Protein Science (1994), *3*:482-492. Cambridge University Press. Printed in the USA.

Modular arrangement of proteins as inferred from analysis of homology

ERIK L.L. SONNHAMMER[1,3] AND DANIEL KAHN[2]

[1] Biophysics and [2] Division of Molecular Biology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

(RECEIVED October 12, 1993: ACCEPTED January 6, 1994)

Abstract

The structure of many proteins consists of a combination of discrete modules that have been shuffled during evolution. Such modules can frequently be recognized from the analysis of homology. Here we present a systematic analysis of the modular organization of all sequenced proteins. To achieve this we have developed an automatic method to identify protein domains from sequence comparisons. Homologous domains can then be clustered into consistent families. The method was applied to all 21,098 nonfragment protein sequences in SWISS-PROT 21.0, which was automatically reorganized into a comprehensive protein domain database, ProDom. We have constructed multiple sequence alignments for each domain family in ProDom, from which consensus sequences were generated. These nonredundant domain consensuses are useful for fast homology searches. Domain organization in ProDom is exemplified for proteins of the phosphoenolpyruvate:sugar phosphotransferase system (PEP:PTS) and for bacterial 2-component regulators. We provide 2 examples of previously unrecognized domain arrangements discovered with the help of ProDom.

Keywords: domain database; domain families; protein domains; protein evolution; protein homology

Many proteins consist of a number of separately evolved, independent structural units, called modules or domains (Baron et al., 1991; Patthy, 1991; Doolittle, 1992; Miklos & Campbell, 1992). The great diversity of protein functions is partly due to the vast number of possibilities to arrange a finite number of these basic units in different combinations (Dorit et al., 1990; Chothia, 1992). The degree of modularity varies from strictly single-module proteins, like small globular proteins, to complex multidomain proteins, like the structural proteins in the extracellular matrix (Bork, 1991). Although the functional and mechanistic interplay between different domains in proteins is diverse, all domains can be regarded as structurally autonomous units and detected as such. The structure of many proteins can thus be understood as a combinatorial arrangement of autonomously structured domains. The combinatorial nature of proteins is a major difficulty when one attempts to classify protein sequences on the basis of homology: homology between 2 proteins implies that they share at least 1 homologous domain, not necessarily that their other domains are homologous. It is therefore a prerequisite to resolve the combinatorial arrangement of protein domains before attempting any protein classification.

Reprint requests to: Daniel Kahn, Laboratoire de Biologie Moléculaire des Relations Plantes-Microorganismes, INRA/CNRS, BP27, F-31326 Castanet-Tolosan Cedex, France; e-mail: dkahn@toulouse.inra.fr.

[3] Present address: Laboratory of Molecular Biology, MRC, Cambridge, England.

A domain that has evolved from one protein to another can often be recognized because of sequence similarity. It is now classical to construct domain families from the manual analysis of sequence similarities (e.g., Bork, 1989; Henikoff et al., 1990; Bengio & Pouliot, 1990), but the scope is usually limited to a few families due to the prohibitively large amount of data to be processed for analyzing all families systematically. Available collections of domains, such as ExCell (Bork, 1991), PROSITE (Bairoch, 1992), and SBASE (Pongor et al., 1993), contain detailed information about each entry. However, because they are hand-assembled on the basis of specialized knowledge, they are not comprehensive. It is therefore desirable to construct such families in a systematic and comprehensive fashion. Such a classification of protein sequences will allow management of the redundancy of protein sequence databases resulting from the occurrence of numerous homologous proteins and bring additional structure to the primary sequence databases. It will thus facilitate protein homology analysis. It could also form the basis for systematic molecular phylogenies.

Here we present a method to find protein domain families automatically on the basis of homology. We started from an all-versus-all comparison of the SWISS-PROT database and exploited this information to construct homologous segment sets. A program was designed to detect domain shuffling, indicative of domain boundaries. This allowed us to resolve the combinatorial arrangement of shuffled proteins, resulting in homogeneous domain families. These constitute the basis for the

ProDom protein domain database, in which each domain family is represented by a multiple sequence alignment and a consensus sequence. The ProDom database appears to be a useful tool for resolving the domain structure of new proteins as well as for the analysis of homology.

Results

The DOMAINER program

To determine the domain organization of proteins we developed the DOMAINER program, which is summarized in the flow diagram of Figure 1. Initially the method was tested and validated on a small set of "2-component" bacterial regulators known to exhibit a combinatorial domain arrangement (see Henikoff et al., 1990; reviewed by Albright et al., 1989; Parkinson & Kofoid, 1992). It was then applied to the entire SWISS-PROT database. Our method consists of the following phases: (1) all-versus-all comparison of the database to generate homologous segment pairs (HSPs); (2) construction of homologous segment sets (HSSs) from HSPs; (3) detection and breakage of domain boundaries; (4) generation of a multiple sequence alignment and a consensus sequence for each domain family.

Phase I: Compare all sequences in the database to generate HSPs

What we required from this phase was (1) that the significance of each alignment could be estimated numerically to avoid human evaluation, and (2) that several significant alignments could be generated between 2 sequences. The fast homology search program BLASTP (Altschul et al., 1990) fulfills these requirements; with a small modification this program can also be used to generate internal alignments between 2 different regions of 1 sequence, which is relevant for proteins with repeats. The significance estimate of BLASTP applies to nongapped alignments and is based on a statistical model applicable to random sequences (Karlin & Altschul, 1990). An alignment was considered to reflect genuine homology only if it was sufficiently improbable in this model. Thus we discarded any alignment expected to occur with a probability higher than 10^{-6} when comparing each sequence with the SWISS-PROT database. Because this operation was reiterated 21,098 times (the number of nonfragmentary sequences in SWISS-PROT 21.0), occurrence of a spurious alignment was expected with a probability of only 0.02.

In order to express homology and sequence information for all sequences in the database, we produced pairwise alignments in 3 ways: (1) internal repeats, (2) normal alignments between different sequences, and (3) entire alignments of each sequence with itself. The last type is somewhat artificial but allowed us to keep track of each sequence, even in the absence of homology. These alignments defined HSPs that were uniquely characterized by 2 protein identity codes, 2 starting, and 2 end positions of the ungapped alignment. An HSP thus has the form:

⟨ID protein 1⟩ ⟨start position⟩ ⟨end position⟩

⟨ID protein 2⟩ ⟨start position⟩ ⟨end position⟩

The protein ID can be protein name or sequence accession number; we used SWISS-PROT protein names, which are more easy to read, but encoded them by integers for higher performance.

Phase II: Overlapping HSPs are used to build HSSs

The sequence relationships defined by HSPs were exploited to cluster homologous protein subsequences into HSSs. An HSS has the form:

⟨ID protein 1⟩ ⟨start position⟩ ⟨end position⟩

⟨ID protein 2⟩ ⟨start position⟩ ⟨end position⟩

⟨ID protein 3⟩ ⟨start position⟩ ⟨end position⟩

⋮

⟨ID protein n⟩ ⟨start position⟩ ⟨end position⟩

where the lengths of all sequence elements are identical. A given protein may be represented several times in the same HSS if it contains internal repeats. Note also that an HSP is a particular case of an HSS with only 2 members.

Each HSP was compared in turn with all other HSPs and previously constructed HSSs. If 2 HSPs or 2 HSSs had at least 1 protein in common, and the sequence segments overlapped, the overlapping regions (which are therefore homologous) were merged into a new HSS. Nonoverlapping segments were put into other HSSs. Sequence positional information between HSSs was kept in links, which indicate that both HSSs are adjacent and which HSS is N-terminal relative to the other.

Merging of 2 HSPs or HSSs was not performed in case their overlap was too short, below a MINOVERLAP value, which is a program parameter. This parameter governs the resolution of the method. Too high a value for MINOVERLAP forbids the generation of small domains. Too small a value made the process very sensitive to short, out-of-frame extensions of alignments, which occur at gaps. A good compromise was MINOVERLAP = 10, which was used throughout the process.

The result of Phase II was a large number of HSSs with links, forming graphs of HSSs. Different domains may initially cluster in a same HSS graph because of multidomain proteins. It was the task of Phase III to detect such situations and split HSS graphs at the links corresponding to domain boundaries.

Phase III: Detection of domain boundaries and splitting of multidomain graphs into single domain graphs

We recognized 3 different situations indicative of domain boundaries.

1. Real N- or C-termini of polypeptide chains. N- and C-termini of real proteins necessarily correspond to domain boundaries. Therefore HSS graphs were broken at the corresponding links. In order to eliminate false termini, we excluded all proteins in SWISS-PROT that were annotated as fragments in the description field.

2. Shuffled domains. Shuffled domains can be detected. For instance, if 4 domains A, B, C, and D occur in different combinations as AB, AC, and DB, this provides evidence that domains A and B are autonomous. To detect shuffling, each link between any 2 HSSs was tested for "mutually exclusive" sets of proteins on either side: the test was positive when both HSSs contained at least 1 protein missing in the other HSS. This is il-

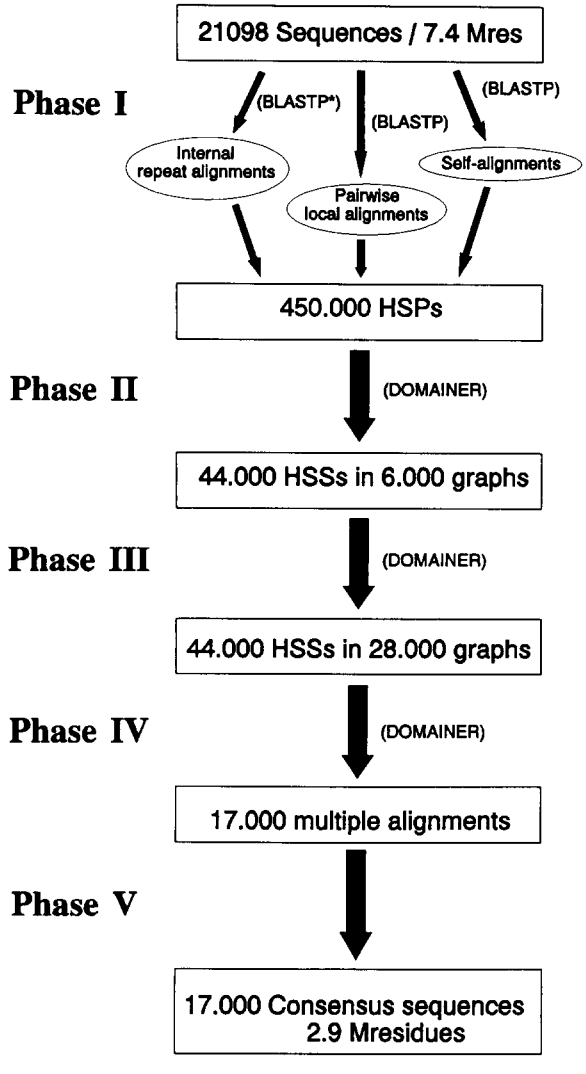


Fig. 1. Flow diagram of the DOMAINER program. As input, all protein sequences in SWISS-PROT 21.0 that were not annotated as fragments were used. The process involves 2 programs, BLASTP for Phase I and DOMAINER for all other phases. Phase I: All-versus-all comparison of SWISS-PROT. Phase II: Clustering of homologous segment pairs (HSPs) into homologous segment sets (HSSs). Phase III: Detection of domain boundaries and cleavage of HSS graphs at these boundaries. Phase IV: Generation of multiple alignments for all domains. Phase V: Construction of consensus sequences for all domains. * Internal repeats are normally not reported by BLASTP; we have enabled this with a small modification of the BLASTP routine ovlap_a. For efficiency, BLASTP was run with 1 sequence as both query and database, with the Z parameter set to the database size to equalize the significance score with the other BLASTP matches.

lustrated in Figure 2 for a few proteins of the bacterial sugar phosphotransferase family, which exhibit extensive domain shuffling (Saier & Reizer, 1992). The graph of HSSs corresponding to these proteins was broken at 3 locations (large arrows). All 3 correspond to N-termini (arrows 1 and 2) or C-termini (arrows 2 and 3) of real sequences. In addition, arrows 2 and 3 correspond to shuffling events; mutual exclusion was indeed verified between the third and fourth HSSs from the left of the graph because they lack the PTHP_ECOLI and PT3F_SALTY sequences, respectively (arrow 2); mutual exclusion was also verified between the fifth and the sixth HSSs, which lack the PT1_ECOLI and PTHP_ECOLI sequences, respectively (arrow 3). In this particular example, different criteria concurred to define the domain boundaries. It occasionally happened that mutual exclusion was verified between 2 HSSs in the absence of genuine shuffling. This occurred in poorly conserved regions, and in such cases the HSS graph showed a complex pattern of diverging and converging branches. To prevent splitting of the graph in such instances, mutual exclusion was validated only in the absence of reconvergence of the diverging branches.

3. Repeat units. If a domain occurs more than once in a protein, the HSS graph will contain a cycle. Cyclic graphs can also arise from "circular permutation," e.g., 2 domains A and B that occur in combinations AB and BA. Because A is C-terminal to B at the same time as B is C-terminal to A, a cycle will result. In order to treat cyclic graphs, we cut and linearized them.

If a junction between 2 or more HSSs satisfied 1 of the 3 criteria above, the graph was split up into 2 disconnected graphs on either side of the junction.

Phase IV: Construction of multiple sequence alignments from HSS graphs

After breakage, each graph of linked HSSs was presumed to correspond to a domain, for which a multiple alignment was generated. There are 2 fundamentally different ways of constructing such an alignment. The first method consists of finding the most representative linear path through the graph and concatenating HSSs in the path (the most representative path was defined as the path of HSSs with maximal number of sequences). Each HSS contained already aligned sequence segments, and a multiple alignment was generated by assembling these. The second method uses hierarchical clustering and dynamic programming (Feng & Doolittle, 1987; Corpet, 1988; Higgins, 1992). Here, for each protein represented in the domain family, the relevant protein segments are aligned. However, available programs for performing multiple sequence alignments, such as Multalin (Corpet, 1988) and Clustalv (Higgins, 1992), can yield incorrect multiple alignments when parameters are not optimized for a particular set of sequences (Argos & Vingron, 1990). It therefore appeared more effective to use the first method.

If a domain boundary was not recognized in Phase III, domains will appear as long divergent branches in the graph. In this case, the most representative path was used to build the initial multiple alignment, and the other branches of the graph were split off. The latter branches, albeit not always perfect domains, were then treated recursively as new graphs until all parts of the graph had been processed.

Because the splitting up of domains may result in large numbers of short linker regions that do not correspond to genuine domains, we eliminated all multiple alignments spanning less than 20 amino acids. This was the only sequence data that was discarded during all phases and should be seen as a mere "trimming" of the domain boundaries.

Consensus sequences were then generated from each multiple alignment as follows. For each position in the multiple alignment, a score for matching each of the 20 amino acids was calculated by summing the coefficients of the PAM120 scoring matrix (Altschul, 1991) over all amino acids occurring at this po-

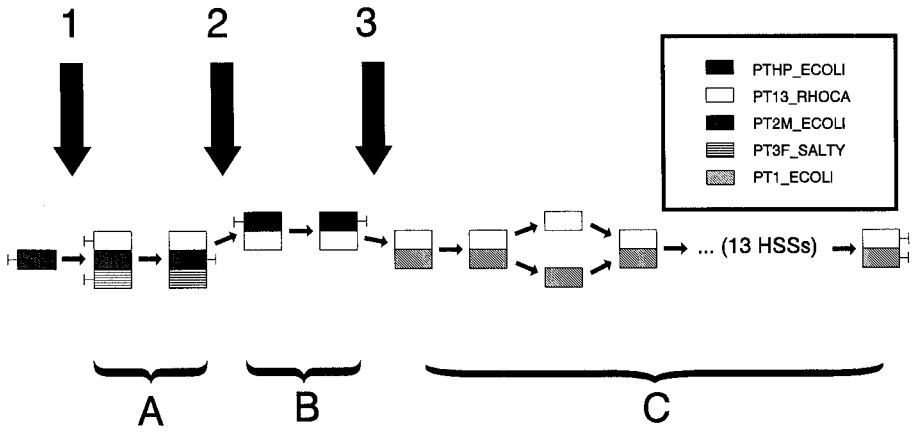


Fig. 2. An HSS graph. This example shows the topology of a graph of HSSs for bacterial PTS proteins after Phase II. Each HSS is represented by a box containing several protein segments symbolized by different hashings. Links are symbolized by small arrows. Sequence termini are symbolized by horizontal Ts. Not all known homologues are shown. This graph will be split at the positions marked by the 3 large arrows during Phase III. Shuffling of domains A and B, and of domains B and C, is apparent on either side of arrow 2 and arrow 3, respectively.

sition in the alignment; the best scoring amino acid was then included at this position in the consensus sequence.

Implementation and performance of the method

DOMAINER was written in the C programming language (Kernighan & Ritchie, 1988) and was run on UNIX machines from SUN and Silicon Graphics. For Phase I we used BLASTP version 1.2.2 and the accompanying PAM 120 scoring matrix (Altschul et al., 1990).

About half the total computing time was spent on Phase I, where BLASTP was run 21,098 times. This was done with a very low level of parallelization, only a factor of 3. If massive parallelization was used, this phase could become much faster.

The performance of Phase II is shown in Figure 3. The redundancy of SWISS-PROT resulted in plateaus during the construction of HSSs. In these plateaus, HSPs fully homologous to existing HSSs were included without creating new HSSs. The big plateau in the middle corresponded to 622 homologous globins, which generated some 100,000 significant HSPs. At the end of the chart, the 21,098 self-alignments were introduced, which were relatively nonredundant (3,810 sequences had no similarity to other sequences at the 10^{-6} threshold) and thus caused fast growth of the number of HSSs. At the end of the process performance dropped dramatically because of memory limitation (16 Mbytes).

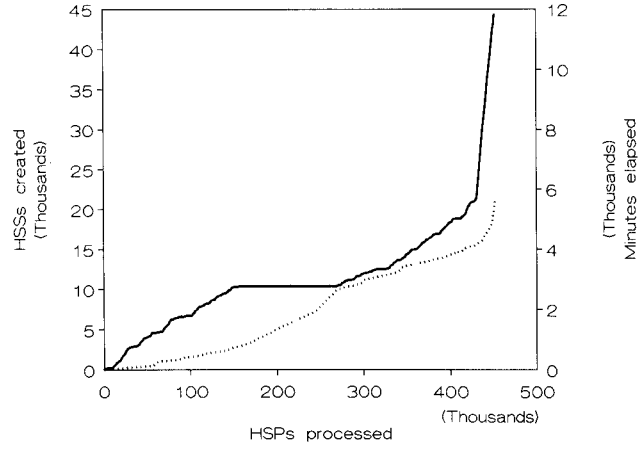


Fig. 3. Clustering HSPs into HSSs. The curves show the number of HSSs (solid line) and time consumption (dotted line) as a function of processed HSPs during Phase II.

Phase III served to detect domain boundaries and to split off the graphs into smaller graphs representing separate domains. Although this phase is complex because multiple criteria are used, the total time spent was only a few hours.

Construction of multiple alignments in Phase IV would be CPU intensive using standard dynamic programming algorithms. Using the fast method described above allowed this phase to be completed in only a few minutes.

The ProDom database

The DOMAINER program, applied to SWISS-PROT 21.0, resulted in the ProDom 21.0 database. ProDom will be distributed and maintained on a regular basis. The complete database, including exploitation programs described below, can be obtained by anonymous FTP from cele.mrc-lmb.cam.ac.uk, in the pub/prodom directory.

Atlas of protein domains generated from SWISS-PROT

The domain families of ProDom were stored as multiple sequence alignments as shown by the examples in Figure 4. These examples correspond to typical modular domains found in bacterial proteins: (1) the HPr domain of the PEP:sugar phosphotransferase system (Wu et al., 1990; Saier & Reizer, 1992) (Fig. 4A) and (2) the phosphorylated regulatory domain of bacterial "2-component" regulators (Albright et al., 1989) (Fig. 4B).

ProDom 21.0 contains 17,162 entries, of which only 5,765 have more than 1 member and can be considered as true families. Linker regions between domains were treated the same way as domains that made the number of families to be overestimated. The average span of a domain in ProDom is 170 residues. The more stringent the cutoff in Phase I, the more families are generated. Because we used a rather stringent threshold to avoid spurious alignments, distantly related domains end up in different families.

As shown in Figure 5, almost half of the proteins in SWISS-PROT contain only 1 ProDom domain. The number of true domains per protein can be overestimated in some cases because linker regions and highly variable termini are treated as sepa-

Fig. 4. Examples of ProDom domains. The figure shows multiple alignments and consensus sequences for 2 ProDom domains. **A:** The HPr domain in the phosphoenolpyruvate:sugar phosphotransferase system. Seven instances were found in SWISS-PROT 21.0. **B:** Phosphorylated domain of bacterial 2-component regulators.

rate domains in ProDom. For example, the SWISS-PROT entry DCTD_RHIME in Figure 6 is split up into 5 ProDom domains, although in principle it has only 3 distinct domains (families 3094, 3098, and 3113). Conversely, because ProDom domains are based on sequence comparisons, some proteins will not be resolved into domains because of lack of homology or shuffling in current sequences.

Using ProDom

There are several ways to access the data in ProDom. One important use of ProDom is to determine the domain organization of a protein. If the protein sequence is in the SWISS-PROT 21.0 database, the answer is already present in ProDom. To this end we created the program ASKDOM, which focuses on a protein and returns the appropriate domain families. Some sample queries are shown in Figure 6.

If the query sequence is not present in SWISS-PROT 21.0, or one is looking for weaker relationships, it is also possible to do a similarity search with ProDom. To this end we stored a consensus sequence of each multiple alignment in a separate file, formatted for database searching with BLASTP (Altschul et al., 1990). The results of a sample search session of ProDom

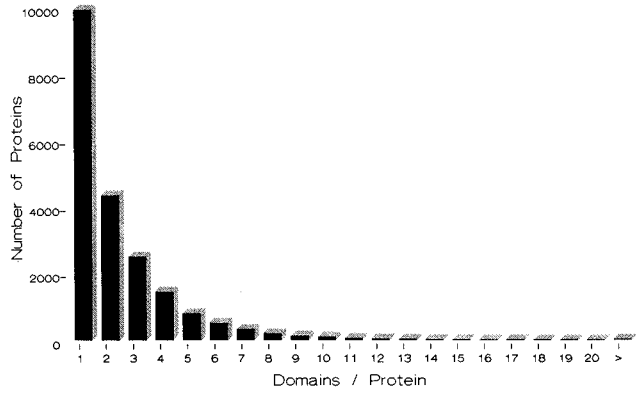


Fig. 5. Many proteins have a multidomain structure that can be revealed by sequence analysis. Histogram of the number of ProDom domains in sequences from SWISS-PROT 21.0.

is shown in Figure 7, together with the results of a search of SWISS-PROT, for comparison. With ProDom, one significant match was reported for each domain in the query. If families are large, the high score list when searching SWISS-PROT becomes prohibitively long, whereas the list obtained when searching ProDom remains short and manageable. Thus ProDom provides the investigator with the suggested domain organization of a protein. To draw similar conclusions from SWISS-PROT takes a trained sequence analyst a substantial amount of time. Note also that sequence clustering in ProDom resulted in a 2.5-fold compression of the database because the ProDom 21.0 consensus file contained 2,862,279 residues versus 7,362,492 residues in SWISS-PROT 21.0. This compression factor accelerated ProDom homology searches accordingly.

Analysis of the domain structure of proteins in the phosphoenolpyruvate:sugar phosphotransferase system and bacterial 2-component regulatory systems

Two examples of protein families with extensive shuffling are illustrated in Figure 8. Bacterial phosphoenolpyruvate:sugar phosphotransferase systems (PEP:PTS) catalyze the concomitant transport and phosphorylation of a wide variety of sugar substrates. Sequence analysis revealed that PTS proteins consist of modular combinations of a few autonomous domains (Saier & Reizer, 1992). These domains appear readily in ProDom, in excellent agreement with previous analyses (Saier & Reizer, 1992) (Fig. 8A).

Two-component regulators are signal transduction devices widely spread among eubacteria (Albright et al., 1989; Parkinson & Kofoid, 1992). They are modular proteins consisting of a conserved phosphorylated regulatory domain and an activator domain, usually involved in transcription activation. The 2 domains are genetically and/or biochemically separable (see for example Simms et al., 1985; Kahn & Ditta, 1991; Huala et al., 1992). Figure 8B shows how the corresponding ProDom domains are combined in a few regulators. We note a good match between this organization into ProDom domains and the known domain organization of these proteins (Henikoff et al., 1990).

We proceeded to investigate the domain structures of newly sequenced 2-component regulators as found in the cumulative update of Genbank archived at the National Center for Biotechnology Information (Bethesda). These sequences were systematically compared with ProDom using the BLASTP program (Altschul et al., 1990). Most regulators exhibited a conventional domain arrangement that had been found previously. However, 2 proteins showed a novel domain arrangement that had not been described (Fig. 9).

The first unusual regulator is the RcaC protein involved in chromatic adaptation in the cyanobacterium *Fremyella diplosiphon* (Chiang et al., 1992). Its N-terminal sequence was reported to be homologous to the *Bacillus subtilis* PhoP 2-component regulator. Comparison of the RcaC sequence with ProDom immediately showed that it contains an additional homologous regulatory domain located at its C-terminus (Fig. 9). Such a re-

```
FIG. ASKDOM

% askdom prodom.mul PT13_RHOCA

Protein      Start  End   Domain Family  Family Members
-------      -----  ---   -------------  --------------
PT13_RHOCA   237    826   12901          9
PT13_RHOCA   164    236   12908          7
PT13_RHOCA   142    163   12909          1
PT13_RHOCA   17     141   12910          4

% askdom prodom.mul PT2N_ECOLI

Protein      Start  End   Domain Family  Family Members
-------      -----  ---   -------------  --------------
PT2N_ECOLI   503    621   11468          8
PT2N_ECOLI   453    488   11469          1
PT2N_ECOLI   413    452   11470          10
PT2N_ECOLI   1      412   11477          3

% askdom prodom.mul PT2S_STRMU

Protein      Start  End   Domain Family  Family Members
-------      -----  ---   -------------  --------------
PT2S_STRMU   506    637   11468          8
PT2S_STRMU   15     95    11470          10
PT2S_STRMU   103    471   11471          7
PT2S_STRMU   472    505   11475          1

% askdom prodom.mul FIXJ_RHIME

Protein      Start  End   Domain Family  Family Members
-------      -----  ---   -------------  --------------
FIXJ_RHIME   10     104   3094           44
FIXJ_RHIME   142    186   3119           12

% askdom prodom.mul PHOP_BACSU

Protein      Start  End   Domain Family  Family Members
-------      -----  ---   -------------  --------------
PHOP_BACSU   9      114   3094           44
PHOP_BACSU   134    210   3123           10
PHOP_BACSU   211    241   3127           1

% askdom prodom.mul DCTD_RHIME

Protein      Start  End   Domain Family  Family Members
-------      -----  ---   -------------  --------------
DCTD_RHIME   11     105   3094           44
DCTD_RHIME   129    150   3096           8
DCTD_RHIME   182    378   3098           25
DCTD_RHIME   381    448   3113           2
DCTD_RHIME   148    168   3137           1
```

Fig. 6. ProDom domain organization of a few selected proteins. The ASKDOM program was applied to three PTS proteins and three 2-component regulators, showing their domain organization.

A

Database: prodom
17,613 sequences; 2,912,708 total residues.

Entry	Segment in query	Score	P(N)	N
12901	237-826	743	3.8e-181	4
12910	17-141	222	1.5e-43	2
12908	164-235	173	2.5e-17	1
14928		65	7.6e-14	6

B

Database: swiss21
23,742 sequences; 7,866,594 total residues.

		Score	P(N)	N
PT13_RHOCA	MULTIPHOSPHORYL TRANSFER PROTEIN (MTP) (CONTAI...	4095	0.0	1
PT1_ECOLI	PHOSPHOENOLPYRUVATE-PROTEIN PHOSPHOTRANSFERASE...	549	9.1e-72	1
PT1_STACA	PHOSPHOENOLPYRUVATE-PROTEIN PHOSPHOTRANSFERASE...	547	1.8e-71	1
PT1_SALTY	PHOSPHOENOLPYRUVATE-PROTEIN PHOSPHOTRANSFERASE...	543	6.9e-71	1
PT1_ALCEU	PHOSPHOENOLPYRUVATE-PROTEIN PHOSPHOTRANSFERASE...	238	3.6e-49	2
PT3F_SALTY	PHOSPHOTRANSFERASE FPR PROTEIN (PSEUDO-HPR).	266	1.8e-30	1
PT2M_ECOLI	PHOSPHOTRANSFERASE ENZYME II, MANNITOL-SPECIFI...	143	2.7e-22	2
PT3M_STACA	PHOSPHOTRANSFERASE FACTOR III, MANNITOL-SPECIF...	160	5.3e-15	1
PTHP_ECOLI	PHOSPHOCARRIER PROTEIN HPR (HISTIDINE-CONTAINI...	155	2.8e-14	1
PT3F_ECOLI	PHOSPHOTRANSFERASE FPR PROTEIN (PSEUDO-HPR) (F...	99	4.8e-14	2
PTHP_KLEPN	PHOSPHOCARRIER PROTEIN HPR (HISTIDINE-CONTAINI...	150	1.5e-13	1
PTHP_BACSU	PHOSPHOCARRIER PROTEIN HPR (HISTIDINE-CONTAINI...	134	3.3e-11	1
PTHP_ENTFA	PHOSPHOCARRIER PROTEIN HPR (HISTIDINE-CONTAINI...	134	3.3e-11	1
PPSA_ECOLI	PHOSPHOENOLPYRUVATE SYNTHASE (EC 2.7.9.2) (PYR...	119	1.6e-10	2
PTHP_STAAU	PHOSPHOCARRIER PROTEIN HPR (HISTIDINE-CONTAINI...	126	4.8e-10	1
PODK_FLATR	PYRUVATE,ORTHOPHOSPHATE DIKINASE PRECURSOR (EC...	112	1.2e-09	5
PODK_MAIZE	PYRUVATE,ORTHOPHOSPHATE DIKINASE PRECURSOR (EC...	109	1.4e-07	1
PTHP_ALCEU	PHOSPHOCARRIER PROTEIN HPR (PROTEIN H).	108	2.0e-07	1
PODK_BACSY	PYRUVATE,ORTHOPHOSPHATE DIKINASE (EC 2.7.9.1).	104	7.8e-07	1

Fig. 7. Using ProDom for similarity search. Search results were obtained using BLASTP with query PT13_RHOCA, the multiphosphoryl transfer protein of *R. capsulatus* (Wu et al., 1990). All scores significant at the $P = 10^{-6}$ level are shown. **A:** ProDom search. Match number 4 originates from an Ala-rich region in the query. **B:** SWISS-PROT search.

markable domain arrangement had not been observed previously in 2-component regulators and had been overlooked in the RcaC protein.

The second unusual regulator is the EvgS protein, an *Escherichia coli* homologue of the BvgS sensor/regulator of virulence genes in *Bordetella* species (Utsumi et al., 1992). BvgS is a transmembrane protein containing both a kinase and a regulator domain (Arico et al., 1991). Comparing EvgS with ProDom indicated that there could be a domain permutation in the periplasmic regions of EvgS and BvgS because significant matches followed a permutated order. Closer inspection showed that this resulted from a duplicated periplasmic domain present in both EvgS and BvgS (Fig. 9). This duplicated domain arrangement of the periplasmic region of BvgS had been overlooked previously.

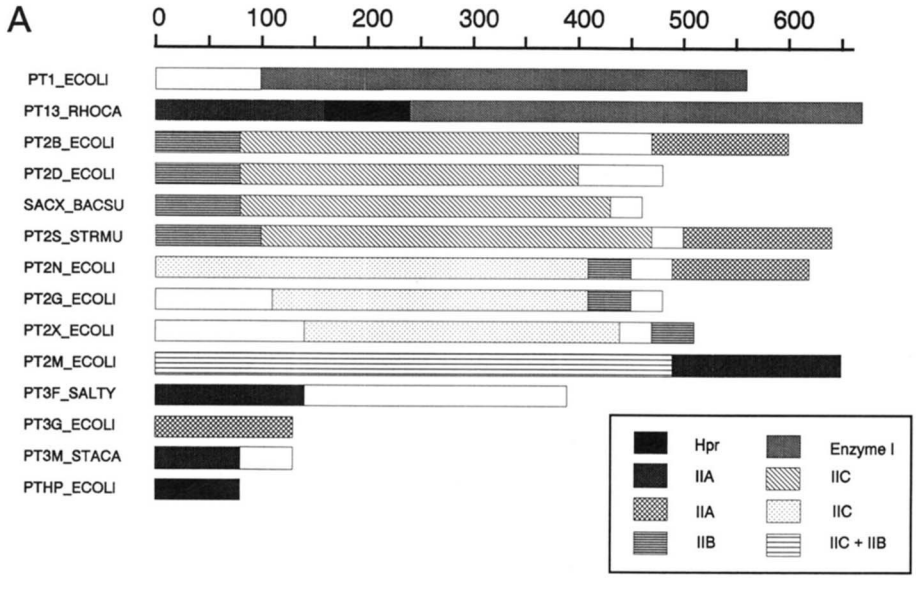
Discussion

In this work we have developed a program to analyze the modular domain organization of proteins in the sequence databases. The procedure was entirely based on the analysis of homology as can be established from sequence comparisons. It relies on the principle that proteins have evolved both by mutational drift and by domain shuffling. Evidence for the latter has been accumulating at an accelerated pace as more protein sequences have been compiled. This principle was often applied on a one-by-one basis, yielding insight into the combinatorial structure of some protein families. However, it had not been applied systematically to all currently available sequences because of the lack of an automated procedure. This we have now achieved with the DOMAINER program, resulting in the ProDom domain database.

The DOMAINER program

The general philosophy of the DOMAINER program parallels the method traditionally used by sequence analysts. It involves the clustering of homologous protein segments and the detection of sequence divergence and shuffling to infer domain boundaries.

Pairwise alignments were generated automatically using the BLASTP program, and filtered according to statistical significance (Karlin & Altschul, 1990). A crucial difference between the manual procedure and DOMAINER at this stage is that the DOMAINER program considers only non-gapped alignments. This was necessary for 3 reasons. First, no good statistical model is available for gapped alignments that would allow DOMAINER to automatically filter out nonsignificant alignments (Karlin et al., 1991). Second, starting from a significant non-gapped HSP, there is no method available to extend this alignment safely with gaps: such an extended gapped alignment frequently intrudes from 1 homologous domain into neighboring domains that are not necessarily homologous, in which case manual editing and biological assessment is required. Third, amino acid correspondences need to be known at each step during the construction of HSSs from HSPs. This is trivial with non-gapped alignments because it is sufficient to archive the positions of the start and end of each of the segments constituting an HSS. It would be considerably more difficult to build HSSs from sequences aligned with gaps. However using non-



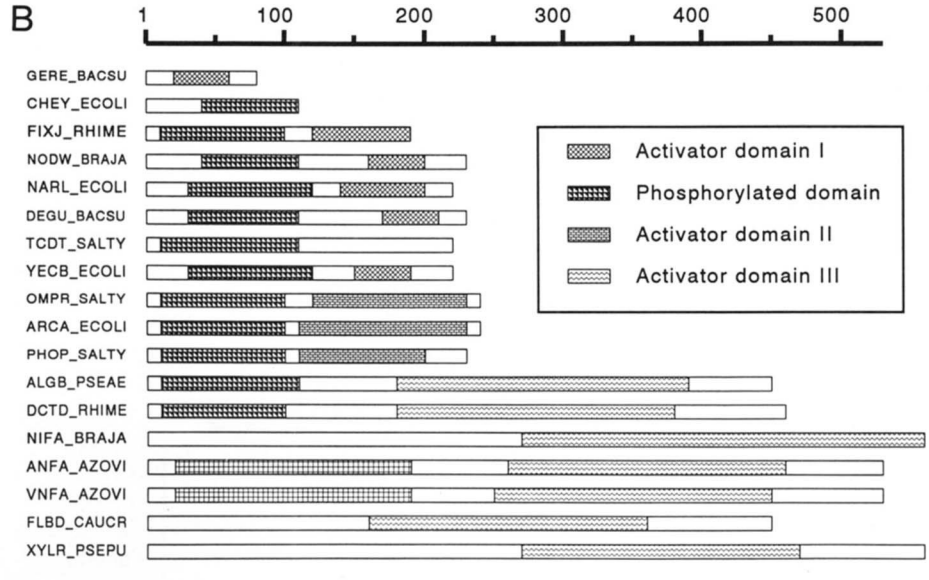


Fig. 8. Modular arrangement of ProDom domains for selected protein families. **A:** PEP:PTS system. **B:** Two-component bacterial regulatory systems. Each ProDom domain is represented by a different hatching. Nomenclature of PEP:PTS domains is from Saier and Reizer (1992). Domain shuffling is apparent for both families. The scale indicates the number of residues in each domain.

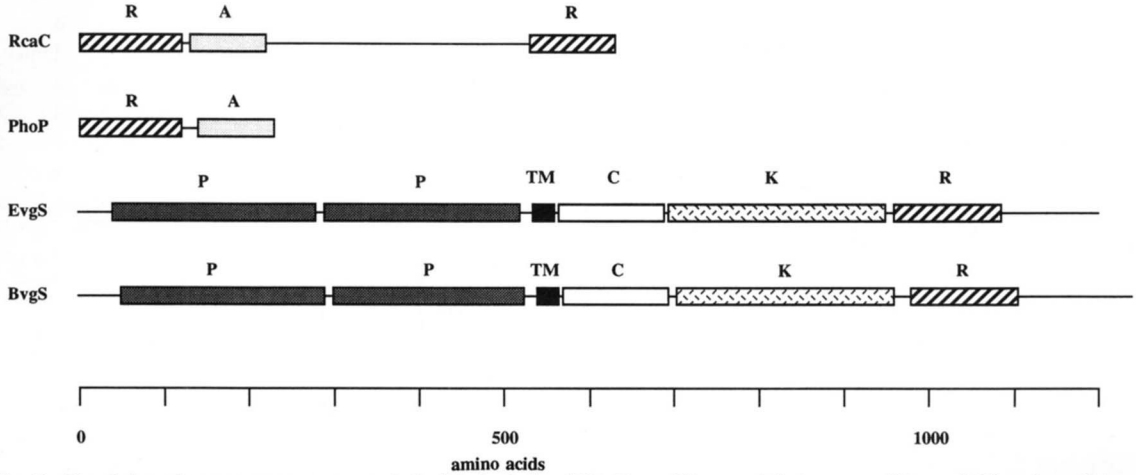


Fig. 9. Novel domain arrangements unraveled with the help of ProDom. The novel features are (1) an additional regulatory domain at the C-terminus of the cyanobacterial RcaC activator and (2) the duplicated periplasmic domain of the BvgS and EvgS proteins. R, Phosphorylated regulatory domain; A, activator domain; P, periplasmic domain; TM, transmembrane sequence; K, kinase domain; C, additional cytoplasmic domain common to BvgS and EvgS.

gapped alignments also had a cost because many more non-gapped HSSs had to be processed than would have been required with gap-aligned HSSs. This made HSS graphs much more complex. Also, with non-gapped alignments, each domain family was represented by a graph usually containing several HSSs, which needed to be processed further in order to generate multiple alignments. If it was possible to handle gap-aligned HSSs, each HSS would correspond to a domain family in a one-to-one relationship. Multiple gapped alignments would then be available as a direct result of HSS construction. This attractive possibility suggests that it is worth directing future efforts toward overcoming the 3 constraints above: (1) to develop a rigorous statistical model of gapped sequence alignments; (2) to develop methods for safely extending them; (3) to design efficient data structures to handle them.

Detection of sequence divergence and shuffling was difficult to encode. For this we derived a heuristic procedure, which used the principle of "mutual exclusion" as an indication of sequence shuffling. This method effectively detected domain boundaries in a number of situations. However it must be emphasized that domain shuffling is not always observed in multidomain proteins (possibly because the putative shuffled proteins remain to be discovered). In those cases the DOMAINER program will not recognize domain boundaries. Also, it must be noted that the determination of domain boundaries by our procedure is only accurate within a margin of 10 amino acids (the value of the MINOVERLAP parameter). This margin may be wider in some unfavorable cases when sequence conservation is poor at the end of domains.

Multiple sequence alignments were generated by pasting together the multiply aligned segments in HSSs. This method is effective and relies on established local alignments. Although all sequences are not aligned throughout their entire lengths, the multiple alignments thus obtained are sufficiently informative that they can be used to generate consensus sequences that appear representative of each domain family. The entire set of domain multiple alignments and consensus sequences, as derived after processing the entire SWISS-PROT database with the DOMAINER program, constitutes the ProDom database.

The ProDom domain database and the protein classification problem

ProDom allows consistent classification of multidomain proteins. If such a classification is attempted without prior separation into domains (Smith & Smith, 1990; Taylor, 1990; van Heel, 1991; Ferran & Ferarra, 1992; Wu et al., 1992), conflicts will arise because otherwise unrelated proteins may share a single homologous domain. It thus appears that it is impossible to classify proteins without resolving their domain arrangement, and that the protein domain is the relevant level for analyzing protein evolution.

A consistent classification of protein sequences also allows management of the ever-increasing redundancy in protein sequence databases. If a given protein (e.g., hemoglobin) has been characterized in many organisms, sequence databases will contain many similar entries, yielding substantial redundancy. This redundancy is not only disadvantageous for database storage and maintenance but also obscures the interpretation of homology searches. This redundancy problem will become ever more serious as large volumes of sequence data are generated by genome sequencing projects, because newly determined sequences will increasingly be combinations of known domains.

With ProDom much of this redundancy is removed. For each domain family, ProDom provides a multiple alignment and a consensus sequence. Other possible legitimate representations include profiles (Gribskov et al., 1987), consensus patterns (Smith & Smith, 1990), or protein blocks (Henikoff & Henikoff, 1991). Using multiple alignments or profiles can in principle make database searching more sensitive (Gribskov et al., 1987). Using consensus sequences makes database searching faster because it results in a compression of sequence information. Thus with ProDom 21.0 we achieved a compression factor of 2.5. In both cases, the output will be not or little redundant (see Fig. 7).

The most obvious usage of ProDom is to determine the domain organization of a query sequence. If the sequence was already in the database, the ASKDOM program provides its organization into ProDom domains within less than a minute. If the sequence is new, a search of ProDom consensus sequences with, for instance, the BLASTP program (Altschul et al., 1990) provides similar information as rapidly. Thus ProDom provides the domain organization of any protein, provided it is homologous to ProDom domains. A good example is shown in Figure 9, which shows the result of domain analysis for 2 recently released protein sequences. The use of ProDom allowed us to detect features in RcaC, EvgS, and BvgS that had been missed previously, probably because of the redundancy problem. Indeed a direct comparison of these sequences with the primary sequence databases yields a considerable amount of partially redundant information, from which it is difficult to extract relevant, nonredundant information. This task is made considerably easier with the help of ProDom.

Another function of the ProDom database is to serve as a repository of clear-cut homologies between known sequences. These homologies are established on the basis of a stringent statistical criterium because each underlying alignment has a chance probability below 10^{-6}. Most of these homologies have been published, but there are instances where homologies may have escaped the attention of some investigators, despite their obvious relevance. In other cases, surprising homologies raise intriguing questions. For instance, the 200 C-terminal amino acids of the POLB_MAIZE protein are unambiguously related to a segment of plasma membrane proton ATPases. This protein is the product of a large open reading frame in the maize *Bs1* element, postulated to be homologous to retro-elements (Johns et al., 1989). We have been unable to confirm the suggested homology of this protein with reverse transcriptase, and thus this sequence needs to be reanalyzed. We expect the release of ProDom to trigger more instances of sequence reevaluation.

An important feature of ProDom is its comprehensiveness. Indeed all nonfragmentary sequences of the SWISS-PROT database were automatically processed. Manually compiled homology databases, such as PROSITE (Bairoch, 1992), contain a wealth of information, but they suffer from lagging inevitably behind the primary sequence databases because they rely on expert analysis of each individual domain family. The formidable nature of the latter task should not be underestimated. From the present study we estimate at 10^4 the order of magnitude of the number of possible protein domain types. A number of these domain types have not yet been sequenced since genomic research reveals that more than half of newly sequenced genes share no detectable homologues in the databases (Bork et al.,

1992). Thus the current scope of available protein sequences is incomplete. In this context, it is to be expected that manually compiled homology databases will lack significant information. Even well-characterized domain families can be occasionally missed in manually compiled databases. For instance, although the phosphorylated domain of 2-component regulators (ProDom domain #3094, Fig. 4) is now extremely well characterized genetically, biochemically, and structurally, and is represented by more than 50 known sequences, it is altogether absent in the latest release of PROSITE (Release 11, October 1993). Another question concerns the relative efficiency of sequence clustering and domain representation in ProDom and PROSITE. As an example, we will consider a large set of RNA-binding domains known as RRMs (Kenan et al., 1991). In the latest version of ProDom (version 24.0), 46 RRMs are clustered in a single domain family. Very divergent RRMs cluster in distinct ProDom families. In PROSITE instead, divergent RRMs are aggregated into a single PROSITE entry (RNP_1). In ProDom the RRM family is represented by a multiple alignment and a consensus sequence spanning 79 amino acids, close to the actual domain span of 80 amino acids (Kenan et al., 1991). In PROSITE instead, the RRM family is characterized by the short 8-amino acid long RNP-1 signature. It should be noted that PROSITE entries do not contain aligned sequences, but only a pattern characteristic of each family. PROSITE also contains short motifs unrelated to homology, such as protein modification sites, etc. The function of the PROSITE database is thus quite different from that of ProDom.

Recently, two other groups independently achieved a systematic classification of protein sequences. In the first such study, Sheridan and Venkataraghavan (1992) clustered homologous protein segments and generated a database of sequence signatures. Some of these signatures are indeed highly informative. However because they derive from non-gapped sequence alignments (akin to HSSs in the present study), these signatures generally do not correspond to entire domains. In the second such study, Harris et al. (1992) designed a clustering algorithm very similar to Phase II of the DOMAINER program. However, these authors did not attempt to infer domain boundaries. Instead their algorithm clusters multidomain proteins into several classes: for instance a protein containing 2 domains A and B would be classified in an AB group, whereas domain A and domain B would also be classified in separate A and B groups. Thus, part of the redundancy remains in their treatment. Finally, it should be noted that Harris et al. (1992) fell short of generating the multiple alignments and consensus sequences that would make their clustering algorithm of general use. The ProDom database, on the other hand, includes such information, and is already operational in our laboratories for routine domain homology analyses. Our long-term objective is now to develop ProDom into an integrated protein database, in which sequences, domains, and 3D structures will point to one another. Because of the combinatorial nature of proteins, it would be difficult to build such a homogeneous set of data directly around the primary sequence databases.

Note added in proof

The ProDom database is now available through anonymous FTP from 2 additional sites: dune.toulouse.inra.fr (192.93.80.7) and bisance.citi2.fr (192.70.98.30), in the pub/prodom directory. In addition, a ProDom electronic mail server has been set up; instructions can be obtained by sending a 1-line HELP message to prodom@toulouse.inra.fr.

Acknowledgments

We are grateful to Huub Stoffers for his input in the initial phase of this project, particularly for defining appropriate data structures for DOMAINER, and for numerous discussions on the art of programming in C. We thank Gilles Kahn and Francis Montagnac for performing the first all-versus-all comparison of SWISS-PROT that we used, Hans Westerhoff and Chris Sander for discussions and helpful suggestions, and Florence Corpet for constructive criticism of the manuscript. This work was supported in part by a computing grant from the Dutch Foundation for National Computing Facilities (NCF SC-234).

References

Albright LM, Huala E, Ausubel FM. 1989. Prokaryotic signal transduction mediated by sensor and regulator protein pairs. *Annu Rev Genet 23*:311–36.

Altschul SF. 1991. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol 219*:555–565.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Argos P, Vingron M. 1990. Sensitivity comparison of protein amino acid sequences. *Methods Enzymol 183*:352–365.

Arico B, Scarlato V, Monack DM, Falkow S, Rappuoli R. 1991. Structural and genetic analysis of the *bvg* locus in *Bordetella* species. *Mol Microbiol 5*:2481–2491.

Bairoch A. 1992. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res 20*:2013–2018.

Baron M, Norman DG, Campbell LD. 1991. Protein modules. *Trends Biochem Sci 16*:13–17.

Bengio Y, Pouliot Y. 1990. Efficient recognition of immunoglobulin domains from amino acid sequences using a neural network. *Comput Appl Biosci 6*:319–324.

Bork P. 1989. Recognition of functional regions in primary structures using a set of property patterns. *FEBS Lett 257*:191–195.

Bork P. 1991. Shuffled domains in extracellular proteins. *FEBS Lett 286*:47–54.

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*:1677–1690.

Chiang GG, Schaefer MR, Grossman AR. 1992. Complementation of a red-light indifferent cyanobacterial mutant. *Proc Natl Acad Sci USA 89*:9415–9419.

Chothia C. 1992. Proteins—1000 families for the molecular biologist. *Nature 357*:543–544.

Corpet F. 1988. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res 16*:10881–10890.

Doolittle RF. 1992. Reconstructing history with amino acid sequences. *Protein Sci 1*:191–200.

Dorit RL, Schoenback L, Gilbert W. 1990. How big is the universe of exons? *Science 250*:1377–1382.

Feng DF, Doolittle RF. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol 25*:351–360.

Ferran EA, Ferrara P. 1992. Clustering proteins into families using artificial neural networks. *Comput Appl Biosci 8*:39–44.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA 84*:4355–4358.

Harris NL, Hunter L, States DJ. 1992. Mega-classification: Discovering motifs in massive datastreams. In: *Proceedings of the Tenth National Conference on Artificial Intelligence*. Menlo Park/Cambridge, Massachusetts: AAAI Press/The MIT Press. pp 837–842.

Henikoff S, Henikoff JG. 1991. Automatic generation of protein blocks for database searching. *Nucleic Acids Res 19*:6565–6572.

Henikoff S, Wallace JC, Brown JP. 1990. Finding protein similarities with nucleotide sequence databases. *Methods Enzymol 183*:111–132.

Higgins DG. 1992. Sequence ordinates: A multivariate analysis approach to analysing large sequence data sets. *Comput Appl Biosci 8*:15–22.

Huala E, Stigter J, Ausubel FM. 1992. The central domain of DctD functions independently to activate transcription. *J Bacteriol 174*:1428–1431.

Johns MA, Babcock MS, Fürstenberg SA, Fürstenberg SM, Freeling M, Simpson RB. 1989. An unusually compact retrotransposon in maize. *Plant Mol Biol 12*:633–642.

Kahn D, Ditta G. 1991. Modular structure of FixJ: Homology of the transcriptional activator domain with the −35 binding domain of sigma factors. *Mol Microbiol 5*:987–997.

Karlin S, Altschul SF. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA 87*:2264–2268.

Karlin S, Bucher P, Brendel V, Altschul SF. 1991. Statistical methods and insights for protein and DNA sequences. *Annu Rev Biophys Chem 20*:175–203.

Kenan DJ, Query CC, Keene JD. 1991. RNA recognition: Towards identifying determinants of specificity. *Trends Biochem Sci 16*:214–220.

Kernighan BW, Ritchie DM. 1988. *The C programming language*. Englewood Cliffs, New Jersey: Prentice-Hall.

Miklos GLG, Campbell HD. 1992. The evolution of protein domains and the organizational complexities of metazoans. *Curr Opin Genet Dev 2*:902-906.

Parkinson JS, Kofoid EC. 1992. Communication modules in bacterial signaling proteins. *Annu Rev Genet 26*:71–112.

Patthy L. 1991. Modular exchange principles in proteins. *Curr Opin Struct Biol 1*:351–361.

Pongor S, Skerl V, Cserzö M, Hatsagi Z, Simon G, Bevilacqua V. 1993. The SBASE domain library: A collection of annotated protein segments. *Protein Eng 6*:391–395.

Saier MH, Reizer J. 1992. Proposed uniform nomenclature for the proteins and protein domains of the bacterial phosphoenolpyruvate:sugar phosphotransferase system. *J Bacteriol 174*:1433-1438.

Sheridan RP, Venkataraghavan R. 1992. A systematic search for protein signature sequences. *Proteins Struct Funct Genet 14*:16–28.

Simms SA, Keane MG, Stock J. 1985. Multiple forms of the CheB methylesterase in bacterial sensing. *J Biol Chem 260*:10161–10168.

Smith RF, Smith TS. 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc Natl Acad Sci USA 87*:118–122.

Taylor WR. 1990. Hierarchical method to align large numbers of biological sequences. *Methods Enzymol 183*:456–474.

Utsumi R, Katayama S, Ikeda M, Igaki S, Nakagawa H, Miwa A, Taniguchi M, Noda M. 1992. Cloning and sequence analysis of the *evgAS* genes involved in signal transduction of *Escherichia coli* K-12. *Nucleic Acids Symp Ser.* pp 149–150.

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

Wu C, Whitson G, McLarty J, Ermongkoncha A, Chang TC. 1992. Protein classification artificial neural system. *Protein Sci 1*:667–677.

Wu LF, Tomich JM, Saier MH. 1990. Structure and evolution of a multidomain multiphosphoryl transfer protein. *J Mol Biol 213*:687–703.