J. Mol. Biol. (1995) **251**, 390–399

JMB

A Sequence Property Approach to Searching Protein Databases

Uwe Hobohm and Chris Sander*

EMBL-European Molecular Biology Laboratory, D-69012 Heidelberg, Germany

Currently available sequence alignment programs are generally not capable of detecting functional and structural homologs in the twilight zone of sequence similarity, i.e. when the sequence identity falls below about 25%. Here we attempt to detect such weak similarities using an approach based on a notion of protein sequence similarity radically different from that used in sequential alignment. The approach defines protein sequence dissimilarity (or distance) as a weighted sum of differences of compositional properties such as singlet and doublet amino acid composition, molecular weight, isoelectric point (protein property search or PropSearch).

With PropSearch, either single sequences can be used for a database query, or multiple sequences can be merged into an ''average'' sequence reflecting the average composition of a protein family.

First, we show that members of structural protein families have a low mutual PropSearch distance when the weights are optimized to discriminate maximally between structural families. Second, we demonstrate the results of database searches using the PropSearch method. Such searches are very rapid when scanning a preprocessed database and do not require alignments.

In cases in which conventional alignment tools fail to detect similarities, PropSearch can be used to generate hypotheses about possible structural or functional relationships between a new sequence and sequences in the database.

**Corresponding author*

Keywords: amino acid composition; database search; structural homologs

Introduction

After sequencing a novel gene or protein of unknown function, one of the first steps is from the laboratory bench to the computer to perform database searches for sequence similarities, in the hope of obtaining an indication, by homology arguments, about the structure or function of the gene product. Widely used protein database search tools such as those based on the dynamic programming algorithms (Smith & Waterman, 1981) or the Blast and Fasta methods (Altschul *et al*., 1990; Pearson, 1991) optimally align two protein sequences keeping strict sequential order, where optimal alignment is defined in terms of maximizing a sequence similarity measure or minimizing sequence distance. In a typical standard database search, the probability of obtaining at least partial information on the function of a newly sequenced gene product is about 50 to 60%, i.e. in about 40% of all cases the researcher is left without functional or structural information (Bork *et al*., 1992). However, in cases where very different amino acid sequences can fold to form very similar structures, detection of a common fold poses severe problems. Sequence alignment methods are not able to discover structural relationships in cases of convergent or far divergent evolution, when the sequence identity is below about 20 to 25%. For example, in the large globin family, sequence similarity between the most distant relatives is not detectable using standard alignment techniques, as mutational drift has erased almost all memory of the original ''sentence'' of amino acid ''letters''. In cases of convergent evolution or in cases of shuffled gene pieces, proteins may share the same overall structure but have a different connectivity of secondary structure elements. In such cases sequential alignment is useless. For such distant relationships the notion of distance in sequence space implied by standard alignment techniques is clearly inadequate.

The concept of distance (or similarity) in sequence space can be generalized in a number of ways. One extreme is to disregard the order of amino acid

Abbreviations used: 3D, three-dimensional; PRH, putative remote homolog; REP-sequence, sequence representing the same overall 3D-fold.

0022–2836/95/330390–10 $12.00/0

residues completely and compare the amino acid composition. Attempts in this direction have addressed the problems of (a) prediction of secondary structural class from sequence, (b) prediction of membership in a protein family from sequence, and (c) automated clustering of the protein sequence database.

(a) Nishikawa *et al*. analyzed the amino acid composition of 135 proteins with known structure and predicted one of five secondary structural classes: alpha, beta, alpha + beta, alpha / beta and ''irregular'', with 70% accuracy (Nakashima *et al*., 1986). Klein & DeLisi (1986) predicted four classes: high alpha, high beta, mixed and low content ordered structure, using four properties and 80% success. Muskal & Kim (1992) predicted the content of alpha and beta structure for 15 proteins with an error rate of about 5%.

(b) Several authors have developed methods to predict membership in particular protein families from amino acid composition. In an early attempt, Nishikawa *et al*. (1983) predicted one of four classes: intracellular enzyme, intracellular non-enzyme, extracellular enzyme and extracellular non-enzyme with 66% success. Klein *et al*. (1984) used four sequence properties (average hydrophobicity, net charge, sequence length and variation in hydrophobic residues along the chain) to predict one of six protein families with 76% success. Dubchak *et al*. (1993) predicted one of five classes: four-alpha-helical bundle, parallel (alpha / beta)-8 barrel, nucleotide binding fold, immunoglobulin fold or none of these, with 87% accuracy, using a neural network system and weighted amino acid composition as properties. Bishop & Thompson (1984) searched a pre-processed EMBL nucleic acid database for homologous sub-sequences of length 7.

(c) Clustering of the protein sequence database based on general sequence properties has been attempted by at least three groups. Blaisdell (1986) clustered 64 DNA-sequences by their expressions of second or third-order Markov chains, i.e. by the distributions of pairs and triples of nucleotides. van Heel (1991) compared sequences using matrices representing the content of dipeptides. An eigenvector projection of the corresponding 400 dimensional sequence space led to a clustering of 10,000 sequences into 600 families, some of which have a reasonable biological interpretation (van Heel, 1991). Wu *et al*. (1992) used a neural network system to classify proteins into 620 families with 90% accuracy.

These approaches represent interesting attempts to use general sequence properties for the classification of a protein based on its sequence. What they have in common is that the cumulative knowledge about protein classes derived from the databases is represented in a set of rules or numerical values that can then be applied to predict properties of a new sequence. However, the usefulness of these methods is limited by the scope of the predefined classes (a, b) and / or by the quality of the precalculated database clustering (c).

Here we describe a sequence comparison method (PropSearch) that uses a notion of similarity in sequence space intermediate between the two extreme approaches of ''strict sequential order'' and ''amino acid content irrespective of order''. Each protein is represented by a set of general sequence properties that include the amino acid content as well as partial information on sequence order in the form of low order *n*-tuplets. Similarity between proteins is then expressed in terms of these general properties. In addition, the PropSearch approach aims to exploit maximally the information present in sequence and structure databases at any given time by recasting the prediction or classification problem as that of finding all significant database neighbors of a newly sequenced protein. In practice this means that for a given protein sequence PropSearch scans the sequence databases for proteins significantly similar to the search protein in terms of general properties such as amino acid residue content, average hydrophobicity and charge, content of some classes of amino acid doublets and the like. Significant hits are those that provide credible information about the structure and / or function of the search protein.

Results and Discussion

The PropSearch approach goes beyond previous attempts to describe sequences by composition, which primarily had led to protein classification schemes (Nishikawa *et al*., 1983; Klein *et al*., 1984; Dubchak *et al*., 1993; van Heel, 1991; Wu *et al*., 1992). Our aim was to develop a database search tool based on general sequence properties, including amino acid composition. In the vast majority of possible protein pairs, similar structure means similar function (though there exist some interesting exceptions from this general rule: Holmes *et al*., 1993; Holm & Sander, 1994). So we asked whether there are conserved properties in similar structures, that could be exploited for database searches. Some of those properties are obvious. For instance, in protein families with S-S bridges the content of cysteine is highly conserved (data not shown). Even more obvious is sequence length, which is highly conserved among members of one protein family. The conservation of other properties is not so obvious. One might, for instance, speculate that the content of charged residues, the content of hydrophobic residues or the average hydrophobicity is conserved in a particular protein family.

Weight optimization

A given set of 144 weights was considered to have a high fitness if it led to a separation in sequence space between 58 protein families (for the selection of protein families see the legend to Table 2), i.e. if the search for a particular query sequence resulted in small distances for members of the same protein family and larger distances for members of different families. Fitness was defined as average rank of members of the same family, averaged over all

Table 1. The 144 properties to characterize a protein sequence as used in PropSearch

1–20	20 amino acid residues (CFILMVWYHNPQSTDEKRAG)
21	Charged residues (ILVHDEKR)
22	Positively charged residues (HKR)
23	Negatively charged residues (DE)
24	Polar residues (WYHNQSTDEKR)
25	Aliphatic residues (ILV)
26	Aromatic residues (FWYH)
27	Tiny residues (MW < 80 D)
28	Bulky residues (MW > 120 D)
29	Sequence length
30	Calculated isoelectric point
31	Average hydrophobicity
32	Molecular weight
33–144	Doublet composition

Values 1 to 20 and 33 to 144 were expressed in percent sequence length, value 29 and 32 as its logarithm to keep all values in about the same numerical range.

To keep the number of possible doublets low, 20 amino acid types were collected into four groups: hydrophobic (CFILMVWY), charged (DEKR), tiny (AG), other (HNPQST).

The grouping of amino acid residues, which was based on the intuition of one of the authors, led to 16 possible doublet groups. We considered doublets gapped by 0 to 6 residues, resulting in 112 (16×7) values for doublet composition per sequence. The content of a particular doublet was counted by sliding a window with a stepsize of 1 over the sequence. For window sizes greater than 2 (i.e. gap length greater than 0) the leftmost and rightmost amino acid residue was used to determine the doublet.

Example: in the sequence ''ACDE'' we find 6 out of 112 possible doublets:

Residue pair	PropSearch doublet group	gaplength	group	Content (percent)
AC	tiny	0	hydrophobic	33
CD	hydrophobic	0	charged	33
DE	charged	0	charged	33
A-D	tiny	1	charged	50
C-E	hydrophobic	1	charged	50
A–E	tiny	2	charged	100

families. The average rank at generation 0, when all 144 weights were initialized to 1.0, was about 135 (Figure 1). The optimization was stopped at generation 50 (average rank of about 17) when no further improvement could be achieved. To avoid artefacts from overtraining, database searches were performed using the weights at generation 30 (average rank of about 25).

Weights after generation 30 are in Table 3. As expected, sequence length and molecular weight are

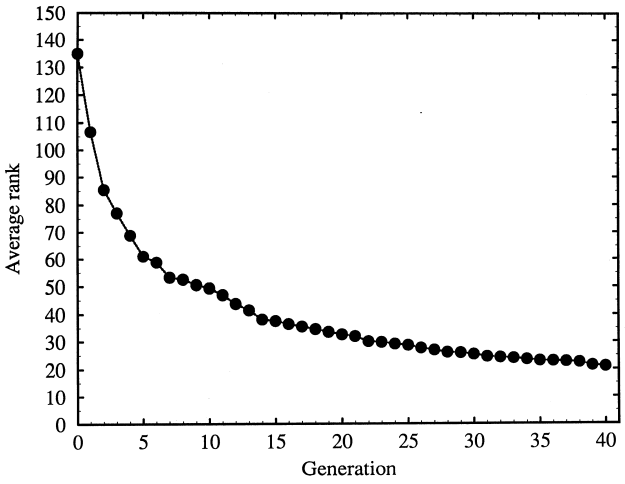

Figure 1. Optimization of property weights using a genetic algorithm.

highly conserved properties within protein families, resulting in high PropSearch weights of 89 and 59, respectively. The highest PropSearch weights for the content of particular amino acid residues are 62 (C), 22 (W), 17 (H), 16 (G), 16 (L), 14 (Y), 12 (Q), 10 (P) for cysteine, tryptophan, histidine, glycine, leucine, tyrosine, glutamine and proline, respectively. The structural importance of these residue types can be rationalized to some extent. Cysteine puts severe constraints on the 3D structure by its ability to form S-S bridges. The aromatic rings of both tryptophan and tyrosine play an important role in maintaining the hydrophobic core structure and, in addition, they contain a polar N and OH, respectively, for the formation of specific interactions. Tryptophan is the most bulky amino acid and sometimes involved in the binding of ligands at protein surfaces and in active sites (e.g. Noble *et al*., 1992). Histidine is often involved in catalytic sites. Glycine has a particular role at the ends of helices (Richardson & Richardson, 1988). Leucine has a high helix forming propensity (Robson, 1976). Proline is particularly important for the formation of tight loops (Sibanda & Thornton, 1991). Regarding the composite properties, the content of aromatic residues, the content of positively charged residues, the content of bulky residues, the average hydrophobicity and the content of polar

Table 2. Collection of protein families representing different 3D folds

1AAJ	1AAK	1AHC	1ARS	1ATX	1BAA	1BFG	1C5A	1CAM	1CBN	1CCR	1CRL
1CSH	1CUS	1ECO	1ENH	1FDX	1FXD	1FHA	1FKB	1FNR	1GKY	1GPS	1HBQ
1HMY	1HUW	1IAG	1IFC	1IPD	1LIS	1MUP	1OFV	1PDA	1PGD	1PHO	1PHP
1POA	1PPN	1PPT	1RCB	1SO1	1TYS	2ACQ	2CDV	2CMD	2HBG	2IHL	2LH7
2LIV	2MHR	2SN3	3CLA	3DFR	3PGM	4ENL	4GCR	4GBP	5P21	7RSA	

From a representative list of 296 protein structures with a pairwise sequence identity below 25% (Hobohm *et al*., 1992) we collected 59 protein families (here assigned as PDB-identifiers (Bernstein *et al*., 1977) which have only a single chain, have a length equal or more than 50 residues and at least five REP-sequences in the respective multiple sequence alignment. To prohibit even subtle sequence similarities, a more rigorous check on sequence similarity was applied. We compared the HSSP multiple sequence alignments (Sander & Schneider, 1991) for all 59 families pairwise: no family pair has a protein sequence in common in the respective multiple sequence alignment.

Ferredoxins 1FDX and 1FXD were merged into one family, since they have the same fold.

The resulting 58 families were chosen to represent protein families with a broad structural diversity, but without inter-familiar sequence similarity.

Table 3. Result of property weight optimization after generation 30

Weight	Property
89	Log sequence length
62	Cys
59	Log molecular weight
22	Trp
20	Aromatic residues
17	His
16	Gly
16	Leu
14	Tyr
12	Gln
12	Charged–0–tiny
11	Isoelectric point
11	Other–2–other
10	Positively charged residues
10	Pro
9	Bulky residues
9	Lys
8	Phe
8	Tiny–5–hydrophobic
7	Val
7	Hydrophob.–0–hydroph.
7	Hydrophob.–0–tiny
7	Other–0–tiny
7	Tiny–4–hydrophob.
6	Average hydrophobicity
6	Glu
6	Hydrophob.–2–hydrophob.
6	Hydrophob.–2–charged
6	Charged–3–other
6	Tiny–4–tiny
5	Polar residues
5	Hydrophob.–6–hydrophob.
5	Charged–5–tiny
5	Charged–6–other
5	Tiny–3–hydrophobic
5	Tiny–5–tiny

Weights were optimized using a genetic algorithm as described in Methods. Only weights higher than 4 are shown. Properties were expressed in percent sequence length, except for molecular weight and sequence length (both as their logarithm), average hydrophobicity (calculated from Eisenberg-scale, Sweet & Eisenberg, 1983), isoelectric point (calculated as described in the GCG-program package; GCG, 1991). Doublets were defined as explained in the legend to Table 1. Example: ''hydro–1–tiny'' represents the content of doublet residues consisting of a hydrophobic residue and a tiny residue gapped by one residue.

residues obtained the highest PropSearch weights with 20, 10, 9, 6, 5, respectively. The conservation of these properties in similar 3D structures is intuitively reasonable. Surprisingly, the weights for content of positively and negatively charged residues were asymmetrical, i.e. 10 and 0, respectively. The reason for this asymmetry is not easily explained.

Several doublets acquired a high weight, higher than some single amino acid residue types. Doublets with a gap length of three and four may reflect alpha-helical content and doublets with a gap length of one may reflect content of beta-sheet secondary structure. Four doublets with a gap length of zero had high weights: ''charged–tiny'', ''hydrophobic–hydrophobic'', ''hydrophobic–tiny'' and ''other–tiny'' with PropSearch weights of 12, 7, 7, 7, respectively. Three other doublets also had high weights: ''other–other'' gapped by two residues, ''tiny–hydrophobic'' gapped by five residues and ''tiny–hydrophobic'' gapped by four residues with PropSearch weights of 11, 8, 7, respectively. With one exception, in all of these high scoring doublets either tiny or hydrophobic or both types of amino acid residues are involved. Thus, the high PropSearch weights for these doublets may be caused in part by the high weights for glycine, tryptophan and tyrosine alone. Taken together, for some of the components of the property vector the distribution of optimized weights makes sense using independent intuitive arguments, but there is no obvious way of deriving the weights from first principles.

Comparison of PropSearch and Fasta

PropSearch was not designed to compete with, but to complement conventional alignment tools. Nevertheless one would like to know how PropSearch compares with a common database searching tool such as Fasta in separating family members from the rest of the SwissProt database. Using PropSearch, the average rank for 58 protein families using 1322 query sequences was about 25 at generation 30 (Figure 1). However, the spread is large. In the average case, the majority of sequences evolutionarily related to the search protein is found at the top of the PropSearch output, but some sequences rank far from the top, leading to a high average rank. Since it is computationally expensive to perform 1322 Fasta database searches for comparison, we used a subset of 10 families with 50 sequences instead. Fasta obtained an average rank of about 1.5. Fasta produces a narrow spread, collecting almost all sequences at the top of the output, which leads to a low average rank. Bearing the wishes of somebody in the laboratory in mind, who uses PropSearch when other searching tools have failed to report a significant relationship, the ability of PropSearch to find some, but not all, family members near the top of the hit list may be useful in proposing a function for the query protein.

Assessment of the capability of PropSearch to find structural homologs without sequence similarity (remote homologs)

The ability of PropSearch to detect remote homologs was investigated in greater detail using a list of protein family pairs with structural homology, but without detectable sequence similarity. Remote homologs in many cases showed up far from the top of the hit list, thus the notion of average rank is not appropriate for comparing PropSearch and Fasta. Instead, we counted how many remote homologs showed up among the first 1000 and 200 hits, respectively. Compared to Fasta, PropSearch found 2.2 and 1.6 times as many remote homologs (Table 4). Thus, the representation of an amino acid sequence as a weighted vector of sequence properties may indeed reflect some generic properties of similar protein structures.

A similar comparison was made between BLAST and PropSearch. However, BLAST cannot be

Table 4. Comparison of PropSearch, Fasta, BLAST for the capability to detect remote homologs

Number of top ranking sequences examined	Variable[a]	200	1000
PropSearch	—	222	771
Fasta	—	134	350
BLAST	0	—	—

To examine the capability to detect remotely homologous protein sequences, protein families with structural homology were selected from Orengo *et al*. (Orengo *et al*., 1993; Orengo, 1994) and Holm & Sander (1994). Only single chain PDB-entries were considered, and only families with at least three REP-sequences in their respective multiple sequence alignment (for definition of REP-sequence see Methods). Family pairs with common protein sequence identifiers in their respective multiple sequence alignment were excluded. Also excluded were family pairs with a difference in sequence length of more than 20%. These selection criteria resulted in a list of 143 pairs of structurally homologous protein families. By searching the database for the top sequence of the multiple sequence alignment of family 1 of a pair, the output was examined for the number of REP sequences belonging to family 2 of a pair. The numbers in the table are the number of remotely homologous sequences detected among the top 200 and top 1000 hits, summed over all searches.

[a] With BLAST the number of hits reported cannot be enforced as with Fasta or PropSearch, but is variable depending on the query sequence and an internal significance threshold (Altschul *et al*., 1990).

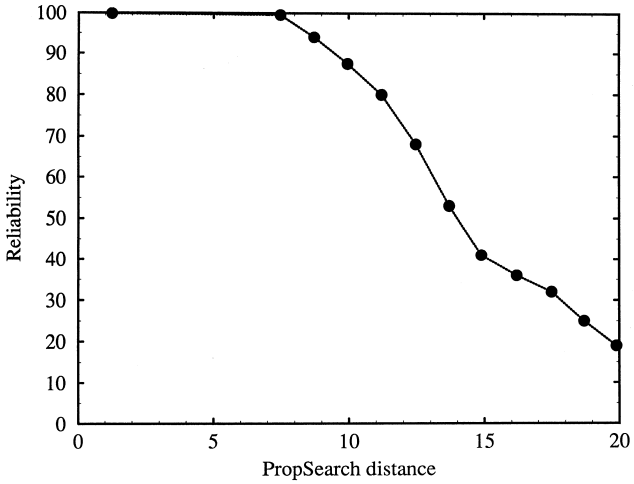

Figure 2. Relationship between PropSearch distance and identity of 3D-structure. Using the same set of 58 protein families and 1322 sequences that were used to train the PropSearch weights, we investigated the reliability of PropSearch database searches with respect to the PropSearch distance. The reliability is expressed as percent number of searches, where the first non-family member showed up. Example: in 991 of 1322 searches (75%) the first non-family member showed up with a PropSearch distance higher than 11.5.

easily forced to report hits below its significant threshold. Thus, the BLAST output examined contained variable numbers of hits depending on the particular query sequence. We did not find a single remote homolog using BLAST.

Significance estimation

For any method calculating similarity or distance between sequences a reliability estimation or significance is required. The structural significance of sequence similarity has been calibrated previously for sequence alignments on the basis of test runs using sequences of known structure (HSSP-threshold, Sander & Schneider, 1991). Briefly, when the percentage of identical residues exceeds 25% in alignments longer than 80 residues, a structural similarity can be inferred. In a similar fashion, a significance estimation for PropSearch similarity was performed here, using the same set of 1322 sequences that were used to train property weights (see Figure 2). According to this estimation, given two proteins a PropSearch distance of 8.0 or lower indicates they are in the same or a structurally similar protein family with 96% reliability; for distances under 10.0 the reliability is about 84%. This is a conservative estimation, i.e. many family members show up at larger distances. The significance threshold of Figure 2 can be applied generally when using PropSearch as an instrument to query a database.

Remote homologs with low PropSearch distance

To examine the capability of PropSearch to find sequences undetectable by alignment methods, 592 different globin sequences were extracted from the SwissProt database and used for an all-against-all alignment. 59,732 of all pairs (34%) had a sequence identity below 24%, i.e. the similarity of those pairs is insignificant and undetectable by alignment. In a second step, the PropSearch distance of such pairs of remote homologs was calculated. Twelve such pairs (0.02%) showed a PropSearch distance below 12, i.e. a reliability of about 70%; 93 pairs (0.16%) showed a PropSearch distance below 13, i.e. a reliability of about 55%, and 414 pairs (0.7%) showed a PropSearch distance below 14, i.e. a reliability of about 50%. Hence, a fraction of globin remote homologs has a low PropSearch distance.

PropSearch can also be used to search for fragments

Optimized PropSearch weights have high values for sequence length and molecular weight, an obvious consequence of choosing protein families for the weight optimization with similar length sequences. Many similarities of interest involve ''mobile'' domains that any bias to overall sequence length matching may greatly limit the power of the method. Therefore a ''fragment search'' option was introduced which can be switched on in those cases, where the sequence of interest is suspected to be part of other, larger sequences. In a fragment search, a window of the length of the query sequence is shifted over sequences in the database, and PropSearch properties are compared. However, for this purpose the unprocessed database has to be queried, and the execution time increases from a few seconds to several minutes.

Predictions

To provide genuine predictions, we applied PropSearch to actins, G-protein coupled receptors (GPCR), cyclins, heat shock proteins, hexokinases, histone-like proteins, HIV-Nef proteins, myosins and prion proteins and predicted putative remote homologs. This collection of protein families is somewhat arbitrary and meant to illustrate the usage of PropSearch.

In a first step we collected family members using multiple sequence alignments. In a second step we calculated an average property vector (PropSearch merge option) as described in Methods, which was aimed at representing the family as an average sequence. This family vector was used to query the database. The output was scanned visually, and the highest ranking sequence without obvious relationship was inspected further as a putative remote homolog (PRH). The distance in sequence space between the query family and the family of the PRH was assessed by running a Fasta database search using the PRH sequence as query sequence. PRH family members were collected applying again the HSSP significance threshold to the Fasta output. Both families were compared for common sequences. If no common sequences were found we assumed that no sequence similarity (in the alignment sense) between the families exists. Such cases represent (non-trivial) PRH predictions: low similarity in the alignment sense, yet low PropSearch distance. The results of our predictions are summarized in Table 5.

Cyclins

The 34 G2/mitosis specific cyclin-A and cyclin-B sequences were merged and used for a PropSearch query. The first sequence in the output following cyclins was TRPE_YEAST, an anthranylate synthase component I. The relationship to cyclins is an open question.

G-protein coupled receptors

G-protein coupled receptors (GPCR) contain seven membrane-spanning helices; 286 sequences of G-protein coupled receptors (GPCR) were used for a PropSearch database family search. The highest ranking protein sequence not (yet) classified as a member of the GPCR family was YNX6_CAEEL, a hypothetical 51.6 kDa protein in chromosome III of *Caenorhabditis elegans*. No function has yet been assigned to this protein, but the hydrophobicity plot of the sequence is consistent with several transmembrane regions. A structural relationship between GPCR-proteins and the hypothetical 51.6 kDa a protein seems likely.

Actins, hexokinases, heat shock proteins

Recently, a remote structural relationship between hexokinase, Hsc-70 and actin was analyzed in detail (Holmes *et al*., 1993). We merged 11 hexokinase sequences for a PropSearch query. The first non-hexokinase was LEPA_ECOLI the GTP-binding protein Lepa from *Escherichia coli*. A similar structure seems possible. Reassuringly, nine Hsc-70 type heat shock proteins were found among the first 200 hits. Searching alone for yeast hexokinase-B, we found 40 heat shock proteins and 14 actins among the first 200 hits. Thus, yeast hexokinase-B appears to be more similar to actin/hsp70 than is the average of all hexokinases.

In addition, 20 actin sequences were merged and used for a PropSearch query. The first sequence in the output following actins was RTG2_YEAST, a retrograde regulation protein required for interorganelle communication between mitochondria. Interestingly, RTG2_YEAST has a sequence identity of 31% with the tubulin beta-2 chain of *Geotrichum candidum*. Thus, a structural relationship between actins, tubulins and RTG2_YEAST appears possible.

A separate search with 40 merged sequences of heat shock proteins showed YME1_YEAST as the first non-heat shock protein. The Yme1 protein of yeast is important for maintaining the integrity of the mitochondrial compartment. It has marginal sequence identity to *E. coli* heat shock protein Hslu in the twilight zone of sequence similarity (27% identical residues over an alignment length of 147 residues). Therefore Yme1 may be a member of the actin/hsp70/hexokinase structural family.

Histone-like proteins

Histone-like proteins of bacteria share some properties with histones of higher organisms, although no sequence identity above the significance threshold can be found. One particular member of this family, the HTa protein from the archaebacterium *Thermoplasma acidophilum*, has been proposed to be an intermediate between eukaryotic histones and eubacteiral histone-like proteins (Nakashima *et al*., 1986) based on a sequence similarity of 25% over 24 residues. However, this similarity lies well below the significance threshold.

Eight histone-like protein sequences were used for a PropSearch family search. The top ranking protein not belonging to the family of histone-like proteins was the L14 protein from the 50 S-ribosomal subunit of *Mycoplasma capricolum*. A similar function, namely compacting DNA or RNA, and a similar structure seems likely. We found 46 histones and 87 ribosomal proteins among the first 200 hits.

Myosins

A total of 22 myosin heavy chain sequences of varying length were used for a PropSearch family search. The first protein without apparent alignment similarity was the human restin protein. It contains a large coiled-coil alpha-helical domain and seems to be an intermediate filament associated protein. Thus,

Table 5. PropSearch predictions

Query family	Number of sequences merged for PropSearch family search	Length	Putative remote homolog	Length	Dist.	Reliab. (%)	
Actin	20	125–381	RTG2_ YEAST	394	7.07	99	Retrograde regulation protein 2
Cyclin A,B G2/mitosis	34	257–530	TRPE_ YEAST	528	6.73	99	Anthranilate synthase comp. I
GPCR	310	297–1171	YNX6_ CAEEL	462	7.32	99	Hypothetical 51.56 kDa protein
Hsc-70 heat shock protein	40	612–681	YME1–YEAST	747	7.31	99	Yme1 protein
Hexokinase	11	465–918	LEPA_ ECOLI	598	5.96	99	GTP-binding protein LEPA
Histone-like protein	8	90–94	RL14_ MYCCA	122	14.12	45	S50 ribosomal protein
HIV-Nef	49	182–309	HA1Y_ MOUSE	298	10.69	80	H-2 class I alpha chain
Myosin heavy chain	22	532–2116	REST_ HUMAN	1427	8.39	94	Restin cytoplasmic linker protein
Prion protein	8	226–273	SER1_ DROME	265	15.43	35	Serine protease

Members of the query family were merged into one average ''sequence'', which was used to query the database. The top scoring protein not obviously belonging to the query family is indicated as putative remote homolog. The reliability of similarity in structure and/or function is a function of the PropSearch distance as shown in Figure 2.

restin may have a function and structure not unlike that of myosin.

HIV-Nef proteins

A total of 49 Nef sequences from HIV-1, HIV-2 and SIV were merged for a PropSearch family search. The most abundant protein species among the first 200 hits were 29 HLA class I alpha chains. The top HLA class-I alpha chain has a PropSearch distance of 10.69 or about 80% probability to have a similar structure, according to our reliability estimation. A hypothesis about a possible structural relationship leading to molecular mimicry has been published recently (Hobohm & Sander, 1993).

Prion proteins

A modified form of the prion protein (PrP) is believed to be the causative agent leading to spongiform encephalopathies such as scrapie, Kuru, Jacob-Creutzfeld-disease and BSE. Its normal function is unknown. The sequence is conserved among species indicating an important function. However, PrP knockout mice show no gross abnormalities (Bueler *et al.*, 1994). Eight prion protein sequences from eight species were collected for a PropSearch family query. The first non-prion protein was a *Drosophila* serine protease with a relatively high PropSearch distance of 15.43. Whether prion protein has a protease activity might be an interesting investigation.

Summary

PropSearch is a protein database searching tool which may be useful in the context of genome analysis in cases for which conventional alignment tools find no sequences significantly similar to the query protein. If a sequence family is available for the query protein, the merge option of PropSearch can be used to scan the database, at the same speed as with a single sequence. If the sequence of interest is suspected to be a fragment, the ''fragment-search''-option can be used to find fragments in larger sequences.

The sets of sequences identified as similar to the query sequence by conventional alignment tools and by PropSearch are, in general, only partially identical: while PropSearch, on the average, finds some known members of a particular protein family, it may fail to find others. On the other hand, PropSearch may find family members not detectable by alignment tools, because no similarity at the level of sequential alignment is present.

A PropSearch server on Internet can be accessed using World Wide Web client software (http://www.embl-heidelberg.de/prs.html). The selection of properties and the optimization of property weights may be subject to further improvement.

Methods

Calculation of a property vector from the amino acid sequence

In the current version of PropSearch, a protein sequence is characterized by 144 numerical values calculated from the amino acid content (1 to 20), the content of some physical properties such as average hydrophobicity, average charge and others (21 to 32), and the content of some gapped and ungapped dipeptides (33 to 144) (Table 1).

Preprocessing a database of properties

From each sequence in the SwissProt database (Bairoch & Boeckmann, 1991) a property vector of 144 numerical values is calculated (see Table 1). Each property except molecular weight, length, isoelectric point and hydrophobicity is expressed as a percentage of the number of residues in the entire sequence. Each property is normalized by dividing by its standard deviation. The standard deviation of properties is calculated over all sequences in the SwissProt database. A database of property vectors is precalculated and stored for fast database access. A typical PropSearch search against the SwissProt database currently (December 1994) takes about four seconds on a standard workstation.

PropSearch database search for a single protein sequence: calculation of property distances

The property distance D between two protein sequences is calculated as the root weighted mean square difference of the components of the property vectors (weighted Euclidian distance):

$$D = \sqrt{\Sigma(|A_i - B_i|)^2 W_i}$$

where A_i is property i of protein A after normalization by database sigma; B_i is property i of protein B after normalization by database sigma; and W_i is weight for property i.

To scan the database of property vectors, the distance between the query vector and each database vector is calculated. Distances are sorted, and high scoring proteins, i.e. those with a small distance relative to the query protein, can be inspected for potential structural and functional homology.

PropSearch database search for a family of protein sequences (merge option)

For a family of query proteins, sequences can be merged into one average property vector. For the merge option the mean for each of 144 properties is calculated, averaged over all sequences of the family. This averaged vector is then used to query the database as described above.

Filtering a multiple sequence alignment for sequences representing the same overall 3D structure (REP-sequences)

To optimize weights and to assess the ability of PropSearch to detect remote homologs, sequence families were needed in which each family member represents the

same overall 3D structure. These are available in the HSSP database of sequence families aligned to proteins of known structure (Sander & Schneider, 1991). Multiple sequence alignments, however, may contain short alignments representing homology limited to domains. To eliminate these, we filtered HSSP files using only sequences which had a sequence identity of more than 35% over an alignment length not less than 75%, and a length difference not more than 25%, relative to the first sequence entry in the multiple sequence alignment. These selection criteria result in a list of sequences with the same overall fold, which we call REP-sequences (representative sequences: sequences representing the same overall 3D structure).

Collection of protein families with different 3D structures for property weight optimization

To optimize the weights of the 144 properties used in calculating we collected a test set of 58 unrelated protein families (Table 2). Details of the selection procedure are in the legend to Table 2. We selected from the multiple sequence alignments of those 58 protein families (Table 2) all REP sequences. To obtain a larger set of sequences with low mutual sequence identity yet similar structure, we attempted to enlarge a family by iteratively performing database alignment searches using the Fasta program with the family member of lowest sequence identity (relative to the original query sequence) as the new query sequence, until no additional sequences with similarity above the threshold of structural homology (Sander & Schneider, 1991) could be detected. This iterative enlargement procedure, a kind of extension walk through sequence space, collected an additional 230 sequences, resulting in an overall collection of 1322 sequences in 58 families.

Weight optimization using a genetic algorithm

The optimization was performed using 100 weight vectors (''genes''), each vector representing the weights of 144 properties. Weights were represented by integer values using an array of length 144. All 100×144 property weights were initialized to 1.0 at generation 0.

The ''fitness'' of a gene was calculated by a four step procedure: (1) Take one sequence out of 1322 and calculate the distance between the query sequence and all other 1321 sequences; (2) order sequences on distance; (3) calculate the average rank of family members:

$$R_{\text{fam}} = \sum \frac{(R_i - (i-1))}{N}$$

where i labels the family member (numbered sequentially starting at the top of the list of hits), R_i is the rank of family member i in the list of hits, the sum is over i from 1 to N where N is the number of family members. (4) Do steps (1) to (3) for all 1322 sequences. The fitness of a set of weights is defined as the sum of 1322 R_{fam}. A low average rank or high fitness results when a query sequence collects its family members at the top of the output (low PropSearch distances), separating them from members of other families with higher PropSearch distances.

In each generation the evaluation of fitness was done for all 100 genes. The ten highest scoring genes were reproduced according to the following scheme:

Gene rank	Number of copies
1	50
2	20
3	10
4	5
5	4
6	3
7	3
8	2
9	2
10	1
	100

After gene reproduction, genes were mutated with a probability of 0.035 and recombined with a probability of 0.2, with the exception of the gene number 1, which was neither mutated nor recombined, but kept. To speed up the optimization, genes 30 to 50 were subjected to a fivefold higher mutation rate.

Acknowledgements

We thank Rüdiger Sültemeyer for his contributions in the early stages of this work in the context of his diploma thesis (Sültemeyer, 1988); Liisa Holm and David Thomas for critically reading the manuscript. U.H. was supported by a grant of the German Ministry of Research (BMFT) in the bioinformatics program under the RELIWE project.

References

Altschul, M. D., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment tool. *J. Mol. Biol.* **215**, 403–410.

Bairoch, A. & Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* **19**, 2247–2250.

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. (1977). The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Bishop, M. & Thompson, E. (1984). Fast computer search for similar DNA sequences. *Nucl. Acids Res.*, **12**, 5471–5474.

Blaisdell, E. A. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl Acad. Sci. USA*, **83**, 5155–5159.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. & Sonnhammer, E. (1992). Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. *Protein Sci.* **1**, 1677–1699.

Bueler, H., Fischer, M., Lang, Y., Bluethmann, H., Lipp, H. P., DeArmand, S. J., Prusiner, S. B., Aguet, M. & Weissmann, C. (1992). Normal development and behaviour of mice lacking the neuronal cell-surface PrP protein. *Nature*, **356**, 577–582.

Dubchak, I., Holbrook, S. R. & Kim, S.-H. (1993). Prediction of protein folding class from amino acid composition. *Proteins: Struct. Funct. Genet.* **16**, 79–91.

GCG Program Manual for the Wisconsin Package Version 8 (1991). Genetics Computer Group, 575 Science Drive, Madison, WI 53711, USA.

Hobohm, U. & Sander, C. (1993). Does the HIV Nef protein mimic the MHC? *FEBS Letters*, **333**, 211–213.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–17.

Holm, L. & Sander, C. (1994). The FSSP database of structural aligned protein fold families. *Nucl. Acids Res.* **22**, 3600–3609.

Holmes, K. C., Sander, C. & Valencia, A. (1993). A new ATP-binding fold in actin, hexokinase and Hsc70. *Trends Cell Biol.* **3**, 53–59.

Klein, P. & DeLisi, C. (1986). Prediction of protein structural class from the amino acid sequence. *Biopolymers*, **25**, 1659–1672.

Klein, P., Kanehisa, M. & DeLisi, C. (1984). Prediction of protein function from sequence properties discriminant analysis of a database. *Biochim. Biophys. Acta*, **787**, 221–226.

Muskal, S. & Kim, S.-H. (1992). Predicting protein secondary structure content—a tandem neural network approach. *J. Mol. Biol.* **225**, 713–727.

Nakashima, H., Nishikawa, K. & Ooi, T. (1986). The folding type of a protein is relevant to the amino acid composition. *J. Biochem.* **99**, 153–162.

Nishikawa, K., Kubota, Y. & Ooi, T. (1983). Classification of proteins into groups based on amino acid composition and other characters. I: angular distribution. *J. Biochem.* **94**, 981–995.

Noble, M., Pauptit, R., Musacchio, A., Wierenga, R. & Saraste, M. (1992). Crystal structure of a SRC-homology 3 (SH3) domain. *Nature*, **359**, 851.

Orengo, C. (1994). Classification of protein folds. *Curr. Opin. Struct. Biol.* **4**, 429–440.

Orengo, C. A., Flores, T. P., Taylor, W. T. & Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng.* **6**, 485–500.

Program Manual for the Wisconsin package Version 8., Genetics Computer Group, 575 Science drive, Madison, Wisconsin, USA 53711.

Pearson, W. R. (1991). Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.

Richardson, J. S. & Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. *Science*, **240**, 1648.

Robson, B. (1976). Conformational properties of amino acid residues in globular proteins. *J. Mol. Biol.* **107**, 327–56.

Sander, C. & Schneider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Sibanda, B. L. & Thornton, J. (1991). Conformation of beta-hairpins in protein structure: classification and diversity in homologous structure. *Methods Enzymol.* **202**, 59–82.

Smith, T. F. & Waterman, M. S. (1981). Comparison of biosequences. *Advan. Appl. Math.* **2**, 482–489.

Sültemeyer, R. (1988). Vergleich von Proteinen anhand von charakteristischen Sequenzkenngrößen. Diplomarbeit, Unversität Heidelberg und Fachhochschule Heilbronn.

Sweet, R. M. & Eisenberg, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.* **171**, 479–488.

van Heel, M. (1991). A new family of powerful multivariate statistical sequence analysis techniques. *J. Mol. Biol.* **220**, 877–887.

Wu, C., Whitson, George, McLarty, Jerry, Ermongkonchai, A. & Chang, T.-C. (1992). Protein classification artificial neural system. *Protein Sci.* **1**, 667–677.

Edited by F. Cohen

(Received 10 January 1995; accepted in revised form 2 June 1995)