Delorme, M.O., Delacroix, H., Henaat, A. Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J. Mol. Biol. 204:1019–1029, 1988.
29. Teplyakov, A.V., Kuranova, I.P., Harutyunyan, E.H., Frömmel, C., Hohne, W.E. Crystal structure of thermitase from Thermoactinomyces vulgaris at 2.2 Å resolution. FEBS-Letter 244:208–212, 1989.
30. Frömmel, C., Sander, C. Thermitase, a thermostable subtilisin: comparison of predicted and experimental structures and the molecular cause of thermostability. Proteins 5:22–37, 1989.
31. Walter, J., Steigemann, W., Singh, T.P., Bartunik, H., Bode, W., Huber, R. On the disordered activation domain in trypsinogen, chemical labelling and low-temperature crystallography. Acta Crystallogr. b 38:1462, 1982.
32. Meyer, E., Cole, G., Radahakrishnan, R., Epp, O. Structure of native porcine pancreatic elastase at 1.65 Ångstroms resolution. Acta Crystallogr. b 44:26, 1988.
33. Dijkstra, B.W., Kalk, K.H., Drenth, J., De Haas, G.H., Egmond, M.R., Slotboom, A.J. Role of the N-terminus in the interaction of pancreatic phospholipase A_2 with aggregated substrates. Properties and crystal structure of transaminated phospholipase A_2. Biochemistry 23:2759–2766, 1984.
34. Dijkstra, B.W., Renetseder, R., Kalk, K.H., Hol, W.G.J., Drenth, J. Structure of porcine pancreatic phospholipase A_2 at 2.6 Ångstroms resolution and comparison with bovine phospholipase A_2. J. Mol. Biol. 168:163–179, 1983.
35. Smith, T.F., Waterman, M.S., Burks, C. The statistical distribution of nucleic acid similarities. Nucleic Acids Res. 13:645–656, 1985.
36. Devereux, J., Haeberli, P., Smithies, O. A comprehensive set of sequence analysis programs for the Vax. Nucleic Acids Res. 12:387–395, 1984.
37. Maxfield, F.R., Scheraga, H.A. Improvement in the prediction of protein backbone topography by reduction of statistical errors. Biochemistry 18:697–704, 1979.
38. Rooman, M., Wodak, S.J. Identification of predictive sequence motifs limited by protein structure data base size. Nature (London) 335:45–49, 1988.
39. Gibrat, J.-F., Garnier, J., Robson, B. Further developments of protein secondary structure prediction using information theory. J. Mol. Biol. 198:425–443, 1987.
40. Gribskov, M., McLachlan, M., Eisenberg, D. Profile analysis: Detection of distantly related proteins. Proc. Natl. Acad. Sci. U.S.A. 84:4355–4358, 1987.
41. Bashford, D., Chothia, C., Lesk, A.M. Determinants of a protein fold. Unique features of the globin amino acid sequences. J. Mol. Biol. 196:199–216, 1987.
42. Smith, R.F., Smith, T. Automatic generation of primary sequence patterns from sets of related protein sequences. Proc. Natl. Acad. Sci. USA 87:118–122, 1990.
43. Staden, R. Methods to define and locate patterns of motifs in sequences. CABIOS 4:53–60, 1988.