Selection of representative protein data sets

U. HOBOHM, M. SCHARF, R. SCHNEIDER and C. SANDER

Protein Sci. 1992 1: 409-417

Protein Science (1992), *1*, 409-417. Cambridge University Press. Printed in the USA.

Selection of representative protein data sets

UWE HOBOHM, MICHAEL SCHARF, REINHARD SCHNEIDER, AND CHRIS SANDER
European Molecular Biology Laboratory, Meyerhofstrasse 1, D-6900 Heidelberg, Germany

(RECEIVED September 24, 1991; ACCEPTED October 31, 1991)

Abstract

The Protein Data Bank currently contains about 600 data sets of three-dimensional protein coordinates determined by X-ray crystallography or NMR. There is considerable redundancy in the data base, as many protein pairs are identical or very similar in sequence. However, statistical analyses of protein sequence–structure relations require nonredundant data. We have developed two algorithms to extract from the data base representative sets of protein chains with maximum coverage and minimum redundancy. The first algorithm focuses on optimizing a particular property of the selected proteins and works by successive selection of proteins from an ordered list and exclusion of all neighbors of each selected protein. The other algorithm aims at maximizing the size of the selected set and works by successive thinning out of clusters of similar proteins. Both algorithms are generally applicable to other data bases in which criteria of similarity can be defined and relate to problems in graph theory. The largest nonredundant set extracted from the current release of the Protein Data Bank has 155 protein chains. In this set, no two proteins have sequence similarity higher than a certain cutoff (30% identical residues for aligned subsequences longer than 80 residues), yet all structurally unique protein families are represented. Periodically updated lists of representative data sets are available by electronic mail from the file server "netserv@embl-heidelberg.de." The selection may be useful in statistical approaches to protein folding as well as in the analysis and documentation of the known spectrum of three-dimensional protein structures.

Keywords: NMR; protein data sets; X-ray crystallography

There is a continuing need for representative lists of proteins, especially in the context of statistical and rule-based approaches to the analysis and prediction of protein structure. However, data banks of protein structures and sequences (Bernstein et al., 1977; Protein Identification Resource, National Biomedical Research Foundation, Georgetown University, Washington, D.C.; Bairoch & Boeckmann, 1991) are very nonhomogeneous in the sense that some protein families are heavily represented (e.g., immunoglobulins), whereas others are only represented by a single entry. In the data base of three-dimensional (3D) protein structure, the Protein Data Bank, the problem is compounded by the fact that the same protein may appear in different crystal forms, with a variety of substrate analogues or with different engineered point mutations. Although all these data sets are useful in general, their blind use in statistical analyses would lead to serious overcounting, perhaps masking otherwise observable regularities. With the current rapid increase in the size of data banks, selection by hand of representative data sets, once enjoyable and feasible (Kabsch & Sander, 1983), becomes an increasingly boring and time-consuming proposition. The need for an automatic procedure for the selection of representative data sets is urgent.

Reprint requests to: Chris Sander, European Molecular Biology Laboratory, Meyerhofstrasse 1, D-6900 Heidelberg, Germany.

Desired properties of nonredundant data

What is a representative data set? One may want one representative per protein family (defined in evolutionary terms) or one representative per protein type (defined according to function or structure). All types or families are to be represented. The precise requirements depend on the scientific question at hand, but in general terms the selection should result in a data set that combines maximum coverage with minimum redundancy.

In this report we focus on the data base of 3D protein structures and on the following requirements. (1) No pair of proteins in the selected set should have more than a given level of sequence similarity. (2) The experimental quality of the protein structures should be optimal or meet given criteria. (3) The number of proteins in the set should be maximal, within the given constraints.

Manual selection of nonredundant data sets

Earlier selections had attempted to fulfill similar criteria. In 1983, along with a dictionary of protein secondary structures, a list of 62 selected proteins with 10,925 residues was published in which no pair of proteins had more than 50% identical residues after optimal alignment (Kabsch & Sander, 1983). Rooman and Wodak (1988), in their attempt to identify predictive sequence motifs in the protein structure data base, used a list of 75 proteins with less than 50% sequence similarity and crystallographic resolution of better than 2.5 A. Niefind and Schomburg (1991) used a list of 69 proteins with a total of 13,563 residues to derive amino acid similarity coefficients for protein modeling and sequence alignment. Unger et al. (1989) used a list of 82 chains for their building blocks approach to the analysis and prediction of protein structures. Heringa and Argos (1991) counted 157 proteins, of which no pair has more than 50% identical residues.

Problems to be solved

The principal difficulty in designing algorithms to solve this problem is combinatorial complexity: the number of potential representative sets of similar quality is very large, and it is impractical to test them all. Other more technical difficulties are due to data base development: any procedure not sufficiently robust to be routinely applied to new updates of the data base would soon leave us with an antiquated selection. Also, single Protein Data Bank data sets can contain multiple chains that have to be treated separately, and accessory information such as crystallographic resolution is not unambiguously coded for in the data sets.

Two solutions

We present two different algorithms (Fig. 1) for the selection of representative data sets from any data base in which similarity relationships can be defined. We apply these to the Protein Data Bank and derive the largest reported set of Protein Data Bank entries nonredundant at a strict level of sequence similarity.

Both solutions are conceptually very simple. Central to each is the concept of distance (or similarity) in sequence space. When two proteins are similar to each other, we will also use the terms "they are close to each other" or "they are neighbors." In its simplest form, assessment of similarity requires a one-bit decision. Two proteins are either similar to each other or they are not. The decision can be made, for example, on the basis of dynamic sequence alignment algorithms followed by application of a length-dependent threshold of similarity. Or, for protein structures, on the basis of optimal 3D alignment, followed by application of an appropriate cutoff.

Outline of algorithm 1. Given a sorted list of candidate proteins, process each protein in turn by selecting or discarding it according to the following criteria: (1) discard proteins that are similar to already selected proteins; (2) discard proteins that fail to meet additional user-specified standards.

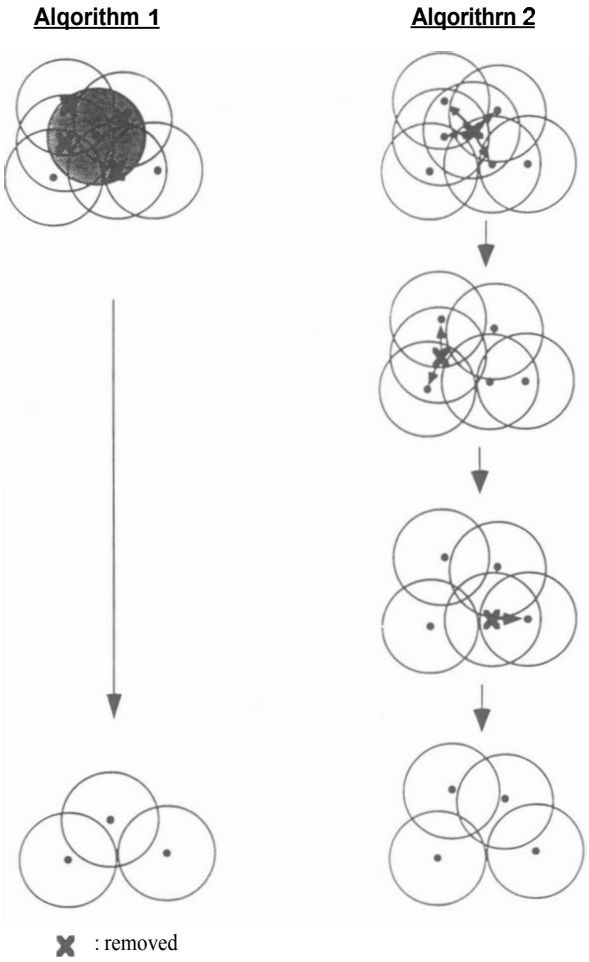

Fig. 1. How do the algorithms for the selection of sets of dissimilar proteins work? In this example, there are seven proteins in the original data base, schematically shown (top) in a two-dimensional projection of some space of properties. Two proteins that are similar to each other are close to each other in this space. A dot marks each protein and a circle centered on the dot the territory of its neighbors, i.e., all proteins considered similar to it. Circles can overlap as the similarity relationship is not transitive, i.e., protein A can be similar to both proteins C and D without C and D being similar. The task of the algorithms is to select a subset of the original set of proteins such that no two proteins in the selected subset are similar, i.e., no circle includes more than one dot.

Algorithm 1 (select until done) works by selecting some protein (center of gray circle) and removing all its neighbors, as they would be similar to an already selected protein. It then goes on to the next protein, until the data base is exhausted. Algorithm 2 (remove until done) works by removing the protein with the largest number of neighbors first, and this protein is no longer counted as a neighbor of any other protein. It then reassesses the number of neighbors and removes the protein with the largest number of neighbors in the new situation, and so on, until the proteins left over have no more neighbors. In this example, algorithm 1 (2) resulted in a nonredundant set of three (four) proteins (bottom), so the performance of algorithm 2 was superior if the goal was to maximize the number of proteins in the selected set.

Outline of algorithm 2. Given a list of candidate proteins and a list of neighbors for each of the proteins, remove one protein at a time from the list until those remaining in the list have no more neighbors in the list. The remaining proteins then represent the selected nonredundant set. As the number of neighbor relations in the original list is a constant, one can attempt to maximize the number of selected proteins by removing at each step proteins with the largest number of neighbor relations.

For either algorithm, proteins definitely to be selected or definitely to be excluded can be specified by the user prior to running the algorithm.

Results and discussion

Three lists of representative proteins

We present several lists for comparison, one generated using algorithm 1 (Fig. 2) and two generated using algorithm 2 (Fig. 3). The first list was generated from a test list ordered in terms of increasing crystallographic nominal resolution, so that its 136 protein chains with 23,295 residues tend to contain representatives of the best available resolution. The cutoff in sequence similarity is at 30% identical residues (Fig. 2). The second and third lists have been exclusively optimized for list size. The second list (Fig. 3A) uses the same cutoff in sequence similarity as the first list but has 14% more chains (155 instead of 136) and 27% more residues (29,615 instead of 23,295). The third list uses a much higher cutoff in sequence similarity, at 50% identical residues, and is correspondingly larger: 35,918 residues in 190 chains. Chains of length less than 20 residues were excluded from any list at the outset.

By construction, the first two lists, with the same cutoff, overlap in that they must contain at least all outliers, i.e., proteins that have no neighbors in the Protein Data Bank, e.g., rhodanese (1RHD) or elongation factor TU (1ETU). Only for families that have several or many members, e.g., immunoglobulins, do the lists differ. In practice, the choice as to which list to use should be carefully considered, and chains that do not suit the purpose of the investigation, e.g., membrane proteins in studies of soluble globular proteins, should be removed.

List size as a function of cutoff in sequence similarity

What cutoff in sequence similarity should be chosen in deriving representative lists? How sensitive is the size of the list to the cutoff? Figure 4, generated with algorithm 2, shows that the size of the list changes only gradually, from 135 protein chains at the threshold for structural homology (Sander & Schneider, 1991), i.e., 25% for aligned subsequences of more than 80 residues, to 155 at 30%, 190 at 50%, and finally, to the full size of the data base, 764 chains, at the extreme limit of 100%, corresponding to treating all sequences as dissimilar. The sharp increase at 100% simply reflects the fact that about one-half of the 764 chains are 100% identical in sequence to another chain in the Protein Data Bank (e.g., sequence-identical subunits; same protein in different crystal forms, etc.).

In practice, the precise value of the cutoff has only weak influence on the size of the sample implied by the list, except for very permissive values of the cutoff. One way to choose the cutoff is to make it as low as possible, consistent with the requirement of having at least one representative from each structural class (see Methods). With current alignment techniques, this leads to a cutoff at about 30% identical residues, which is used in the first (Fig. 2) and second (Fig. 3A) list.

Advantages and disadvantages

The first algorithm (select until done) optimizes a user-defined property of the selected set of proteins, such as crystallographic resolution, by sorting the candidate list according to that property. It is faster in that not all pair comparisons have to be calculated, as proteins in the skip list do not need to be compared to all other proteins. The second algorithm (remove until done) maximizes the number of proteins in the final selection. It is, however, more time-consuming, as all pair relations are needed. Preprocessing filters can be applied to the lists from either algorithm, in order to impose additional requirements. For example, when low-resolution crystal structures are removed from the outset, say structures with resolution not better than 3.0 Å, the second algorithm will aim at generating the longest possible representative list of structures with better than 3.0 Å resolution. In practice, the algorithms, applied to the current Protein Data Bank, can run to completion on a work station computer (e.g., SPARCstation 2, DECstation 5000) in a matter of 2–3 h (algorithm 1) and 2–3 days (algorithm 2, including about $3*10^5$ sequence comparisons).

Sequence-unique or structure-unique?

The similarity relationship in the set of all proteins required for posing this problem can be based on sequence alignment, optimal superposition of 3D coordinates (Taylor & Orengo, 1989; Vriend & Sander, 1991), or other criteria. If the goal is to have a set of structurally unique proteins, then explicit structural superposition should be used, rather than sequence alignment. A fundamental limitation of current sequence alignment algorithms is that they can only establish, beyond reasonable doubt, that two proteins are similar in structure, when the sequence similarity exceeds a certain threshold. But they cannot establish that two proteins are dissimilar in structure when sequence similarity is very low. For example, we can be certain that endothiapepsin 4APE and rhizo-

PDB	C	Exclusion	NAA	RES	%STR	%NCS	%NHE	SwissProt	Name, Function, Species
451C	_		82	1.6	52	0	7	C551$PSEAE	CYTOCHROME C551, ELECTRON TRANSPORT (PSEUDOMONAS AERUGINOSA)
256B	A		106	1.4	81	0	6	C562$ECOLI	CYTOCHROME B562, ELECTRON TRANSPORT (ESCHERICHIA $COLI)
2AAT	_	Res	396	2.8	47	0	1	AAT$ECOLI	ASPARTATE AMINOTRANSFERASE, TRANSFERASE (ESCHERICHIA COLI)
1ABP	_		306	2.4	49	0	0	ARAF$ECOLI	L-ARABINOSE-BINDING PROTEIN,(ESCHERICHIA COLI)
1ACX	_	Sec	107	2.0	34	4	0	ATXA$STRGL	ACTINOXANTHIN, ANTIBACTERIAL PROTEIN (ACTINOMYCES GLOBISPORUS)
8ADH	_		374	2.4	51	0	0	ADHE$HORSE	APO-LIVER ALCOHOL DEHYDROGENASE, OXIDOREDUCTASE (EQUUS CABALLUS, LIVER)
3ADK	_		194	2.1	66	0	1	KAD1$PIG	ADENYLATE KINASE, TRANSFERASE(PHOSPHOTRANSFERASE) (SUS SCROFA)
9API	A	Res	339	3.0	62	0	4	A1AT$HUMAN	ALPHA-1-ANTITRYPSIN, PROTEINASE INHIBITOR MODIFIED (HOMO SAPIENS)
9API	B	SizRes	36	3.0	62	0	4	A1AT$HUMAN	ALPHA-1-ANTITRYPSIN, PROTEINASE INHIBITOR MODIFIED (HOMO SAPIENS)
8ATC	A		310	2.5	54	0	0	PYRB$ECOLI	ASPARTATE CARBAMOYLTRANSFERASE, TRANSFERASE (E. COLI)
8ATC	B		146	2.5	54	0	0	PYRI$ECOLI	ASPARTATE CARBAMOYLTRANSFERASE, TRANSFERASE (E. COLI)
2AZA	A		129	1.8	52	2	1	AZUR$ALCDE	AZURIN, ELECTRON TRANSPORT PROTEIN (ALCALIGENES DENITRIFICANS)
3B5C	_		85	1.5	64	0	6	CYB5$BOVIN	CYTOCHROME B5, ELECTRON TRANSPORT (BOS TAURUS, LIVER)
1BDS	_	SecSs	43	NMR	28	14	0	BDS1$ANESU	ANTI-VIRAL PROTEIN, ANTI-HYPERTENSIVE (ANEMONIA SULCATA)
3BLM	_		257	2.0	58	0	0	BLAC$STAAU	BETA-LACTAMASE, HYDROLASE (STAPHYLOCOCCUS AUREUS)
1BMV	1	Res	185	3.0	43	0	0	----------	BEAN POD MOTTLE VIRUS (MIDDLE COMPONENT) BOUNTIFUL BEAN
1BMV	2	Res	374	3.0	43	0	0	----------	BEAN POD MOTTLE VIRUS (MIDDLE COMPONENT) BOUNTIFUL BEAN
1BP2	_	Ss	123	1.7	58	11	2	PA2$BOVIN	PHOSPHOLIPASE A2, HYDROLASE (BOS TAURUS, PANCREAS)
2CA2	_		256	1.9	50	0	0	CAH2$HUMAN	CARBONIC ANHYDRASE, LYASE (HOMO SAPIENS, ERYTHROCYTES)
1CBH	_	SizSs	36	NMR	39	11	0	GUX1$TRIRE	CELLOBIOHYDROLASE I (C-TERMINAL DOMAIN), HYDROLASE (TRICHODERMA REESEI)
1CC5	_		83	2.5	47	2	8	CYC5$AZOVI	CYTOCHROME C5 (OXIDIZED), ELECTRON TRANSPORT (AZOTOBACTER VINELANDII)
1CCR	_		111	1.5	48	0	6	CYC$ORYSA	CYTOCHROME C, ELECTRON TRANSPORT (ORYZA SATIVA)
2CCY	A		127	1.67	76	0	5	CYCP$RHOMO	CYTOCHROME C, ELECTRON TRANSPORT (RHODOSPIRILLUM MOLISCHIANUM)
1CD4	_		173	2.3	43	2	0	CD4$HUMAN	T-CELL SURFACE GLYCOPROTEIN CD4 (HOMO SAPIENS)
2CDV	_	Het	107	1.8	37	0	21	CYC3$DESVM	CYTOCHROME C3, ELECTRON TRANSPORT (DESULFOVIBRIO VULGARIS)
3CLA	_		213	1.75	54	0	1	CAT3$ECOLI	CHLORAMPHENICOL ACETYLTRANSFERASE, TRANSFERASE (E. COLI)
4CPA	I	SizSs	37	2.5	52	16	0	ICBP$SOLTU	CARBOXYPEPTIDASE A INHIBITOR
5CPA	_		307	1.54	55	1	0	CBPA$BOVIN	CARBOXYPEPTIDASE A-ALPHA, HYDROLASE (BOS TAURUS, PANCREAS)
2CPP	_		405	1.63	61	0	2	CPXA$PSEPU	CYTOCHROME P450CAM, OXIDOREDUCTASE (PSEUDOMONAS PUTIDA)
4CPV	_		108	1.5	56	0	0	PRVB$CYPCA	PARVALBUMIN, CALCIUM BINDING CALCIUM-BINDING (CYPRINUS CARPIO)
1CRN	_	Ss	46	1.5	50	13	0	CRAM$CRAAB	CRAMBIN, PLANT SEED PROTEIN (CRAMBE ABYSSINICA)
1CRO	_		66	2.2	55	0	0	RCRO$LAMBD	CRO REPRESSOR, GENE REGULATING PROTEIN, BACTERIOPHAGE (LAMBDA)
1CSE	E		274	1.2	53	0	0	SUBT$BACLI	SUBTILISIN CARLSBERG, SERINE PROTEINASE (BACILLUS SUBTILIS)
1CSE	I		63	1.2	53	0	0	ICIC$HIRME	EGLIN-C (HIRUDO MEDICINALIS)
1CTF	_		68	1.7	69	0	1	RL7$ECOLI	RIBOSOMAL PROTEIN L7/L12 (C-TERMINAL DOMAIN) (ESCHERICHIA COLI)
2CYP	_		293	1.7	56	0	2	CCPR$YEAST	CYTOCHROME C PEROXIDASE, OXIDOREDUCTASE (SACCHAROMYCES CEREVISIAE)
3DFR	_		162	1.7	58	0	6	DYR$CHICK	DIHYDROFOLATE REDUCTASE, OXIDOREDUCTASE (GALLUS GALLUS, LIVER)
1ECN	_		136	1.4	75	0	4	GLB3$CHITH	HEMOGLOBIN, OXYGEN TRANSPORT (CHIRONOMOUS THUMMI THUMMI)
2ER7	E		330	1.6	52	1	2	CARP$CRYPA	ENDOTHIAPEPSIN, HYDROLASE (ACID PROTEINASE) (ENDOTHIA PARASITICA)
1ETU	_	Res	177	2.9	59	0	2	EFTU$ECOLI	ELONGATION FACTOR TU, TRANSPORT AND PROTECTION PROTEIN (ESCHERICHIA COLI)
2FB4	H		229	1.9	53	3	0	HV3K$HUMAN	IMMUNOGLOBULIN FAB, IMMUNOGLOBULIN (HOMO SAPIENS)
1FC2	C	Res	43	2.8	51	0	6	PROA$STAAU	PROTEIN A (FRAGMENT B) (STAPHYLOCOCCUS AUREUS)
1FC2	D	Res	206	2.8	51	2	6	GC1$HUMAN	IMMUNOGLOBULIN FC, IMMUNOGLOBULIN (HOMO SAPIENS)
1FD2	_		106	1.9	48	0	2	FER1$AZOVI	FERREDOXIN, ELECTRON TRANSPORT (AZOTOBACTER VINELANDII)
1FXB	_	Sec	81	2.3	25	0	1	FERS$BACTH	FERREDOXIN, ELECTRON TRANSPORT (BACILLUS THERMOPROTEOLYTICUS)
3FXC	_	Sec	98	2.5	26	0	1	FER$SPIPL	FERREDOXIN, ELECTRON TRANSPORT (SPIRULINA PLATENSIS)
4FXN	_		138	1.8	59	0	3	FLAV$CLOSP	FLAVODOXIN, ELECTRON TRANSPORT (CLOSTRIDIUM MP)
3GAP	A		208	2.5	52	0	1	CRP$ECOLI	CATABOLITE GENE ACTIVATOR PROTEIN, GENE REGULATORY PROTEIN (E. COLI)
2GBP	_		309	1.9	65	0	1	DGAL$ECOLI	GALACTOSE/GLUCOSE BINDING PROTEIN, PERIPLASMIC BINDING PROTEIN (E. COLI)
1GCN	_	SizRes	29	3.0	62	0	0	GLUC$PIG	GLUCAGON, HORMONE (SUS SCROFA, PANCREAS)
1GCR	_		174	1.6	48	0	0	CRGB$BOVIN	CRYSTALLIN GAMMA-II, (BOS TAURUS, EYE LENS PROTEIN)
1GD1	O		334	1.8	60	0	2	G3P$BACST	D-GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE (BACILLUS STEAROTHERMOPHILUS)
2GLS	D	Res	92	3.5	50	0	0	GLNA$SALTY	GLUTAMINE SYNTHETASE, LIGASE(AMIDE SYNTHETASE) (SALMONELLA TYPHIMURIUM)
2GN5	_	Sec	87	2.3	14	0	0	VHED$BPFD	GENE 5 DNA BINDING PROTEIN (VIRAL) FILAMENTOUS BACTERIOPHAGE FD (M13)
1GOX	_		350	2.0	59	0	1	2HAO$SPIOL	GLYCOLATE OXIDASE, OXIDOREDUCTASE (SPINACIA OLERACEA)
1GP1	A		183	2.0	49	0	1	GSHP$BOVIN	GLUTATHIONE PEROXIDASE, OXIDOREDUCTASE (BOS TAURUS, ERYTHROCYTE)
1HIP	_		85	2.0	39	0	1	HPIS$CHRVI	OXIDIZED HIGH POTENTIAL IRON PROTEIN,ELECTRON TRANSFER (CHROMATIUM V.)
5HIR	_	SecSs	49	NMR	29	12	0	ITH1$HIRME	HIRUDIN, COAGULATION INHIBITOR (HIRUDO MEDICINALIS)
3HLA	B		99	2.6	61	2	0	HA1H$HUMAN	HUMAN CLASS I HISTOCOMPATIBILITY ANTIGEN, (HOMO SAPIENS)
3HMG	A	Res	328	2.9	50	3	2	HEMA$INAAI	HEMAGGLUTININ, INFLUENZA VIRUS
3HMG	B	Res	175	2.9	50	2	2	HEMA$INAAI	HEMAGGLUTININ, INFLUENZA VIRUS
1HOE	_		74	2.0	46	5	0	IAA$STRTE	ALPHA-AMYLASE INHIBITOR, AGLYCOSIDASE INHIBITOR (STREPTOMYCES TENDAE)
4HVP	A		97	2.3	47	0	3	POL$HIV1A	HIV-1 PROTEASE, HYDROLASE (ACID PROTEINASE)
2I1B	_		153	2.0	49	0	0	IL1B$HUMAN	INTERLEUKIN-1 BETA, CYTOKINE (HOMO SAPIENS)
3ICB	_		75	2.3	59	0	1	CAB1$BOVIN	CALCIUM-BINDING PROTEIN (VITAMIN D-DEPENDENT) (BOS TAURUS)
4ICD	_		414	2.5	56	0	0	IDH$ECOLI	ISOCITRATE DEHYDROGENASE, OXIDOREDUCTASE (ESCHERICHIA COLI)
1IL8	A		71	NMR	52	6	0	IL8$HUMAN	INTERLEUKIN 8, CYTOKINE (HOMO SAPIENS, RECOMBINANT IN E. COLI)
3INS	A	SizSs	21	1.5	56	19	0	INS$PIG	INSULIN, HORMONE (SUS SCROFA)
3INS	B	Siz	30	1.5	56	7	0	INS$PIG	INSULIN, HORMONE (SUS SCROFA)
1L24	_		164	1.7	65	0	0	LYCV$BPT4	LYSOZYME (MUTANT), HYDROLASE (O-GLYCOSYL), BACTERIOPHAGE T4
2LBP	_		346	2.4	61	1	0	LIVK$ECOLI	LEUCINE-BINDING PROTEIN, PERIPLASMIC BINDING PROTEIN (E. COLI)
6LDH	_		329	2.0	60	0	0	LDHM$SQUAC	APO-LACTATE DEHYDROGENASE, OXIDOREDUCTASE (SQUALUS ACANTHIAS, MUSCLE)
1LH3	_		153	2.0	75	0	4	LGB2$LUPLU	LEGHEMOGLOBIN (CYANO,MET), OXYGEN TRANSPORT, (LUPINUS LUTEUS L)
2LTN	A		181	1.7	54	0	0	LEC$PEA	LECTIN (PISUM SATIVUM, SEEDS)
2LTN	B		47	1.7	54	0	0	LEC$PEA	LECTIN (PISUM SATIVUM, SEEDS)
1LZ1	_		130	1.5	56	6	0	LYC$HUMAN	LYSOZYME, HYDROLASE (O-GLYCOSYL), (HOMO SAPIENS)
1MBD	_		153	1.4	76	0	4	MYG$PHYCA	MYOGLOBIN, OXYGEN STORAGE (PHYSETER CATODON)
4MDH	A		333	2.5	58	0	2	MDHC$PIG	CYTOPLASMIC MALATE DEHYDROGENASE, OXIDOREDUCTASE (SUS SCROFA, heart)
2MEV	4	Res	58	3.0	36	0	0	POLG$ENMGO	MENGO ENCEPHALOMYOCARDITIS VIRUS COAT PROTEIN MONKEY BRAIN (MENGO VIRUS)
2MHR	_		118	1.7	68	0	1	HEMM$THEZO	MYOHEMERYTHRIN, OXYGEN BINDING (THEMISTE ZOSTERICOLA, RETRACTOR MUSCLE)
2MLT	A	SizMem	26	2.0	77	0	3	MEL1$APIME	MELITTIN, TOXIN (HEMOLYTIC POLYPEPTIDE) (APIS MELLIFERA, VENOM)
2MRB	_	SecSiz	31	NMR	3	0	1	MT2$RABBIT	CD-7 METALLOTHIONEIN-2A, METALLOTHIONEIN (ORYCTOLAGUS CUNICULUS. liver)
1MRT	_	SecSiz	31	NMR	10	0	1	MT2$RAT	CD-7 METALLOTHIONEIN-2, METALLOTHIONEIN (RATTUS RATTUS,liver)
1NXB	_	Ss	62	1.38	37	13	1	NXS1$LATSE	NEUROTOXIN B, (LATICAUDA SEMIFASCIATA)

Fig. 2. *Continues on facing page.*

```
2OVO _ Ss          56  1.5   36   11    0 IOVO$TYMCU OVOMUCOID THIRD DOMAIN, PROTEINASE INHIBITOR (LOPHURA NYCTHEMERA)
2PAB A             114  1.8   56    0    0 TTHY$HUMAN THYROXINE, TRANSPORT (RETINOL) IN SERUM PREALBUMIN (HOMO SAPIENS)
9PAP _             212 1.65   48    3    4 PAPA$CARPA PAPAIN, HYDROLASE (SULFHYDRYL PROTEINASE) (CARICA PAPAYA)
1PAZ _             120 1.55   51    0    0 AZUP$ALCFA PSEUDOAZURIN, ELECTRON TRANSFER, (ALCALIGENES FAECALIS, STRAIN S-6)
1PCY _              99  1.6   44    0    0 PLAS$POPNI PLASTOCYANIN, ELECTRON TRANSPORT (POPULUS NIGRA, leaves)
4PEP _             326  1.8   56    2    0 PEPA$PIG   PEPSIN, HYDROLASE (SUS SCROFA)
2PFK A             301  2.4   65    0    0 K6P1$ECOLI PHOSPHOFRUCTOKINASE, TRANSFERASE(PHOSPHOTRANSFERASE) (E. COLI)
3PGM _ Res         230  2.8   37    0    1 PMGY$YEAST PHOSPHOGLYCERATE MUTASE, TRANSFERASE (SACCHAROMYCES CEREVISIAE)
1PHH _             394  2.3   56    0    2 PHHY$PSEFL p-HYDROXYBENZOATE HYDROXYLASE, OXIDOREDUCTASE (PSEUDOMONAS FLUORESCENS)
2PLV 2 Res         268 2.88   40    0    1 POLH$POL1M POLIOVIRUS, PICORNAVIRUS POLIOVIRUS (HOMO SAPIENS)
2PLV 3 Res         235 2.88   40    0    1 POLH$POL1M POLIOVIRUS, PICORNAVIRUS POLIOVIRUS (HOMO SAPIENS)
2PLV 4 Res          62 2.88   40    0    1 POLH$POL1M POLIOVIRUS, PICORNAVIRUS POLIOVIRUS (HOMO SAPIENS)
1PPT _ Siz          36 1.37   53    0    0 PAHO$CHICK AVIAN PANCREATIC HORMONE  (MELEAGRIS GALLOPAVO, PANCREAS)
1PRC C Mem         332  2.3   58    0    8 CYCR$RHOVI PHOTOSYNTHETIC REACTION CENTER (RHODOPSEUDOMONAS VIRIDIS)
1PRC H Mem         258  2.3   58    0    8 CYCR$RHOVI PHOTOSYNTHETIC REACTION CENTER (RHODOPSEUDOMONAS VIRIDIS)
1PRC L Mem         273  2.3   58    0    8 CYCR$RHOVI PHOTOSYNTHETIC REACTION CENTER (RHODOPSEUDOMONAS VIRIDIS)
1PRC M Mem         323  2.3   58    0    8 CYCR$RHOVI PHOTOSYNTHETIC REACTION CENTER (RHODOPSEUDOMONAS VIRIDIS)
5PTI _ SecSs        58  1.0   34   10    1 BPT1$BOVIN TRYPSIN INHIBITOR, PROTEINASE INHIBITOR (BOS TAURUS, PANCREAS)
4PTP _             223 1.34   47    5    0 TRYP$BOVIN BETA TRYPSIN, HYDROLASE (SERINE PROTEINASE) (BOS TAURUS, PANCREAS)
1PYP _ SecRes      280  3.0   30    0    0 IPYR$YEAST INORGANIC PYROPHOSPHATASE, ACID ANHYDRIDE HYDROLASE (SACCHAROMYCES C.)
1R69 _              63  2.0   54    0    0 RPC1$BP434 434 REPRESSOR (N-TERMINAL DOMAIN), GENE REGULATING PROTEIN (PHAGE 434)
1RHD _             293  2.5   45    0    0 THTR$BOVIN RHODANESE, TRANSFERASE (BOS TAURUS, LIVER)
2RHE _             114  1.6   48    2    0 ---------- BENCE-JONES PROTEIN (LAMBDA, V-DOMAIN), IMMUNOGLOBULIN  (HOMO SAPIENS)
1RMU 4 Res          40  3.0   41    0    0 POLG$HRV14 RHINOVIRUS 14 COAT PROTEIN  (HOMO SAPIENS)
2RNT _ Het         104  1.8   48    4    9 RNT1$ASPOR RIBONUCLEASE T1, HYDROLASE (ENDORIBONUCLEASE) (ASPERGILLUS ORYZAE)
7RSA _             124 1.26   51    6    1 RNP$BOVIN  RIBONUCLEASE A, HYDROLASE (PHOSPHORIC DIESTER) (BOS TAURUS, PANCREAS)
2RSP A             115  2.0   49    0    0 GAG$RSVP   ROUS SARCOMA VIRUS PROTEASE, HYDROLASE (ASPARTYL PROTEINASE)
5RXN _              54  1.2   44    0    0 RUBR$CLOPA RUBREDOXIN, ELECTRON TRANSFER (CLOSTRIDIUM PASTEURIANUM)
4SBV A Res         199  2.8   51    1    0 COAT$SOBMV SOUTHERN BEAN MOSAIC VIRUS COAT PROTEIN , COAT PROTEIN (VIRAL)
2SGA _             181  1.5   60    2    0 PRTA$STRGR PROTEINASE A, HYDROLASE (SERINE PROTEINASE)  (STREPTOMYCES GRISEUS)
4SGB I              51  2.1   55    0    1 IPR2$SOLTU SERINE PROTEINASE-INHIBITOR (SOLANUM TUBEROSUM)
1SN3 _ Ss           65  1.8   40   12    2 SCX3$CENSC SCORPION NEUROTOXIN (CENTRUROIDES SCULPTURATUS EWING)
2SNS _             141  1.5   47    0    2 NUC$STAAU  NUCLEASE, HYDROLASE (PHOSPHORIC DIESTER) (STAPHYLOCOCCUS AUREUS)
2SOD O             151  2.0   41    1    0 SODC$BOVIN SUPEROXIDE DISMUTASE, OXIDOREDUCTASE (BOS TAURUS, ERYTHROCYTE)
2SSI _             107  2.6   46    4    0 ISUB$STRAO SUBTILISIN INHIBITOR, PROTEINASE INHIBITOR (STREPTOMYCES ALBOGRISEOLUS)
2STV _             184  2.5   51    0    0 COAT$STNV  TOBACCO NECROSIS VIRUS COAT PROTEIN OF SATELLITE TOBACCO NECROSIS VIRUS
2TBV A Res         283  2.9   38    0    0 COAT$TBSVB TOMATO BUSHY STUNT VIRUS
1TGS I Ss           56  1.8   48   11    0 IPST$PIG   PORCINE PANCREATIC SECRETORY TRYPSIN INHIBITOR (SUS SCROFA, PANCREAS)
1THB A             141  1.5   74    0    5 HBA$HUMAN  HEMOGLOBIN, OXYGEN TRANSPORT (HOMO SAPIENS)
5TMN E             316  1.6   56    0    1 THER$BACTH THERMOLYSIN, HYDROLASE (BACILLUS THERMOPROTEOLYTICUS)
2TMV P Res         154  2.9   52    0    0 COAT$TMV   INTACT TOBACCO MOSAIC VIRUS
5TNC _             161  2.0   64    0    0 TPCS$MELGA TROPONIN-C, CONTRACTILE SYSTEM PROTEINS (MELEAGRIS GALLOPAVO)
1TNF A             152  2.6   42    1    0 TNFA$HUMAN TUMOR NECROSIS FACTOR-ALPHA, LYMPHOKINE (HOMO SAPIENS)
2TS1 _             317  2.3   63    0    0 SYY$BACST  TYROSYL-TRANSFER RNA SYNTHETASE, LIGASE (BACILLUS STEAROTHERMOPHILUS)
1UBQ _              76  1.8   55    0    0 UBIQ$HUMAN UBIQUITIN, CHROMOSOMAL PROTEIN (HOMO SAPIENS, ERYTHROCYTES)
1UTG _              70 1.34   73    0    0 UTER$RABIT UTEROGLOBIN, STEROID BINDING (ORYCTOLAGUS CUNICULUS)
9WGA A Ss          171  1.8   41   19    0 AGI2$WHEAT AGGLUTININ, LECTIN (AGGLUTININ) (TRITICUM VULGARIS, GERM)
2WRP R             104 1.65   70    0    1 TRPR$ECOLI TRP REPRESSOR, DNA BINDING REGULATORY PROTEIN (ESCHERICHIA COLI)
1WSY A             248  2.5   62    0    0 TRPA$SALTY TRYPTOPHAN SYNTHASE, LYASE (CARBON-OXYGEN) (SALMONELLA TYPHIMURIUM)
1WSY B             385  2.5   62    0    0 TRPA$SALTY TRYPTOPHAN SYNTHASE, LYASE(CARBON-OXYGEN)  (SALMONELLA TYPHIMURIUM)
4XIA A             393  2.3   58    0    0 XYLA$ARTS7 D-XYLOSE ISOMERASE, ISOMERASE (INTRAMOLECULAR OXIDOREDUCTSE)
1YPI A             247  1.9   59    0    0 TPIS$YEAST TRIOSE PHOSPHATE ISOMERASE (SACCHAROMYCES CEREVISIAE)
```

Fig. 2. *Continued from facing page.* Selected set of 136 nonredundant proteins with optimal crystallographic resolution according to algorithm 1. The chains contain a total of 23,295 residues. No pair of chains in the list exceeds 30% sequence identity after optimal alignment (more precisely, five percentage points above the length-dependent threshold for structural homology [Sander & Schneider, 1991], i.e., 29.8% for alignments of length 80 or longer, higher cutoff for shorter alignments). Column notation is as follows: PDB, Protein Data Bank four-letter data set identifier, with proteins sorted on the first letter (second byte); C, chain identifier, in cases where the data set contains more than one protein chain, or else "-"; NAA, number of amino acids in the protein chain; RES, crystallographic resolution in Å, or NMR for structures determined by nuclear magnetic resonance spectroscopy; %STR, number of backbone hydrogen bonds involved in secondary structure, per 100 residues (in parallel and antiparallel bridges, and in $(i, i + 4)$ and $(i, i + 3)$ type H-bonds defined as in Kabsch and Sander (1983); %NCS, number of Cys residues in disulfide bonds, per 100 residues; %NHE, ratio of the number of heteroatoms (HETATM records) to the number of protein atoms (ATOM records) in the PDB data set (not just in the chain), times 100; Swissprot, sequence identifier in the Swissprot protein sequence data base (Bairoch & Boeckmann, 1991), where sequence identity between SWISS-PROT and PDB entry is more than 98% and total sequence length is within three residues; Name, Function, Species, protein name, function, and species as taken from the COMPND, HEADER, and SOURCE records in the Protein Data Bank; Exclusion, criteria by which a protein may be excluded by the end-user (Res, insufficient resolution: RES > 2.6 Å; Sec, insufficient amount of secondary structure: %STR < 35%; Siz, too small: NAA < 40; Ss, too many SS bonds: %NCS > 8%; Mem, membrane protein or part of a membrane complex). No firm recommendation for removing particular proteins is implied here.

puspepsin 2APR have the same structure on the basis of observing 37% identities in 320 aligned residues. But we cannot assert that phage 434 repressor (1R69) and phage lambda cro repressor (1CRO) do not have similar structure on the basis of observing only 16% identities in 62 aligned residues; in fact, parts of their structures are similar in that over a three-helix motif of 38 residues they have a root mean square position deviation (rmsd) of C_α atoms of less than 2.9 Å.

For these reasons the lists derived here do contain

A

351C	1BMV-2	1CSC	3GAP-A	1IL8-A	2MRT	1RED	4TMN-E
256B-A	2BP2	1CTF	2GBP	2INS-A	2OR1-L	1RMU-1	2TMV-P
2AAT	1BRD	1CY3	1GCN	4INS-B	1P09-A	1RNS-S	5TMC
1ABP	2C2C	2CYP	1GCR	2KAI-B	2PAB-A	1RNT	1TNF-A
6ACN	3CA2	3DFR	2GD1-O	1L12	1PAZ	2RR1-4	2TPI-I
2ACT	7CAT-A	4DFR-A	2GLS-A	5LDH	2PCY	2RSP-A	4TS1-A
1ACX	1CBE	2DHF-A	2GN5	1LH1	1PFK-A	5RXN	1UBQ
8ADH	1CC5	5EBX	1GP1-A	2LIV	3PGK	2SBT	2UTG-A
3ADK	2CCY-A	1ECA	3GRS	1LRD-3	3PGM	4SBV-A	9WGA-A
4AIT	1CD4	4ER4-E	1HIP	2LTN-A	1PHH	4SGB-I	1WRP-R
8API-A	1CHO-I	1ETU	6HIR	2LTN-B	2PKA-A	1SGT	1WSY-A
8API-B	1CLA	1F19-H	2HLA-B	1LZ1	1PPT	1SN3	1WSY-B
2AT1-B	1CMS	1F19-L	2HLA-B	1MBA	1PRC-C	2SNS	5XIA-A
2ATC-A	1COH-B	1FC1-A	3HMG-A	5MBN	1PRC-H	2SOD-B	2YHX
2AZA-A	5CPA	1FC2-C	3HMG-B	4MDH-A	1PRC-L	2SSI	2YPI-A
3B5C	4CPA-I	1FCB-A	1HME-A	2MEV-1	1PRC-M	2STV	
3BCL	2CPP	2FD2	4HVP-A	2MEV-3	1PYP	2TAA-A	
1BDS	5CPV	1FX1	2I1B	2MEV-4	2R06-3	2TBV-A	
3BLM	1CRN	1FXB	3ICB	1MHU	1R08-2	1TEC-I	
1BMV-1	1CRO	3FXC	3ICD	2MLT-A	1RBB-A	1TGS-I	

B

351C	1BDS	4CPA-I	1ETU	3GRS	2KAI-B	2MHR	1PRC-H	6RXN	1TPA-I
155C	3BLM	3CPP	1FC1-A	3RHB-A	1L18	1MHU	1PRC-L	4SBV-A	4TS1-A
256B-A	1BMV-1	5CPV	1FC2-C	1HIP	1LDB	2MLT-A	1PRC-M	4SGB-I	1UBQ
2AAT	1BMV-2	1CRN	1FCB-A	5HIR	3LDH	2MRT	2PRK	1SGC	2UTG-A
1ABP	1BP2	1CRO	2FD2	1HKG	1LH5	1P06-A	1PSG	1SGT	7WGA-A
2ABX-A	1BRD	2CRO	1FDX	3HLA-A	2LHB	2PAB-A	1PYP	1SN3	3WRP
5ACN	2BOS	1CSE-I	1FX1	3HLA-B	2LIV	2PAD	2R06-2	2SNI-E	1WSY-A
2ACT	3C2C	1CTF	1FXB	3HMG-A	1LRD-3	2PAZ	2R07-4	2SNS	1WSY-B
1ACX	1CA2	5CTS	3FXC	3HMG-B	2LTN-A	4PCY	1R08-1	2SOD-B	4XIA-A
8ADH	7CAT-A	1CY3	3FXN	1HMQ-A	2LTN-B	4PFK	1R1A-1	2SSI	1YPI-A
3ADK	1CBH	2CYP	3GAP-A	1HNE-E	2LYM	3PGK	1R69	2STV	
1ALC	1CC5	3CYT-I	2GBP	1HOE	3MBA	3PGM	1REI-A	2TAA-A	
9API-A	2CCY-A	3DFR	4GCH	4HVP-A	5MBN	1PHH	1RHD	2TBV-A	
9API-B	1CD4	8DFR	1GCN	1I1B	2MCG-1	2PKA-A	1RN3	1TEC-E	
4APR-E	2CDV	4DFR-A	1GCR	3ICB	1MCP-H	2PLV-1	1RNS-S	1TGS-I	
7AT1-A	1CHO-I	2DHB-B	2GLS-A	3ICD	4MDH-A	2PLV-3	3RNT	1TLD	
7AT1-B	2CI2-I	5EBX	2GN5	2IG2-H	2MEV-1	2PLV-4	3RP2-A	3TLN	
1AZU	3CLA	1ECH	1GOX	1IL8-A	2MEV-2	1PP2-L	2RS3-3	2TMV-P	
3B5C	1CM1-A	2ER0-E	1GP1-A	2INS-A	2MEV-3	1PPT	2RSP-A	4TNC	
3BCL	4CPA	1EST	4GPD-1	2INS-B	2MEV-4	1PRC-C	4RXN-O	1TNF-A	

Fig. 3. Two selected sets of nonredundant protein chains according to algorithm 2. **A:** A total of 155 chains with 29,615 residues, using a cutoff in sequence similarity at 30% identical residues (more precisely, five percentage points above the threshold for structural homology [Sander & Schneider, 1991], see Fig. 1 and Methods). **B:** A total of 190 chains with 35,918 residues, using a cutoff in sequence similarity at 50% identical residues (25 percentage points above the threshold). Only the four-letter Protein Data Bank identifiers and the one-letter chain identifiers are given. When no chain identifier is given, the chain with a blank character (" ") in the chain column of the atomic coordinate lines in the Protein Data Bank data set is used.

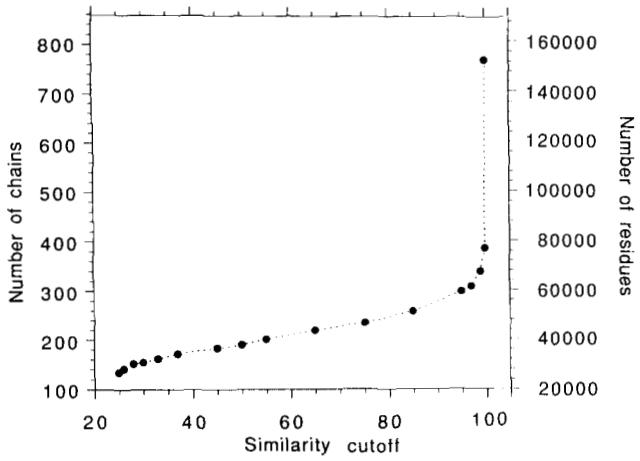

Fig. 4. Size of the lists of representative proteins extracted from the Protein Data Bank as a function of the cutoff in sequence similarity. The graph is a guide for choosing a particular list for a particular problem. Each point represents a list in which no two proteins have sequence similarity higher than the given cutoff. The more severe the cutoff, the shorter the list. List size is defined as the number of protein chains (left vertical axis) or as the total number of residues in these chains (right). The cutoff in sequence similarity is in percentage of identical residues after optimal alignment of all protein pairs. For aligned sequence pairs shorter than 80 residues, a higher cutoff is applied, following the length dependence of the threshold for structural homology (Sander & Schneider, 1991) by adding a fixed number of percentage points to the cutoff. For example, adding plus five percentage points to the length-dependent threshold corresponds to 29.8% on the horizontal axis. The precise size of the list may vary from one run to the next, using different random number seeds, as the algorithm is not mathematically guaranteed to find the global optimum. In practice, the size of the list varies by less than 0.5% (0.47% standard deviation of the number of residues in 50 lists at the 29.8% cutoff).

groups of proteins of similar structures, although they are minimally redundant in sequence. For example, the following sets of proteins of similar 3D fold are in the first list (Fig. 2): immunoglobulin-like proteins 1CD4 (human T-cell receptor), 2RHE (human lambda immunoglobulin variable domain), and 2FB4 (human immunoglobulin heavy chain), 1 FC2 (human gamma immunoglobulin FC region); globins 1LH3 (plant leghemoglobin), 1THB (human hemoglobin), 1MBD (sperm whale myoglobin), and 1ECN (insect erythrocruorin); small copper-binding proteins 1PAZ (bacterial azurin), 1PCY (plant plastocyanin), and 2AZA (bacterial pseudoazurin); helix–turn–helix repressors 2WRP (bacterial TRP repressor), 1R69 (phage 434 repressor), 1CRO (bacteriophage cro repressor); spherical virus coat proteins from 2MEV (mengo virus), 1RMU (rhinovirus), 2PLV (poliovirus), 2STV (tobacco necrosis virus), and 2TBV (tomato bushy stunt virus). Structural similarity in these examples can be established by optimal structural alignment. For example, T-cell receptor 2CD4 and immunoglobulin 2RHE have an rmsd of C_α atoms of 1.2 Å over 79 residues with a sequence identity of 24%; or, sperm whale myoglobin 1MBD and insect erythrocruorin 1ECN have an rmsd of 1.7 Å over 118 residues with a sequence identity of 18%.

Is any structural family not represented in the lists? If it is desirable to have at least one representative from each structural family, the cutoff in sequence similarity has to be chosen appropriately, using independent knowledge of what constitutes structural families. We find that a cutoff of five percentage points above the (length-dependent) threshold of structural similarity (Sander & Schneider, 1991) (see Methods) avoids rejection of structurally unique proteins on the grounds of spurious sequence similarity to another protein in the list. An interesting borderline case is the pair 2WRP (TRP repressor) and 1ETU (elongation factor) for which the FASTA alignment algorithm (Pearson & Lipman, 1988) detects 37% identical residues in a 49-residue overlap, 4.1 percentage points above the threshold, with four gaps of total gap length of 5 residues. This alignment does reflect some similarity of secondary structure (primarily helices),

but not similarity of tertiary structure (rmsd = 9.8 Å). Clearly, adjustment of gap penalty parameters or a more refined definition of the threshold for structural homology (Sander & Schneider, 1991) could move this particular case out of the gray zone. However, the general point remains that a set of structurally unique proteins cannot be defined in terms of sequence criteria alone as long as the problem of protein structure prediction from sequence is not solved.

Extensions

Variants of the algorithms can easily be developed. For example, the algorithms could be used to generate many nonoverlapping sets of proteins, by defining the output list of a run as the exclusion list of subsequent runs. Such sets may be useful in testing the stability of statistical procedures. For algorithm 1, the test list could be ordered according to any desired property, e.g., species origin, so that if possible all selected proteins come from a limited set of species.

Looking beyond protein structures, the algorithms are sufficiently general so that they can be applied to any data base of entities for which a similarity relationship and a threshold of similarity can be meaningfully defined. For example, one could take the data base of protein sequences, currently at about 30,000 proteins, and extract a nonredundant set of proteins in similar fashion. Note, however, that algorithm 2 would require the calculation of 900 million pair relationships. Algorithms that exploit hash tables probably can deal with this problem in finite time. In the much smaller data base of protein 3D coordinates one could extract a representative set of folding units by defining an appropriate similarity relationship between protein 3D structures, exploiting perhaps the fast algorithms for detection of similar 3D substructures (Orengo & Taylor, 1990; Vriend & Sander, 1991).

Future development

The Protein Data Bank is continuing to grow, and it does so at a fast rate. The list of Kabsch and Sander (1983) had 10,925 residues in 62 proteins of which no two proteins had more than 50% identical residues. In 1991, that number increased more than threefold to about 36,000 residues in 190 chains (Fig. 4) (algorithm 2). Statistical analyses of sequence–structure relations can become more reliable as a result. However, a data base increase to, say, 1,000 nonredundant chains is many years away, assuming the use of current technology and current levels of funding.

Here we have restricted ourselves to an application useful in protein structure research by producing lists that are maximally dispersed in sequence space, yet contain at least one representative of each structural family. The lists of representative protein chains reported here, with under 30,000 residues, will soon be out of date. Plans of an ongoing project are to supply updated lists to the scientific community as more solved protein structures become available.

Methods

The algorithms are now described in more detail. Note that the problem and algorithm could be neatly described in the language of graph theory (a protein is a vertex; two vertices are connected by an edge if the two proteins are similar; the matrix of all pair relationships is the adjacency matrix; the problem is to find the largest subgraph that has no edges; and so on [e.g., Sedgewick, 1983]). However, for ease of communication we choose here not to use the language of graph theory.

Algorithm 1: Select until done

The first algorithm proceeds by simultaneous processing of three lists of protein identifiers, the test list of all candidate proteins (or protein chains), the skip list, and the select list. The test list can be sorted according to user-defined criteria, such as resolution (for proteins of known 3D structure), so that certain types of proteins have a higher probability of being selected. The skip list contains proteins that are similar to a previously processed protein from the test list and may also contain a priori unwanted proteins. The select list (initially empty) contains proteins chosen as part of the nonredundant data set.

In detail, the three lists are processed as follows: (1) read one protein identifier from the test list and check if this protein is a member of the skip list; if so, process the next protein in the test list, i.e., repeat step 1; otherwise (2) check if the protein satisfies user-specified requirements, such as minimum sequence length, maximum number of unknown residues, and the like. If the requirements are satisfied, append the protein to the select list; otherwise, process the next protein in the test list, i.e., repeat step 1; (3) with the selected protein, start a FASTA search (Pearson & Lipman, 1988) against all remaining sequences in the test list; (4) scan the FASTA output file and append to the skip list proteins with a higher similarity than the specified threshold (e.g., five percentage points above the threshold for structural homology corresponding to the length of the FASTA alignment [Sander & Schneider, 1991]). Finally, step 1 is repeated until all proteins in the test list are processed.

Algorithm 2: Remove until done

The second algorithm is computationally more expensive, as it requires a complete matrix of pair relations among all proteins in the candidate list. The goal of the algo-

rithm is to remove, in the smallest number of steps, one protein at a time, together with its pair relations, until the matrix of the remaining proteins contains no more pair relations. This global optimization problem is far from trivial. We solve the problem in practice by a procedure of the "greedy" type: removal of the protein with the largest number of pair relations at a particular step tends to minimize the total number of steps needed to remove all pair relations in the matrix.

For the examples presented here, the matrix is generated by aligning each protein chain with all other protein chains using a dynamic sequence alignment algorithm (Smith & Waterman, 1981), using program Swalign. After application of the threshold for structural homology (Sander & Schneider, 1991), the matrix contains only one bit for each protein pair, 1 if the two proteins are similar (are neighbors, are related) and 0 otherwise.

In detail, the algorithm proceeds as follows. In each iteration step, the protein with the largest number of relations (neighbors) is removed by setting all its pair relation bits to zero. If the largest number of pair relations is shared by more than one protein, the choice of protein to be removed is made arbitrarily, using a random number generator. The algorithm terminates when all pair relations in the matrix have been set to zero, i.e., all remaining proteins are mutually dissimilar. The proteins remaining are considered as the selected set. In a final pass over all removed proteins, a protein is reinstated (added to the selected set) if it has no neighbors in the selected set. In practice, this final pass rarely increases the size of the selected set.

In principle, this algorithm can be reformulated so as to guarantee the global optimum of the largest number of unrelated chains, namely by a complete tree search to arbitrary depth. However, execution times of such an algorithm would be prohibitive. This algorithm, as the first one, can also be used to optimize a particular property, such as crystallographic resolution, in the following sense. By initially removing all proteins with, say, a resolution less than a certain cutoff, the algorithm then aims at generating the largest list of proteins with a resolution better than the given cutoff.

Choice of cutoff parameter

The choice of cutoff parameter for sequence similarity between two proteins depends on the purpose of the list. Here, we wanted to have the protein representatives spaced as widely as possible in sequence space (minimum redundancy), which requires the strictest possible cutoff, yet not miss any structurally unique family (maximum coverage), which puts a lower bound on the cutoff. The lower bound was determined by raising the cutoff parameter from 25% (for length 80, higher percentages for shorter sequences [Sander & Schneider, 1991]) until all structurally unique protein families are represented in the list. With the current data base and our assessment of what constitutes unique folds, the cutoff in sequence similarity was set at plus five percentage points above the length-dependent threshold of structural homology, i.e., roughly at the upper edge of the "twilight zone" (Doolittle, 1986). One may argue that this level is too permissive. However, even if the cutoff in sequence similarity is set very low at, say, 20% identical residues, the resulting list would still contain pairs of proteins that have very low sequence similarity yet are identical in basic fold. So, raising the cutoff until each structural protein family is just represented is a reasonable objective criterion for a representative list based on sequence similarity.

List distribution via electronic mail

The lists of selected proteins can be obtained from the EMBL file server by electronic mail. There is one file per list. File names are, e.g., pdb_select_56_cut30.pid, where pdb stands for Protein Data Bank, 56 refers to the PDB release number, cut30 refers to the cutoff in sequence identity, and pid stands for protein identifiers. The following mail message sent to NETSERV@EMBL-Heidelberg.DE should result in the above list being sent by return email: send proteindata:pdb_select_56_cut30.pid. To obtain general information send the message: help proteindata. Files for release 58 of the PDB are currently available. These are provided on the Diskette Appendix (see \SUPLEMNT\Hobohm.doc for listing).

Acknowledgments

We value highly the willingness of many crystallographers and NMR spectroscopists to deposit their structures soon after they have been determined. We are grateful to Martin Vingron for alignment software code used in algorithm 2, to Peter Rice for maintaining the Protein Data Bank and related files, to Roy Omond and Rainer Fuchs for implementing and maintaining the network file server, and to the members of the Protein Design Group for valuable discussions.

References

Bairoch, A. & Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucleic 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.

Doolittle, R.F. (1986). *Of Urfs and Orfs: A Primer on How to Analyze Derived Amino Acid Sequences*. University Science Books, Mill Valley, California.

Heringa, J. & Argos, P. (1991). Side chain clusters in protein structures and their role in protein folding. *J. Mol. Biol. 220*, 151–171.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure. Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers 22*, 2577–2637.

Niefind, K. & Schomburg, D. (1991). Amino acid similarity coefficients for protein modelling and sequence alignment derived from main-chain folding angles. *J. Mol. Biol. 219*, 481–497.

Orengo, C.A. & Taylor, W.R. (1990). A rapid method of protein structure alignment. *J. Theor. Biol. 147*, 517–551.

Pearson, W.R. & Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA 85*, 2444–2448.
Rooman, J.M. & Wodak, S.J. (1988). Identification of predictive sequence motifs limited by protein structure data base size. *Science 335*, 45–49.
Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins 9*, 56–68.
Sedgewick, R. (1983). *Algorithms*. Addison Wesley, Reading, Massachusetts.
Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular subsequences. *J. Mol. Biol. 147*, 195–197.
Taylor, W.R. & Orengo, C.A. (1989). Protein structure alignment. *J. Mol. Biol. 208*, 1–22.
Unger, R., Harel, D., Wherland, S., & Sussman, J.L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins 5*, 355–373.
Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins 11*, 52–58.