BIOINFORMATICS

Vol. 15 no. 9 1999
Pages 741–748

Associative database of protein sequences

Jens Hanke, Gerrit Lehmann, Peer Bork and Jens G. Reich

Max-Delbrück-Center for Molecular Medicine, Department of Bioinformatics, Robert-Rössle-Straße 10, D-13125 Berlin-Buch, Germany

Received on August 27, 1998; revised and accepted on April 19, 1999

Abstract

Motivation: *We present a new concept that combines data storage and data analysis in genome research, based on an associative network memory. As an illustration, 115 000 conserved regions from over 73 000 published sequences (i.e. from the entire annotated part of the SWISSPROT sequence database) were identified and clustered by a self-organizing network. Similarity and kinship, as well as degree of distance between the conserved protein segments, are visualized as neighborhood relationship on a two-dimensional topographical map.*

Results: *Such a display overcomes the restrictions of linear list processing and allows local and global sequence relationships to be studied visually. Families are memorized as prototype vectors of conserved regions. On a massive parallel machine, clustering and updating of the database take only a few seconds; a rapid analysis of incoming data such as protein sequences or ESTs is carried out on present-day workstations.*

Availability: *Access to the database is available at http://www.bioinf.mdc-berlin.de/unter2.html*

Contact: *(hanke,lehmann,reich)@mdc-berlin.de; bork@embl-heidelberg.de*

Introduction

At present, any novel sequence obtained in the laboratory undergoes extensive database comparison. Resulting matches to existing entries facilitate interpretation and further experimental exploration. However, the current huge increase in database size tends to produce adverse effects. An avalanche of text information, in flat file or table format, has become typical for each query sequence, the basic items of which are sequence strings or identifiers accompanied by coded or plain text information. This makes semantic cross-referencing of information as well as cross-comparison of sequence features very tedious and difficult to automate. Moreover, the intensity of background noise due to high-scoring but biologically irrelevant matches will grow exponentially with the increasing size of current databases. Prohibitively redundant information obtains when databases contain many identical or slightly varying entries (which is expedient for certain other applications). Finally, a match is usually established by pairwise similarity scores without embedding it into a consistent distance metric within the *N*-dimensional Euclidean sequence space.

To overcome these difficulties, we propose here a concept of integrated sequence handling and distance analysis based on pictorial representation that focuses on neighborhood relationships defined by the module substructure of protein sequences.

We wanted to free molecular biologists from monotonous screening of large text collections and allow them to carry out more advanced analyses of significant genome regions. Artificial intelligence methods are very well suited for that type of task. In addition to the actual recognition, which consists of filtering particular characteristics from dense data collections, neural networks are increasingly being used for knowledge-based data management. Their great advantage lies in the conceptual collection of information recognition and information structuring. The collection of primary structure information (on the basis of conserved motifs) and secondary structure information [on the basis of the homology-derived secondary structure protein database (HSSP); Schneider *et al.*, 1997] can be presented very impressively using such a data management strategy. This database is a systematically derived motif database, allowing the classification of the majority of the newly appearing protein sequences into known families.

System and methods

An associative network is a computer program that classifies and stores a large number of information items. As a basis of this network, we used a self-organizing map (SOM; Kohonen, 1996). It consists of a two-dimensional (2D) field of vectors (often called weight vectors or neurons). Individual neurons are ordered in a lattice (map), and the input consists of *n*-dimensional vectors. The dimension *n* of the input space is significantly larger than the dimension of the SOM. The result of the training is a mapping of the *n*-dimensional input space onto a 2D map. Between different neurons, there is a map distance defined (we selected the Euclidean distance). Neighboring vectors (vectors of similarity) should thereby be mapped onto neighboring neurons on the SOM. Neighbourhood is defined by a suitable distance metric, which we obtain by transforming each amino acid in a sequence

segment into a column vector of scores taken from a similarity matrix (like PAM or BLOSUM; Dayhoff *et al*., 1978; Henikoff and Henikoff, 1992). This representation of an amino acid by a vector constitutes a conceptual difference from scoring methods like PROFILE (Gribskov *et al*., 1987, and many others), where only one scalar is assigned to one amino acid in one given position. The coding scheme and the data presentation for the network are described in detail in Hanke and Reich (1996).

The algorithm is structured in a rather simple manner. First, every neuron on the map is assigned a random weight vector, w_{ij}, which has the same dimension as the vectors of the input space. The following scheme will then be carried out for a given number of training steps: a vector x from the input space is chosen at random, and the entire map is searched for the neuron that (we selected the Euclidean distance, D_{ij}) has stored the weight vector next to the input vector.

$$D_{ij} = \sum_{K=1}^{n} (w_{ij}k(t) - xk(t))^2$$

The training function is calculated with this neuron and the weights in a particular area around the excitement center changed accordingly (adapted). During the course of the training process, the strength of the learning and the width of the excitation center are reduced to zero, and the training process ends. The algorithmic details of the training strategy are described in the specialized literature (e.g. Hertz *et al*., 1991; Ritter *et al*., 1992).

This whole principle is termed topology preservation and has beneficial side-effects. The SOM is a productive, capable classification tool. It represents a generalization of the linear principal component analysis used in statistics. Instead of linear main axes or levels, SOM uses non-linear hyperplanes. Their position and orientation are chosen so that every data point from the input space lies as close as possible to a point in the hyperplane, whose position is defined by the weight vectors stored on the 2D map.

Vector quantization

In spite of continuous growth of sequence collections, the number of (super-) families and constituent motifs will probably saturate as the sequence space is not unlimited (Green *et al*., 1993; Gerstein *et al*., 1994). Hence, we expect that the number of motifs occupying our network will likewise converge to a limit.

The SOM allows the representation of many homologous sequences on a weight vector (a type of consensus production) through its projection rules. Average data reduction is ~10 sequence motifs per reference sequence.

Large amounts of data can be effectively processed based on the fundamental idea that every input vector can be approximated through a reference vector (weight vector) of the same dimension.

For that, one must define a smallest possible number, R, of weight vectors (independent of distance function) so that, for every input vector, an approximation vector can be found that sufficiently represents the qualities of the input vector. The areas with minimum distance to the weight vectors correspond to the partition of the input space into Voronoi cells (polygons). The resulting data compression, in the form of a self-constructing mean-reference vector, is referred to as vector quantization (Lynch, 1985).

Results

The associative database as analysis source

In preparation for later protein classification, we chose conserved patterns of amino acids (motifs) as the smallest sequence information units. Such sequence motifs are signatures of protein families, usually corresponding to <300 amino acids in length, and can often be used as tools for the prediction of protein function.

In order to set up the associative database of motifs, we first needed a systematic collection of all known protein patterns. Such patterns are collected in specialized databases such as BLOCKS (Henikoff and Henikoff, 1996), PROSITE (Bairoch *et al*., 1997), PRINTS (Attwood *et al*., 1997), PFAM (Sonnhammer *et al*., 1998), and others, which are routinely updated from the publicly available sequence collections. After the collection and comparison of the complete data set, we decided, in order to avoid too much redundancy between the databases, to fill the associative database only with patterns from the PROSITE database. We also included self-generated patterns chosen by an established pattern recognition method (Hanke *et al*., 1996). The complete method can be describe briefly as follows:

1. Using the pattern recognition method, only informative sequences (conserved regions) are filtered out of the protein database (from SWISSPROT Release 35; Bairoch and Apweiler, 1998).
2. The significant sequence areas (without gaps) are divided into partial patterns.
3. The entire collection of partial ungapped patterns is then classified (see System and methods). Using projection rules, an attempt is made to build a 2D neighborhood map for all partial patterns (Figure 2C; similarities are shown in the form of distances).
4. The complete similarity comparison for all partial patterns is then shown through classes. Every class builds a type of consensus pattern during the learning process (see the previous section on vector quantization). The possible variations, i.e. distantly related or similar patterns, are shown by neighborhood clusters on a 2D map (Figure 2C).

Table 1. Systematic collection and comparison between the Prosite database and our own pattern database

No. of PROSITE patterns	No. of own patterns	Coinciding/overlapping patterns with PROSITE	New patterns	Hits in HSSP
1355	11750	1294	10 456	1265

Table 2. Pattern recognition (receiver domain) in transcriptional regulatory proteins starts from query sequence RESD_BACSU

// ADZ-Pattern signature I

```
TRANSCRIPTIONAL REGULATORY PROTEIN
ID H000513
LN 20
ARCA_ECOLI (50) LVIMDINLPGKNGLLLAREL
TORR_ECOLI (46) LILLDINLPDENGLMLTRAL
VIRG_AGRT5 (46) VVVVDLNLGREDGLEIVRSL
OMPR_ECOLI (45) LMVLDLMLPGEDGLSICRRL
VANR_ENTFC (45) LAILDIMLPGTSGLTICQKI
YC27_PORAE (50) LVVLDLMMPKLDGYGVCQEL
CPXR_ECOLI (45) LLLLDVMMPKKNGIDTLKAL
PETR_RHOCA (47) LIVLDVMMPGEDGLSLTRDL
PHOB_ECOLI (45) LILLDWMLPGGSGIQFIKHL
BAER_ECOLI (48) LILLDLMLPGTDGLTLCREI
AFQ1_STRCO (47) LIVLDVMLPGIDGFEVCRRI
YC27_PORPU (49) LVVLDVMMPKLDGYGVCQEL
PHOB_HAEIN (34) LILLDWMLPGRSGIQFIQYI
YV17_MYCLE (17) IVLLDLMLPGMSGTDVYKQL
RESD_BACSU (45) LILLDLMMPGTDGIEVCRQI
CHVI_AGRTU (45) LAIFDIKMPRMDGMELLRRL
CREB_ECOLI (50) VMILDVGLPDISGFELCRQL
YRKP_BACSU (43) LVILDIMMPGISGIETCQHI
KDPE_ECOLI (31) LIILDLGLPDGDGIEFIRDL
YXDJ_BACSU (45) VVLLDINLPAYDGYYWCRQI
RSTA_ECOLI (45) LVLLDIMLPGKDGMTICRDL
RCAC_FREDI (45) LIILDIMLPNLDGISLCKRF
CUTR_STRLI (45) VVVLDRDLPLVHGDDVCRKI
COPR_PSESM (45) LLILDVMMPGLDGWEVIRRL
YGIX_ECOLI (45) AVILDLTLPGMDGRDILREW
BASR_SALTY (45) LMVLDLGLPDEDGLHFLTRI
YGIX_HAEIN (45) AVVLDLTLPKLDGLEVLQQW
TCTD_SALTY (45) LAVLDINMPGMDGLEVVQRL
SPAR_BACSU (2)  LILLDVMMPDIDGFELCKQI
NISR_LACLA (45) LILLDIMMSNIEGTEICKRI
YC29_CYAPA (46) LIICDIIMPGMGGFNFLHQL
HNR_ECOLI  (45) LMICDIAMPRMNGLKLLEHI
```

// ADZ-Pattern signature II

```
TRANSCRIPTIONAL REGULATORY PROTEIN
ID H000514
LN 28
```

Table 2. *Continued*

// ADZ-Pattern signature II

```
ARCA_ECOLI (77) LMFLTGRDNEVDKILGLEIGADDYITKP
TORR_ECOLI (73) IILVTGRSDRIDRIVGLEMGADDYVTKP
VIRG_AGRT5 (73) IIIISGRLEEADKVIALELGATDFIAKP
OMPR_ECOLI (70) IIMVTAKGEEVDRIVGLEIGADDYIPKP
VANR_ENTFC (72) IIMLTGKDTEVDKITGLTIGADDYITKP
YC27_PORAE (77) IIMLTALSDVSDRITGLELGADDYIVKP
CPXR_ECOLI (72) VIMLTARGSELDRVLGLELGADDYLPKP
PETR_RHOCA (74) ILLLTARGETRERIEGLEAGADDYLPKP
PHOB_ECOLI (72) VVMLTARGEEEDRVRGLETGADDYITKP
BAER_ECOLI (75) IVMVTAKIEEIDRLLGLEIGADDYICKP
AFQ1_STRCO (74) IILLTARNDDIDVVVGLESGADDYVVKP
YC27_PORPU (76) IIMLTALGEVCDRITGLEIGADDYVVKP
PHOB_HAEIN (61) IIMLTAKSTEEDCIACLNAGADDYITKP
YV17_MYCLE (44) VIMVTARDSEIDKVVGLELGADDYVTKP
RESD_BACSU (72) IIMLTAKGEEANRVQGFEAGTDDYIVKP
CHVI_AGRTU (72) VIFLTSKDEEIDELFGLKMGADDFITKP
CREB_ECOLI (70) VLFLTARSEEVDRLLGLEIGADDYVAKP
YRKP_BACSU (70) ILFLTARSSTLDKTEGLLAGGDDYMTKP
KDPE_ECOLI (58) VIVLSARSEESDKIAALDAGADDYLSKP
YXDJ_BACSU (72) IIFISARSGEMDQVMAIENGGDDYIEKP
RSTA_ECOLI (72) IVLLTSLDSDMNHILALEMGACDYILKT
RCAC_FREDI (70) ILLLTAQDNITAKVQGLDAGADDYVVKP
CUTR_STRLI (67) VLMLTASGDVSDRVEGLEIGADDYLPKP
COPR_PSESM (68) VLFLTARDGVDDRVKGLELGADDYLVKP
YGIX_ECOLI (68) VLILTARDALAERVEGLRLGADDYLCKP
BASR_SALTY (72) VLILTAHDTLNDRITGLDVGADDYLVKP
YGIX_HAEIN (69) VLILTARDTLDERVKGLQSGADDYLCKP
TCTD_SALTY (72) PVLLTARSAVADRVKGLNVGADDYLPKP
SPAR_BACSU (29) ILFLTAKTEEEAIVKGLITGGDDYITKP
NISR_LACLA (72) IIFVSAKDTEEDIINGLGIGGDDYITKP
YC29_CYAPA (73) VILLTTRGLTQDRIIGYKTGCDSYISKP
HNR_ECOLI  (70) VLVISATENMADIAKALRLGVEDVLLKP
```

The comparison with both databases produced a catalog of 11 750 different patterns (ftp.bioinf.mdc-berlin.de/database/SWISSPROT/motiv.res). From our pattern recognition method and PROSITE database, we chose patterns containing at least a minimum length of 17 amino acids without gaps.

Out of a total of 73 459 protein sequences, 59 178 proteins could be classified into 3124 protein families from our pattern method [the algorithmic details of the pattern method are described in Hanke *et al.* (1996)]. The rest of the protein sequences were less well represented (they were either too closely related as a group or were only represented by a limited number of sequences).

Redundant sequences, both coinciding and overlapping with the PROSITE patterns, were filtered out (see Table 1). We used the PROSITE pattern annotation and chose complete sequences for our new, self-generated patterns using

SWISSPROT annotation. For every new pattern, we generated a signature according to the PROSITE rules.

Finally, the complete data set is divided into subpatterns of 17 amino acids (total 115 000 subpatterns).

In addition, secondary structure information from the pattern collection was added, for which we used the HSSP database. For every protein sequence pattern found, an HSSP database comparison was made. Of the 11 750 pattern classes, at least 1265 held a sequence–structure homology.

Our goal was both to include a clear overview of the relationship between primary and secondary structure information, and to include two distinct forms of information (normally requiring two different databases) in the associative network. This allowed not only a simultaneous sequence analysis and secondary structure prediction for protein sequences, but also a detailed observation of distantly related sequence motifs (structural motifs) on the 2D map.

Our method attempts to derive a sensitive pattern- or module-domain recognition method. The strength of this method lies in the recognition or summary of variable protein modules or protein domains supported by each of the secondary structure statements. Out of the 12 000 classes contained in the associative database, we chose two representative examples of domain recognition.

The method is as follows. The pattern generated in the first step (in this example, derived from the transcriptional regulatory proteins RESD_BACSU and YJDG_ECOLI) provided 47 homologous sequences for RESD_BACSU. Two conserved regions (ADZ pattern signature I and ADZ pattern signature II) could be extracted from that data set (Table 2).

The receiver domain

The receiver domain is a good example for the performance of this two-level pattern-oriented recognition and classification method, and is not annotated in PROSITE. We want to describe the two-level method in detail using this example. These domains are found in transcriptional regulatory proteins, but show extraordinarily low sequence homology within this family (<14% sequence identity). Sequence analysis methods, like Gap-Blast (Altschul *et al.*, 1997), provided no evidence of common functional domains or common sequence families.

Using the other query sequence, YJDG_ECOLI, seven similar protein sequences, not included in the previous results from RESD_BACSU, were found. These seven sequences all contained a conserved region of 59 amino acids length (Table 3).

All three conserved regions found were cut into fragments of 17 amino acids length and added to the complete data set of 115 000 partial patterns. After the classification of all partial patterns, we found one cluster (11 311 position 94 30), that contained both families (ADZ pattern signature I and the AQH pattern). We show here a sample from the total cluster of the 54 sequences (Table 4).

Table 3. Pattern recognition (receiver domain) in transcriptional regulatory proteins starts from query sequence YJDG_ECOLI

```
// AQH-Pattern TRANSCRIPTIONAL REGULATORY PROTEIN
ID H000119
LN 59
CITB_KLEPN (52) LILLDNFLPDGKGIDLIRHAVSTHYKGRIIFITADNHMETISEALRLGVFDYLIKPVHY
YJDG_ECOLI (41) LILLDIYMQKENGLDLLPVLHNARCKSDVIVISSAADAATIKDSLHYGVVDYLIKPFQA
LUXO_VIBHA ( 3) LILLDLRLPDMTGMDVLHAVKKSHPDVPIIFMTAHGSIDTAVEAMRHGSQDFLIKPCEA
PILR_PSEAE (31) LCLTDMRLPDGSGLDLVQYIQQRHPQTPVAMITAYGSLDTAIQALKAGAFDFLTKPVDF
HYDG_ECOLI (21) LVLCDVRMAEMDGIATLKEIKALNPAIPVLIMTAYSSVETAVEALKTGALDYLIKPLDF
HYDG_SALTY (21) LVLCDVRMAEMDGIATLKEIKALNPAIPILIMTAFSSVETAVEALKAGALDYLIKPLDF
ATOC_ECOLI ( 3) VVLMDIRMPEMDGIKALKEMRSHETRTPVILMTAYAEVETAVEALRCGAFDYVIKPFDL
```

Table 4. Receiver domain classified in one neuron

Position 94 30 (11311)						
Seq-ID	Position	Sequence	Description	Score	Code	Structure
YJDG_ECOLI	41	LILLDIYMQKENGLDLL	Hypothetical 27.4 kDa protein	24.7	ADZ	EEEE S SSSS TTH
PILR_PSEAE	31	LCLTDMRLPDGSGLDLV	Fimbriae expression regulat	27.2	ADZ	EEEE S SSSS TTH
ATOC_ECOLI	3	VVLMDIRMPEMDGIKAL	Acetoacetate metabolism reg	23.4	ADZ	EEEEES SSS HHHHH
ARCA_ECOLI	50	LVIMDINLPGKNGLLLA	Aerobic respiration control	25.8	AQH	–
ARCA_HAEIN	50	LVVMDINLPGKNGLLLA	Aerobic respiration control	25.9	AQH	EEEE S SSSS TTH
RESD_BACSU	45	LILLDLMMPGTDGIEVC	RESD protein	23.2	AQH	EEEEES SSS HHHHH
HNR_ECOLI	45	LMICDIAMPRMNGLKLL	HNR protein	24.7	AQH	EEEEES SSS HHHHH

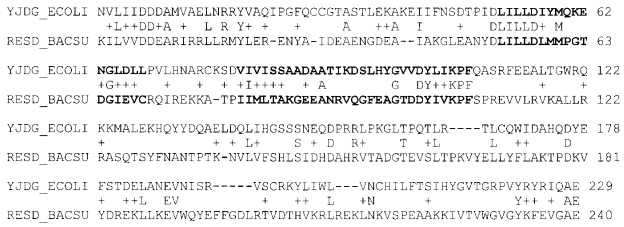

Fig. 1. Weak pairwise alignment of the two query transcriptional regulatory proteins (RESD_BACSU and YJDG_ECOLI). The conserved regions (shown in bold) extracted from the pattern method are classified as 'receiver domain' and contained within two typical neurons (one receiver domain neuron example shown).

The first line shows the position of the cluster on the 2D map, as well as its cluster number. The most important information found using the associative databank is shown in the form of a table for each cluster. This provides the molecular biologist with a quick overview of function, structure and sequence (a more detailed description of the table is given in Figure 2C). Distantly related sequences can be found by observing the neighborhood on 2D map.

Both previous search results showed similarity in primary and secondary structure information for this cluster. [In Figure 1, we examine the pairwise alignment of the two query sequences; a weak agreement is shown over the entire length of the sequences identities 42/239 (17%).]

Alignment is difficult with such a low sequence homology. The only relatively homologous areas were found using the pattern recognition method and also summarized in the classification step. The second conserved region was also found in a cluster at position 30 34 (not shown here) The classification step associated the two query families. The common, multiple appearance of different families confirmed, therefore, the theory that we were dealing with one domain. It was also possible to show a common relationship using the iterative blast method (PSI-Blast; Altschul *et al.*, 1997) after the third iteration (E-value = 10–28).

Sushi module (CCP)

This module is abundant in complement control proteins and also in the complement system itself. It has also been found in viruses, the enzyme thyroxide peroxidase and in protein families such as selectins and mucins. Its functions are still unclear and may well be different in the various proteins where it has been found so far (Bork *et al.*, 1996). CCPs occur most frequently in tandem arrays, but also a few single copies have been identified. An example is the classification of class 1626. We show here a sample from the total cluster of the 72 sequences. The complement receptor CR2, necessary and sufficient for binding natural ligands, the cell adhesion selectin precursor proteins and the apolipoprotein r precursor protein fell into one class.

Such a module is too short to be found with standard homology tools. Generally, it needs a lot of experience to find such small similarities in these diverse families (Table 5).

The associative database as application

The actual version of our associative memory of protein motifs consists of 115 000 conserved segments distributed over 12 000 classes. About 92% of all SWISSPROT sequences match to at least one submotif in this memory. The rest (total 8932 sequences) consist of either single sequences or of sequences without a marked conserved region in the primary structure. These are stored in separate classes of the memory.

Table 5. Sushi module classified in one neuron

Position 13 65 (1626)						
Seq-ID	Position	Sequence	Description	Score	Code	Structure
APAR_PIG	59	CDEGYTLVGEDRLSCRS	Apolipoprotein r precursor	32.5	AKO	–
CR2_MOUSE	31	CDPSFTLIGEKTISQKN	Complement receptor type	31.7	AKO	–
CR2_HUMAN	30	CKTNFSLIGESTIRCTS	Complement receptor type	31.5	AKO	–
CFAH_MOUSE		CDDGYQLVGQDYLRCTA	Complement factor H precur	34.1	AKO	–
LEM2_HUMAN	23	CEEGFELMGAQSLQCTS	E-selectin precursor (endot	31.2	CIT	–
LEM2_RABIT	23	CEEGFTLLGARSLQCTS	E-selectin precursor (endot	31.3	CIT	–
LEM2_PIG	13	CKEGFELIGPEHLQCTS	Selectin precursor (endothe	30.2	CIT	–
CO2_HUMAN	23	CSSNLVLTGSSSRLCKS	Complement C2 precursor	30.1	CIT	–
CFAB_HUMAN	62	CPSGFYPYPVQTRTCRS	Complement factor B (precur	31.3	CIT	–

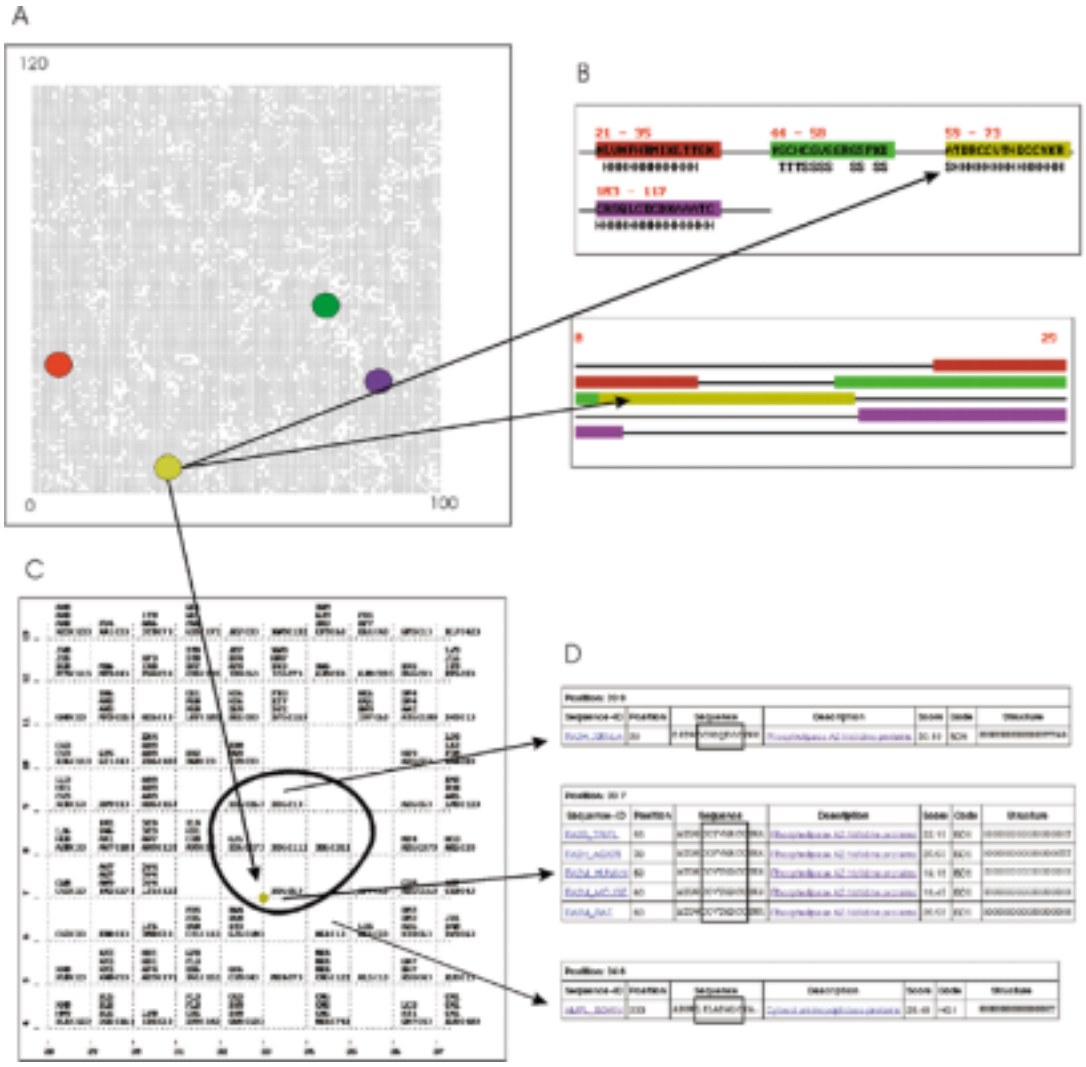

Fig. 2. Associative memory of protein sequence motifs displaying the fingerprint of the phospholipase A2 family, Ca-dependent subgroup (Dennis, 1997). (**A**) A grid of 120 × 100 memory cells ('neurons') is displayed highlighting spots (each comprising a cluster of 3–6 neurons) that indicate the position of four characteristic sequence motifs of the PA2 family. Spots (each covering several neurons) appear when 'phospholipase A2' is called, whereas a single query sequence (such as PA2M_HUMAN) will result in the same pattern, but with finer spots (as then only one cell per spot is activated). The spots identify characteristic motifs as present in all true family members: red, the LFI motif (L-X(2)-F-X(3)-I); ochre, the CCHC motif (C-C-X(2)-H-X(2)-C-C); green, the YGCYCG motif (Y-G-C-Y-C-G(4)-X-G-X-P-X-D); purple, the CCDC motif ([LIVMA]-C-[ADEGHKNQR]-C-D-X(5)-C). (**B**) PA2M_HUMAN as query has produced four spots. Shown is a schematic alignment of the hit motifs in the sequence, including position, together with a tentative prediction of the secondary structure (H, helix; T, turn; S, sheet). (**C**) Zoom into the region of the CCHC motif. For each individual memory cell of the grid, a bracketed number states how many motifs are stored there, identifying the first few (up to four) of them by a short code. Further information may be obtained by clicking into the quadratic domain pertinent to a cell [see (D)]. Five memory cells (encircled) are being occupied by 91 motifs of the queried PA2 family (code: BD6). (**D**) Tables of sequence motif information are provided when individual cells of (C) are opened. Shown is the content of cell (33,7) with five unambiguous PA2 motifs (with the CCHC motif, boxed). Cell (33,9) contains a 'maverick' PA2 member from *Xenopus laevis*, the sequence of which, though being closely related to the other members in the neighboring cells, has replaced by Q the obligatory H of the active center, hence its catalytic function is lost. Cell (34,6) contains a segment of bovine aminopeptidase (code: HG1), biochemically quite unrelated, with faint similarity to the PA2 motif, but, interestingly, also predicted to form an α helix. Clicking on any table item (not shown) established links to background databases such as SWISSPROT, PROSITE and further, in the usual way. In this manner, the whole vicinity of PA2M_HUMAN and its family, as well as neighbors, may be systematically studied on the conspicuous 2D map supported by hyperlinked text information. The recently detected other subgroups (Dennis, 1997) of PA2 get quite different fingerprints in the memory, pointing to evolutionary unrelatedness. PAF acylhydrolase, secreted into human plasma, for instance, gets only one hit containing the characteristic G-X-S-X-G motif of the active center of esterases and lipases. Obviously, the superfamily is functionally defined, not by common ancestry.

A query sequence (protein, putative exon, EST, translated genomic region, etc.) is divided into segments which are projected and shown as a 'hit' on the map. The resulting hit pattern (a 'motif fingerprint'; see Figure 2A) is typical for that sequence and may serve to identify its class affiliation as well as to strengthen the memory by updating the prototype vector. An alternative approach to the memory is to specify a sequence family in order to find its common motifs.

The hit pattern, consisting of segments typical for that protein in its family, may be schematically aligned to show its position in the query sequence (Figure 2B). A different application is a zoom into a certain region of cells containing a family of similar motifs (Figure 2C). Each memory cell may present its content and offer (by a double-click on the WWW map) links to pertinent text information (Figure 2D) or to structural information as present in the background databases supporting the memory.

The sensitivity and selectivity of a query are illustrated by the phospholipase A2 family of proteins (PA2). We take the subgroups of Ca-dependent enzymes (groups I–III; Dennis, 1994) as an example. It has two signatures in the PROSITE database relating to its catalytic center (His-Asp pair). Having trained the 120 × 100 cell memory with all conserved segments available, we look up where PA2 is stored and obtain a fingerprint of four spots (Figure 2A) characteristic of this family. Any single true member, when submitted as a query, will show the same unambiguous pattern of spots, which may be complemented by additional information (see Figure 2B–D) on that item as well as on its neighbors in the sequence space. Thus, the map allows the study of sequences by partial motifs, in great detail and in quite conspicuous form: as fingerprinted spots of families, with subtle variants in different cells, including outsider (single sequence without homology) at the fringe, closely related motifs and their character of faint sequence similarity, including prediction of higher structure. All this can be done in turn, motif by motif. Lengthy domains are divided into convenient submotifs of suitable length. The memory is capable of storing all motifs (many thousands) as known to be conserved in the sequence space.

Discussion

The prevailing method of sequence analysis in huge files of information items rests on text linking (often hampered by sloppy nomenclature) and on similarity study. Several tools depart one step into the 2D world by storing index matrices of pre-processed 'hits' between raw data (Mewes *et al*., 1997; Vingron and von Haeseler, 1997). Another way is to envisage proteins as arrays of motifs (being signatures standing for domains or modules) and to establish links between sequences in terms of shared motifs or domains (Bork *et al*., 1996).

With the associative data memory, we offer to push ahead further into the 2D world of shared motifs. We make use of the neighborhood concept that emerges from the notion of a map. A geographic map is superior to a table of distances between selected points in a landscape. Likewise, the display of a sequence memory as a 2D map gives a clear picture of cluster structure as well as of neighborhood 'topology'. As we make use of the 'motif' concept of sequence study and transform similarity into distance, we circumvent the common problem of non-transitivity in sequence analysis (A and B may both be close to C, but completely alien to each other, because the closeness relates to different domains). The size of motifs has to be chosen as a compromise: sufficiently short to display sequence conservation as a small distance, but sufficiently long to avoid random clustering. For protein sequences, this means that motifs of about 17 residues are expedient. Sequences thus envisaged as a succession of conserved motifs implies sequence diagnostics in the form of fingerprints. In the case of multidomain proteins (Bork and Koonin, 1996), this may result in quite a bunch of motif hits.

The associative memory is a tool for retrieval, diagnosis and comprehensive visualization of biomolecular sequences that effectively complements (not replaces) the hitherto dominating list-and-table software. Beyond sequence relationships, it is possible to evoke prediction of secondary structure and of function or regulation (if this is deducible with sufficient confidence from the sequence). Apparent dissimilarities between amino acid sequences within a cluster can be explained by similar secondary structure (see Figure 2D).

The accompanying WWW server allows an extremely rapid analysis of either peptide or nucleotide sequences (translated in all six reading frames) (see http://www.bioinf.mdc-berlin.de/unter2.html). The interactive network is on-line and was implemented on a parallel computer (Sparc-20 with four processors).

In conclusion, we propose the further development of visually oriented analysis of the sequence space. Extension to the performance of associative memories for genomic repeats and for ESTs is being studied at present. We are convinced that the rapid worldwide progress in genome decoding necessitates the integration of associative data storage into the toolbox of information analysis.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Attwood,T.K. *et al*. (1997) The PRINTS database of protein fingerprints: a novel information resource for computational molecular biology. *J. Chem. Inf. Comput. Sci*., **37**, 417–424.

Bairoch,A. and Apweiler,R. (1998) The SWISS-PROT protein sequence data bank and its supplement TREMBL in 1998. *Nucleic Acids Res.*, **26**, 38–42.

Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.*, **25**, 217–221.

Bork,P. and Koonin,E.V. (1996) Protein sequence motifs. *Curr. Opin. Struct. Biol.*, **6**, 366–376.

Bork,P., Downing,A.K., Kieffer,B. and Campbell,I.D. (1996) Structure and distribution of modules in extracellular proteins. *Q. Rev. Biophys.*, **29**, 119–167.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, pp. 345–352.

Dennis,E.A. (1994) Diversity of group types regulation and function of phospholipase A2. *J. Biol. Chem.*, **22**, 13057–13060.

Dennis,E.A. (1997) The growing phospholipase A2 superfamily of signal transduction enzymes. *Trends Biochem. Sci.*, **22**, 1–2.

Gerstein,M., Sonnhammer,E.L. and Chothia,C. (1994) Volume changes in protein evolution. *J. Mol. Biol.*, **236**, 1067–1078.

Green,P., Lipman,D., Hillier,L., Waterston,R., States,D. and Claverie,J.M. (1993) Ancient conserved regions in new gene sequences and the protein databases. *Science*, **259**, 1711–1716.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.

Hanke,J. and Reich,J.G. (1996). Kohonen map as a visualization tool for the analysis of protein sequences: multiple alignments, domains and segments of secondary structures. *Comput. Appl. Biosci.*, **12**, 447–454.

Hanke,J., Beckmann,G., Bork,P. and Reich,J.G. (1996) Self organizing hierarchic networks for pattern recognition in protein sequence. *Protein Sci.*, **5**, 72–84.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Henikoff,J.G. and Henikoff,S. (1996) Blocks database and its applications. *Methods Enzymol.*, **266**, 88–105.

Hertz,J., Krogh,A. and Palmer,R.G. (1991) *Introduction to the Theory of Neural Computation*. Addison Wesley, Redwood City.

Kohonen,T. (1996) *Self-organization and Associative Memory*. Springer-Verlag, Berlin.

Lynch,Th.J. (1985) *Data Compression Techniques and Applications*. Lifetime Learning Publications, Belmont.

Mewes,H.W., Albermann,K., Heumann,K., Liebl,S. and Pfeiffer,F. (1997) MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.*, **25**, 28–30.

Ritter,H., Martinetz,Th. and Schulten,K. (1992) *Neural Networks: An Introduction to the Neural Informations Process of Self-Organized Networks*. Addison-Wesley, Bonn.

Schneider,R., De Daruvar,A. and Sander,C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **25**, 226–230.

Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 320–322.

Vingron,M. and von Haeseler,A. (1997) Towards integration of multiple alignment and phylogenetic tree construction. *J. Comput. Biol.*, **4**, 23–34.