Protein Science (1996), 5:72–82. Cambridge University Press. Printed in the USA.

Self-organizing hierarchic networks for pattern recognition in protein sequence

JENS HANKE, GEORG BECKMANN, PEER BORK, AND JENS G. REICH

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

(RECEIVED August 14, 1995; ACCEPTED October 10, 1995)

Abstract

We present a method based on hierarchical self-organizing maps (SOMs) for recognizing patterns in protein sequences. The method is fully automatic, does not require prealigned sequences, is insensitive to redundancy in the training set, and works surprisingly well even with small learning sets. Because it uses unsupervised neural networks, it is able to extract patterns that are not present in all of the unaligned sequences of the learning set. The identification of these patterns in sequence databases is sensitive and efficient.

The procedure comprises three main training stages. In the first stage, one SOM is trained to extract common features from the set of unaligned learning sequences. A feature is a number of ungapped sequence segments (usually 4–16 residues long) that are similar to segments in most of the sequences of the learning set according to an initial similarity matrix. In the second training stage, the recognition of each individual feature is refined by selecting an optimal weighting matrix out of a variety of existing amino acid similarity matrices. In a third stage of the SOM procedure, the position of the features in the individual sequences is learned. This allows for variants with feature repeats and feature shuffling.

The procedure has been successfully applied to a number of notoriously difficult cases with distinct recognition problems: helix-turn-helix motifs in DNA-binding proteins, the CUB domain of developmentally regulated proteins, and the superfamily of ribokinases. A comparison with the established database search procedure PROFILE (and with several others) led to the conclusion that the new automatic method performs satisfactorily.

Keywords: amino acid sequences; multiple alignment; neural network; pattern recognition; self-organizing maps

In sequencing projects, a database search for similar sequences is an inexpensive first attempt at suggesting the biological function of newly sequenced primary structures. More and more sequences are assignable to families, and there are a number of published procedures for the heuristic recognition of local sequence patterns using the information inherent in a set of related specimens or in a consensus model. All such strategies have a circularity problem, however, in that pattern recognition presupposes a valid alignment of the sequences, whereas the construction of an alignment requires previous knowledge of the pattern. Although in the case of a very clear-cut and distinct pattern this difficulty may be alleviated by a skillful iteration procedure, serious problems may arise when one or several of the following situations apply: the presence of a fuzzy pattern (difficult to distinguish from noise), very liberal alignment (too many possible insertions/deletions), or undersampling (prohibiting statistical analysis of pattern elements). An additional and very common practical problem is that extensive searches in voluminous databases become prohibitive, because unrelated noise accumulates even in the neighborhood of true samples, and automatic routines become time-consuming and dependent on laborious interactive decisions after inspection of intermediary results. The most popular search method (Gribskov et al., 1987, 1990; Lüthy et al., 1994) is essentially a statistical approach that requires a preset alignment and derives pattern descriptors from the single-position frequency of amino acids, combined with downweighting of redundant sequences and evaluation according to a mutation table. A different strategy (Smith et al., 1990; Henikoff & Henikoff, 1991; Neuwald & Green, 1994; Tatusov et al., 1994) circumvents the intricate alignment and looks instead for ungapped partial blocks whose common presence leads to the identification of a motif.

Pattern recognition in biopolymer sequences is a task that may be tackled with a recently developed class of algorithms based on distributed computational capability (Rumelhart & McClelland, 1986; Hertz et al., 1991). These methods are known as artificial neural networks (ANN). In one variant, the learning

Reprint requests to: Jens Hanke, Max-Delbrück-Center for Molecular Medicine, Department of Bioinformatics, Robert-Rössle-Str. 10, D-13125 Berlin-Buch, Germany; e-mail: hanke@bioinf.mdc-berlin.de.

program is called "supervised," because it is confronted with two or several classes of objects, each of them carrying its unique class pertinence as a label. The goal is to learn correct classification of objects. In a second variant, named "unsupervised," the learning process includes both definition of classes and their correct classification. ANNs that learn under supervision have been applied to prediction of secondary structure from the information supplied by a database of known structures (Bohr et al., 1988; Qian & Sejnowski, 1988; McGregor et al., 1989; Rost & Sander, 1994), and also in motif definition and recognition in protein sequences (Bengio & Pouliot, 1990; Hirst & Sternberg, 1991; Frishman & Argos, 1992).

An unsupervised, self-classifying strategy is appropriate for treatment of patterns. It may support or even replace the often dubious previous definition (including classification and alignment) of patterns. This paper proposes the use of Kohonen maps (1989, 1990) for this purpose. These are self-organizing neural networks that compress a more or less compact training set of high-dimensional vectors to low-dimensional ones, arranging them on a (usually two-dimensional) map. Such a strategy has the advantage of tolerating low accuracy of signal representation and of synaptic weights. The approach has been applied to detecting signal peptides (Arrigo et al., 1991) and to clustering protein sequences into families according to their degree of sequence kinship (Ferrán & Ferrara, 1991).

Here we want to apply Kohonen maps to the task of motif definition in a training set and to subsequent recognition of query objects. In contrast to previous applications, but by analogy with brain functional mapping, we arrange Kohonen maps in a hierarchic structure of self-organizing maps (SOMs). We chop all sample sequences into the set of constituent words (4–16 letters, a selectable parameter), encode them as a property-describing or scoring vector, and present this information to an unsupervised self-organizing Kohonen map. This leads to a map in which words are portrayed and agglomerate locally (as features, a feature being a characteristic subsequence occurring in the set) according to their similarity. Each feature so obtained is introduced into a tuning procedure that takes place in a second "floor" of Kohonen maps, which selects the "best" evaluation principle for that class of objects (emphasizing, for example, evolutionary relationship, physicochemical properties of residues, information content, or structural propensity). Only after this step is the position of the feature specimens in the sequence block considered; if desirable, a common alignment is then suggested. The ordering of partial stages is to some extent similar to the procedure applied in a different context by Pearson and Lipman (1988), which first finds accumulations of small "tuples" in order to align them afterward. The strategy involving Kohonen maps as presented here is applicable to motif families with sequence lengths of more than 15 residues.

We tested this method on several sets of sample proteins (difficult cases, different types of problems) and compared it to other methods.

Results

We examined the performance of a set of hierarchic arrayed Kohonen maps at the recognition of subtle motifs in protein sequences. The search strategy, as described in the Materials and methods, involves three stages: (1) finding features in chopped sequences; (2) tightening pattern definition; and (3) evaluating the occurrence, array, and position of features. We did three series of studies, taking notoriously intricate test problems from the literature, and compared the performance of our SOM hierarchy with that of some well-established recognition methods.

Helix-turn-helix motif in DNA-binding proteins

The helix-turn-helix (HTH) motif belongs to a class of protein sequences that are bacterial transcription factors or similar proteins (Brennan & Matthews, 1989; Anderson, 1992; Dodd & Egan, 1990; Chapter 7 of Branden & Tooze, 1991; Pabo & Sauer, 1992). The HTH motif is the site that binds to regulatory sites at DNA. It is about 20 amino acid residues long, and there is 3D structural information available. There is no alignment problem between most of the specimens, because the two helices and the short turn in between are superimposable without gaps. On the other hand, the sequence information is extremely fuzzy, because a wide spectrum of only slightly related sequence variants is compatible with the requirement of a HTH structure. The PROSITE database (release 12) lists several hundred sequences classified into 10 families with different sequence "signatures."

In addition to the bacterial transcription factors, there is a further superfamily of many hundred sequences of approximately the same length, originating in eukaryotic sources: the so-called "homeodomains" or "homeoboxes" (Chapter 8 of Branden & Tooze, 1991; Pabo & Sauer, 1992; Treisman et al., 1992). They perform a similar regulatory function and also contain a HTH motif. Structural information is likewise available. The HTH motif of homeoboxes is in the twilight region of similarity to the bacterial transcription factors, and they may be aligned together.

Recognition and analysis of the HTH motifs is a very tough matter that has led to much controversy (Yudkin, 1987, 1988; Dodd & Egan, 1988, 1990; Lawrence et al., 1993). We took a subset of HTH motifs as selected and studied in detail by Lawrence et al. (1993) and Neuwald and Green (1994). The neural networks were trained on this set to identify these and other HTH motifs in the whole protein sequence database.

The training set consisted of 12 HTH patterns (each 20 amino acids long) and included 1 homeobox and 11 bacterial transcription sites. These motif examples are so diverse that none of the mutual pairwise PAM120 scores surpassed a value of 35, which is a faint similarity at best.

The SOM procedure chopped each of the 12 learning sequences into five tetrapeptides, collected them in five files, and subjected each file to the learning cycle of a specialized Kohonen map. The map contained 16 neurons; that is, slightly more than the number of sequence segments. After training, we performed a run through the SWISSPROT database (Bairoch & Boeckmann, 1993) and collected in every 20-window of all sequences the quantization error (q.e.) (a distance measure) reported by the five maps. A file was collected of all 20-windows where the five maps reported a moderately low q.e. (threshold: positive if < 3.0). The results (Table 1) may be summarized as follows:

1. All training sequences were retrieved as a set of positive reports from all five maps with extremely low q.e. (≤ 0.003). This means that they all were redetected with high fidelity (which is no surprise, because the neurons were numerous enough to "learn by heart").

Table 1. *Self-organizing hierarchic networks for pattern recognition in protein sequence*

ID	Position	N_1	N_2	N_3	N_4	N_5
NIFA_KLEPN	495	**0.0003**	**0.0001**	**0.0001**	**0.0001**	**0.0004**
HMAN_DROME	326	**0.0002**	**0.001**	**0.0001**	**0.0001**	**0.0004**
DICA_ECOLI	22	**0.002**	**0.001**	**0.0001**	**0.002**	**0.004**
DNIV_ECOLI	160	**0.003**	**0.0003**	**0.001**	**0.0002**	**0.0001**
DEOR_ECOLI	23	**0.002**	**0.0002**	**0.001**	**0.0002**	**0.0002**
ARAC_ECOLI	196	**0.0002**	**0.004**	**0.001**	**0.001**	**0.002**
NAHR_PESPU	22	**0.001**	**0.002**	**0.0003**	**0.0001**	**0.0002**
MERD_SERMA	5	**0.001**	**0.0004**	**0.003**	**0.0001**	**0.0004**
FIS_ECOLI	73	**0.003**	**0.0003**	**0.002**	**0.002**	**0.0003**
RPC2_LAMBD	25	**0.0001**	**0.002**	**0.0001**	**0.0002**	**0.0001**
RP32_ECOLI	252	**0.002**	**0.0001**	**0.0001**	**0.001**	**0.001**
LEXA_ECOLI	27	**0.0001**	**0.0001**	**0.002**	**0.0003**	**0.0001**
RPSK_BASCU	94	0.8	0.8	0.14	1.2	1.1
RPSB_BASCU	223	0.9	1.9	0.14	0.9	0.7
NTRC_BRASR	449	1.2	1.0	0.64	1.05	0.82
MTA1_YEAST	99	1.1	0.6	0.6	1.6	0.4
CRP_ECOLI	169	0.9	1.9	1.1	0.9	1.4
RCRO_LAMBD	15	0.9	0.9	1.1	0.8	1.1
RCRO_BPP22	12	1.1	0.6	0.1	1.1	0.4
FNR_ECOLI	196	1.4	1.5	0.0001	1.0	0.7
NTRC_KLEPN	444	1.2	0.0001	1.0	1.0	0.8
CYTR_ECOLI	11	0.8	1.0	0.9	1.1	0.7
GALR_ECOLI	3	0.9	1.0	0.001	1.2	1.0
LACI_ECOLI	5	1.3	0.9	0.001	1.2	1.4
TER2_ECOLI	26	0.9	0.6	0.8	0.8	1.1
TRPR_ECOLI	67	0.8	1.3	1.4	1.2	1.6
RPSE_BASCU	205	0.8	1.2	0.4	1.0	0.7
PURR_ECOLI	3	1.0	0.4	1.1	1.0	0.9
EBGR_ECOLI	3	0.6	0.9	0.0001	0.7	1.1
RPC1_BP22	25	0.9	1.0	0.4	0.6	0.0004
HXA7_XENLA	27	1.1	0.7	1.5	1.2	1.4
HMAA_DROME	164	0.8	1.4	1.2	1.1	2.5
HXB6_HUMAN	172	1.1	0.8	1.5	1.8	1.2
HMAN_DROME	323	1.2	1.3	2.3	1.0	1.7
HM90_APIME	34	1.6	1.2	1.1	1.2	1.5
HMSC_APIME	34	1.4	1.2	1.4	1.7	1.8
HXB5_HUMAN	220	0.9	1.8	2.3	2.1	1.9
HXA5_MOUSE	221	1.0	1.9	2.1	2.4	2.1
HXB4_MOUSE	187	1.2	1.5	2.3	2.4	2.2
HXB7_HUMAN	163	1.3	1.7	2.6	2.3	2.1
HXC6_MOUSE	167	2.2	2.1	2.4	2.1	2.5
HXD4_CHICK	170	2.1	2.4	2.5	2.6	2.3
HXB6_CHICK	33	2.5	2.3	2.6	2.7	2.6

[a] Recognition of features contained in HTH patterns of DNA-binding proteins by a parallel array of SOMs. The first 30 rows of the table refer to individual HTH patterns from the collection assembled as a test set by Lawrence et al. (1993). SWISSPROT acronyms are reported in the first column. Column 2 identifies the position of the first amino acid of the HTH motif (20 residues long) in the complete sequence. The upper block of the table displays the 12 patterns used as a training set, and the 18 patterns in the middle block are the application (test) set. *Training stage*: Each HTH pattern was chopped into five nonoverlapping 4-words and presented to the corresponding SOM (labeled N_1–N_5) for training. *Application stage*: All window positions in the protein database were chopped into five consecutive 4-words and presented in turn for diagnosis to the trained SOMs. The table shows the "hits" obtained (identified by q.e. < 3.0 pointing to similarity). It is evident that each SOM redetected the complete learning set (q.e. ≤ 0.003), which by itself is a difficult task, because of its diversity (see text), and recognized (with less similarity, q.e. around 1.0) the application test of bacterial HTH sites (block II). A final block of 13 patterns shows a subset of additionally detected homeoboxes (q.e. < 3.0). Many further binidng sites are detected with a looser acceptance criterion (e.g., not all subfeatures require q.e. < 3.0), but then false positives would also appear.

2. Such an extremely low q.e. was only occasionally reported for other sites in SWISSPROT (examples in Table 1, second block), but never by all five maps.

3. Many bacterial HTHs (in particular, all those of the paper by Lawrence et al., 1993) showed up with moderately low q.e. in all five tetrapeptide maps and in proper arrangement (i.e., are identified as HTHs by the network).

4. Many homeoboxes appeared also at moderately low q.e. (Table 1, last group), with no suspected false positives in between (identification of a false positive was not always sure, as several proteins of unkown function appeared in the list).

We consider this an excellent result. It compares satisfactorily with that of the PROFILE study we undertook with the same training set. The PROFILE procedure was not only blindly used: a wide spectrum of parameters was studied, sequence weights were incorporated, and different evaluation matrices were applied as described by Lüthy et al. (1994).

The result of a PROFILE search through SWISSPROT was in all cases essentially the same (not much influenced by variants of technique). A list of around 100 window sites appeared with high to moderate Z scores (>5.0, pointing to similarity with the profile). The profile identified mainly homeoboxes (very numerous in the database); many fewer of the equally numerous bacterial sites were identified (although they made up 11 of the 12 training HTHs). There were always a few definite false positives interspersed with the true positives (some of them with high Z scores), which is an indication that the "detection radius" of the profile is so large that a number of fortuitous hits appear, pointing to the high "volume" of the loose HTH cluster. The 12 training sequences were typically retrieved only partly: usually about one third with a high Z score, one third with a moderate Z score (dubious statistical significance), and one third with a clear "random score," sometimes below 2.0.

We also performed a database search with the HTH profile obtained from 21 sequences by Gribskov et al. (1990) as available in the public domain. The result was a similar list of about 100 HTHs with significant Z scores. Its composition emphasized the bacterial origin of the profile: more bacterial transcription factors, only a few homeoboxes, again a few dubious or false positives, and again the typical fact that only part (10 of 21) of the input sequences to the file were retrieved with high Z scores.

The fact that a profile retrieves only part of its own input is the clearest indication that HTH motifs are only loosely similar. The profile constructs the analogon of a "center of gravity" of a family, and it can retrieve only those input sequences that are close enough to the bulk. Weighting (Lüthy et al., 1994) may somewhat redress the effect, but it cannot change the large distance of the periphery from the center of gravity.

By contrast, the SOM has a special property here. It can "learn" a distant HTH with the same precision as a central one. Therefore, the input sample members and their closest relatives are always redetected with high fidelity, and hits by pure random coincidence have a much smaller chance of getting a low q.e. in all partial maps. Chopping an alignment into oligopeptides increases this effect, as we concluded from an experiment with training one Kohonen map by the full HTH motif (20 amino acids). This attempt produced a worse separation of HTH variants and non-HTHs. A further experiment was designed to detect HTH motifs without specifying their position in the total protein sequence. Even in this case, the "feature-extraction" method was able to recognize the motifs in 7 of 12 cases.

For comparative purposes, we undertook a retrieval study of HTH patterns with the program BLIMPS, available in the public domain (Wallace & Henikoff, 1992). This program is block-oriented and therefore suitable for diagnosis of cases where there is a well-conserved substructure in the family. The result confirmed what the ASSET study (Neuwald & Green, 1994) had concluded already. Four of the 12 representative motifs were redetected by BLIMPS (the output contained bacterial DNA-binding proteins in the first place), and the one homeobox was also identified, but in two further cases the correct assignment appeared on the lower part of the list only, and in five cases, the correct assignment was missed entirely. Again, this result points to the very vague cluster structure of the HTH region.

Ribokinase family

The ribokinase family is one of several sugar kinase families that catalyze ATP-dependent phosphorylation of sugars. Bairoch's PROSITE database (Bairoch, 1993) lists the current established set of 22 sequences (between 280 and 430 residues long) under the name "pfkb-family." He includes, in the form of regular expressions, two consensus signatures as typical for this set. Bork et al. (1993) have studied this family in more detail. For sequence searching, the particular intricacy of this motif is that its elementary features are separated by long unrelated regions of variable length (up to 150 residues).

Because this conservation pattern is apparently different from that of the HTH family, we tested an SOM network on this family. Only three full-length specimens were used as learning set (Fig. 1). They were presented without any alignment to a first-stage map, which distinguished 15 features, and retrained on a parallel array of 15 second-stage SOMs, resulting in the set of typical features displayed in Figure 1.

The 15 SOM networks so instructed looked into the whole SWISSPROT database and found (as a set of maps "flashing," as it were, in the correct position) all 22 specimens of the family contained in the database. Without second-stage refinement (by matrix retraining), identification was possible in only 17 cases. The PROFILE software, with a CLUSTAL-W-derived alignment (Higgins et al., 1992), was also able to "learn" on these three sequences and produced (without refinement) 19 of 22 specimens present in the database. Also, BLIMPS (Wallace & Henikoff, 1992) is able to identify three blocks of the ribokinase family, but not in all cases (a notable exception is LACC_ST AAU, from which only one of the motifs was found).

Figure 1 shows sections of the sequences and their typical features. A part of the 15 features is identical with the two "signatures" of PROSITE and with the conserved blocks of Bork et al. (1993). These are rather well conserved (Gs, Ds, and Ns present in all three specimens at identical positions). The other features as found by SOM are more fuzzy (e.g., only a few residues identical in two of the learning sets, and none identical in all three). A problem in all standard methods is to find all this in the bulky learning set without previous alignment. The oligoword strategy of SOMs performs satisfactorily.

In particular, the extraction of the distantly related INGK_ECOLI is remarkable. The overall sequence similarity of the 11 features, as identified (see Fig. 1) to any one of its analoga in the three learning sequences, is in the twilight zone below 25% letter identity. The identification is complicated by the facts that the features in INGK_ECOLI are in completely unrelated positions of the sequence and that several features of the training set are lacking in INGK_ECOLI.

It is noteworthy that, after inclusion of the newly found five specimens, the block shows much more sequence identity than the original three ones, suggesting (in a situation where one would not know this in advance) that a cluster has been collected rather than just five more random sequences. This ability to identify rather fuzzy features from an undersampled set (if indeed a feature is present) is one of the peculiar capacities of the Kohonen map hierarchy.

CUB domain

The widespread CUB domain occurs in various developmentally regulated extracellular proteins and in complement proteins (for details, see Bork & Beckmann, 1993). It contains nine repeated feature blocks of length ~6–8 residues, separated by unrelated variable regions of flexible length (2–8 residues). Only two cysteine residues seem to be invariant in all of the 18 modules present in SWISSPROT, release 31. The secondary structure attained by these patterns has so far not been rigorously determined, but prediction analysis suggests all β structures. The sequence features are vague, consisting mainly of hydrophobic and partly aromatic amino acids. A unique multiple alignment and hence similarity analysis is very difficult to establish. PROFILE was offered a training set of five specimens and found afterward 11 of 18 CUB examples present in the database with a Z score above 6.0 and ahead of the first false positive. We did not pursue PROFILE further with sophisticated parameters, because we wanted only to show the CUB problem to be intricate enough. Our SOM hierarchy, having learned on the same set of five specimens, detected all 18 sequences in a fully automated run. Figure 2 shows a detail of the CUB set that illustrates that three characteristic patterns (7- or 8-words) were found by an SOM, without previous supervised instruction and without alignment.

Discussion

Heuristic recognition of patterns in protein sequence sets is a notoriously difficult task because different factors influence the appearance of a sequence: (1) the evolutionary origin of the sequence; (2) functional requirements (e.g., the need to form an active center); and (3) structural demands on the chain (e.g., the need to form a certain globular structure or to fit into a membrane medium). The sequence contains all this information, but in a more or less veiled form. Hence, some features in a family of related proteins may be quite evident (e.g., two cysteine residues at a certain distance), whereas others are extremely fuzzy or even obscure at the level of primary structure.

There exist a number of procedures for heuristic prediction. They rely on numerical indicators evaluated for a given position of the sequence, such as the frequency of an amino acid or amino acid subgroup (expressing the degree of "conservation" during evolution), the propensity of a certain amino acid to support certain structural or functional demands, and the evolutionary or functional "distance" between amino acids. In most cases, a sum of indicator weights over all positions (a "score") or similar statistic is defined for this evaluation. A learning stage (during which weights are assigned) may be distinguished from an application phase (when diagnostic "scores" for candidates are

RBSK_ECOLI
MQNAGSLVVLGSINADHILNLQSFPTPGETVTGNHYQVA**FGGKGA NQAV**AAGRSGANIAFIAC**TGDDS**IGESVRQQ**LATDNID**ITPVSVIKG**ESTGV**ALIFV**NGEGENVIG**IHAG
ANAALSPALVEAQRERIANASALLMQLESPLESVMAAAKIAHQNKTIVALNPAPARELPDELLALVDIITPNE**TEAEKL**TGIRVEN**DEDAAK**AA**QVLHE**KGIRT**VLITL GSRGV**W
ASVNGEG**QRVP**GFRVQ**AVDTI AAGDTFN GALITAL**LEEKPLPEAIRFAHAAAAIAVTRKGAQPSVPWREEIDAFLDRQ

LACC_STAAU
MILTLTLNPSVDISYPLTALKLDDVNRVQEVSKT**AGGKGL NVTR**VLAQVGEPVLASGF**IGGEL**GQFIAKK**LDHADIK**HAFYNIKGE**TRNCI**AILH**EGQQTEILE**QGPEIDNQEAA
GFIKHFEQLLEKVEAVAISGSLPKGLNQDYYAQIIERCQNKGVPVILDCSGATLQTVLENPYKPTVIK**PNISEL**YQLLNQPL**DESLES**LK**QAVSQ**PLFEGIEW**IIVSL GAQGA**FA
KHNHTF**YRVN**IPTIS**VLNPV GSGDSTV AGITSAI**LNHENDHDLLKKANTLGMLNAQEAQTGYVNLNNYDDLFNQIEVL

SCRK_SALTY
MNAKVWVLGDAVVDLLPESEGRLLQC**PGGAPA NVAV**GVARLGGNSGFIGA**VGGDP**FGRYMRHT**LQQEQVD**VSHMYLDDQ**HRTST**VVVDL**DDQGERTFT**FMVRPSADLFLVEEDLP
QFAAGQWLHVCSIALSAEPSRSTTFAAMESIRSAGGRVSFDPNIRPDLWQDQALLLACLDRALHM**ANVVKL**SEEELVFISS**SNDLAY**GI**ASVTE**RYQPEL**LLVTR GKAGV**LAAFQ
QKF**THFN**ARPVA**SVDTT GAGDAFV AGLLASL**AANGMPTDMTALEPTLTLAQTCGALATTAKGAMTALPYQRDLNRQF

feature extraction :

```
1  FGGKGA NQAV-14-TGDDS-8-LATDNID-9-ESTGV-5-NGEGENVIG-77-TEAEKL-7- DEDAAK-2-QVLHE-5-VLITL GSRGV-8-QRVP-5-AVDTI  AAGDTFN  GALITAL
2  AGGKGL NVTR-14-IGGEL-8-LDHADIK-9-TRNCI-4-EGQQTEILE-79-PNISEL-8- DESLES-2-QAVSQ-8-IIVSL GAQGA-8-YRVN-5-VLNPV  GSGDSTV  AGITSAI
3  PGGAPA NVAV-14-VGGDP-8-LQQEQVD-9-HRTST-5-DDQGERTFT-82-ANVVKL-10-SNDLAY-2-ASVTE-6-LLVTR GKAGV-8-THFN-5-SVDTT  GAGDAFV  AGLLASL

    1       2       3       4       5       6       7       8       9       10      11      12      13      14      15
 BLO62   TAYLOR  PAM120  PAM50   BLO62   TAYLOR  PAM120  BACON   BLO62   TAYLOR BLO62   BLO75   PAM120  BLO62   BLO70

4 AAGKGI NVAK-14-LGKDN-7-FSELGIA-9-TRINV-4-KDGEVTDFN-85-PNRREL-10-MKDVIE-2-HALRE-5-VVISL GAEGA-10-PPSV-1-VVSTV GAGDSMV GGLIYGL
5 PGG.GI NVAR-14-AGGAT-7-LADENVP-9-TRQNL-5-ASGEQYRFV-82-PNQKEL-10-PDDVRK-2-QEIVN-6-VVVSL GPQGA-10-QVAL-1-SQSTV GAGDRLV GAMTLKL
6 PGGAPA NVAV-14-VGDDP-8-LAQEQVD-9-QRTST-5-DSHGERTFT-83-ADAIKL-10-SDDIVS-2-ARLNA-6-LLVTQ GKAGV-10-ARPV-1-AVDTT GAGDAFV AGLLAGL
7 PGGAPA NVAV-14-VGDDP-8-LAQEQVD-9-QRTST-5-DDQGERSFT-82-ADVVKF-10-STSMAQ-2-QQIAA-5-VLVTQ GAKGV-10-GQVV-1-PIDTT GAGDAFV GGLLACL
8 AGGTIG NTMH-17-SNIEG-8-LVIEDDV--------19-IGNTMHNYS-------------123---------------VLCTA GPIGL-56-GGPE-1-IMNTN GAGDGAL AALPDDI
```

Fig. 1. Recognition of ribokinase motif by an SOM parallel array. The three sequences at the top are learning sequences taken from the SWISSPROT database. The boldface sections are 15 features as extracted by Kohonen map analysis of the sequences. They are aligned in the middle part of the figure, and the weight matrices best representing them are stated. A SWISSPROT search (release 31) revealed the presence of 19 sequences, all recognized by the presence of the 15 features. Five of them are aligned as an example in the bottom part of the figure. Numbers between features indicate the residue length of intervening sections (IS) not typical and therefore not shown in detail. A point stands for a single gap in a position of a letter in the other sequences. At query sample 5, it is seen that even distantly related sequences (missing several features and having quite different ISs) may be recognized. Explanation of SWISSPROT acronyms: RBSK_ECOLI, ribokinase *Escherichia coli*; LACC_STAAU, 6-phosphogalactokinase *Staphylococcus aureus*; SCRK_SALTY, fructokinase *Salmonella typhimurium*; SCRK_KLEPN, ditto *Klebsiella aerogenes*; SCRK_VIBAL, ditto *Vibrio alginolyticus*; K1PF_ECOLI, 1-phosphofructokinase *E. coli*; K6P2-ECOLI, minor 6-phosphofructokinase *E. coli*; INGK_ECOLI, inosine guanosine kinase *E. coli*. Reference of the features to the PROSITE signatures (Bairoch, 1993) and to the conserved blocks of that family indicated by Bork et al. (1993, Fig. 5): signature 1, features 1 (except first residue), 2, and 3 (except two last residues), including IS; signature 2, features 13 (except first two residues), 14, and 15 (except last three letters); Bork block 1, features 1 and 2 (except last residue); Bork block 2, features 2 (one last residue) and 3 (first four residues), including IS; Bork block 3, features 10 (including four residues from IS to the left) and 11 (plus three residues from IS to the right); Bork block 4, features 13 (including three residues from IS to the left), 14, and 15 (first four residues). It is seen that initial chopping of the learning sequences for analysis ultimately leads to the synthesis of a multiple alignment of application sequences.

Fig. 2. Recognition of CUB features by an SOM array. Shown are the first three (of nine) features, labeled I–III, as extracted by the "feature extraction" network (see Fig. 3) from the learning set of five sequences. "Best matrix" was obtained by retraining in the "feature tuning" layer (Fig. 6) and is indicated below each feature. The bottom part of the figure emphasizes the feature occurrences detected in some "unknown" sequences. The whole set of neural maps may be envisaged as an array of signal lamps that flash when a common feature comes by. A pattern of flashing signals identifies a sequence belonging to the set of CUB domains.

calculated). The procedures require a multiple alignment of the learning set. Our results permit the general conclusion that a level hierarchy of Kohonen maps is able to solve this same task even in notoriously intricate cases. Hence, a well-trained neural network is as good a tool for the experimentalist as is a well-designed statistical procedure.

At first glance, the two principles – statistical arithmetics and neural computation – seem to be rather different. On closer examination, one sees items of similar information processing. The most important one is numerical encoding of sequence information (usually in the form of scores assigned to "window" sections). Furthermore, both approaches use the strategy of replacing the crude letter information by a vector of physicochemical properties and/or by a vector of numerical values representing the distance between amino acids. One of the best established arithmetic methods of sequence recognition is the program named PROFILE (Gribskov et al., 1987, improved as iterative optimization by Lüthy et al., 1994, and Thompson et al., 1994). It starts from a set of prealigned sequences and calculates, position by position, scores for any amino acid in a given place. The scores are derived from the position-specific frequency of amino acids and from similarity coefficients between them.

Our Kohonen map is able to process the same information, but in a different way. Instead of one scalar element, the whole vector of coefficients of an amino acid in a weight matrix is presented to the learning procedure. A Kohonen map may be trained with a very small training set (indeed, even from a single specimen). This would create problems in any frequency-dominated arithmetic procedure.

The Kohonen network learns and works with chopped, ungapped partial sequences. This principle is similar to that of methods like FAST (Pearson & Lipman, 1988: first stage, search of "tuples") or BLAST (Altschul et al., 1990, looking for significant common "segments"), and to the strategy of "block" representation of sequence families (Henikoff & Henikoff,

1991). It is obvious that such a strategy does not require prealignments, at least not in the first stage, and thus circumvents principal obstacles of sequence analysis in the presence of insertions and/or deletions. The price to be paid is additional consideration of number and position of ungapped features. In the Kohonen network, this necessitates higher levels of hierarchy beyond simple recognition of features.

We have compared (not shown here) our Kohonen map strategy with the more common feed-forward networks (see, for example, Frishman & Argos, 1992). It turned out that Kohonen hierarchies discriminate better, which may be caused by the topology-preserving learning principle of lateral inhibition in Kohonen networks (Ritter et al., 1992).

The claim of this paper is that suitably scaled Kohonen maps arranged in an appropriate architecture are able to memorize and recognize motifs and patterns in protein sequences with a quality that is not worse, and is often better, than that of established methods. The test cases of ribokinases and of CUB domains demonstrate the ability to discern related sequences by common short subsequences without previous alignment. The HTH example demonstrates the ability to be trained with a restricted learning set of a structural motif that is extremely fuzzy on the level of primary structure (see analyses of Lawrence et al., 1993, and Neuwald & Green, 1994). From a Kohonen map hierarchy as reported here, one may expect the following. (1) A detection performance occurs as in any other arithmetic procedure working on the same principle (blockwise detection, blockwise optimizing of mutation table, discarding out-of-block information as "noise." (2) An SOM is able, without special effort, to "subclassify" an inhomogenous family by seeing subclusters of features (which in several established procedures would be merged by averaging). (3) An SOM learns without prealignment on rather small learning sets. With such performance established, and given access to a massively parallel computer, the method should be applicable to search tasks of enormous volume and may thereby overcome, or at least alleviate, the well-known capacity limitation of sequence analysis against large databases. At present, we prepare a systematic application of the method to sets of superfamilies.

Materials and methods

We describe the Kohonen mapping strategy as applied to partial sequences and, later on, the hierarchic array of Kohonen maps for pattern recognition.

Feature recognition by a Kohonen map (SOM)

A feature is a set of characteristic ungapped partial sequences of a certain length that are close neighbors of each other according to a scoring measure. The Kohonen map is trained to recognize such features within a more or less numerous collection of partial sequences presented to it. The sequences are coded such that either distance or similarity between them is defined, and the feature specimens appear as a cluster of small mutual distance in the sequence space, whereas nonfeatures are at a large distance from that cluster.

The algorithmic details of the training strategy are described in the specialized literature (e.g., Hertz et al., 1991; Ritter et al., 1992). We sketch only the application to our task.

The basic model of the Kohonen map is outlined in Figure 3. For a certain set of codebook vectors, the signal space is uniquely mapped into domains of the pertinent neurons. "Learning" is an iterative process, during which all input signals are "shown" in turn and many times to the mapping, and each time the codebook vector of the pertinent neuron (the "winner") and those of its closest neighbors become updated to better memorize this signal (shifting the codebook vector somewhat closer to it). Over a prolonged cycle, this procedure results in the codebook vectors being in the center of the feature specimens around them.

As a result, we obtain a map of neurons whose codebook vectors populate the signal space such that densely populated input regions also have a dense representation on the map (in the sense that codebook vectors of neurons that are neighbors on the map have a small distance in the signal space), whereas in sparsely populated input regions the codebook vectors occur rarely (i.e., codebook vectors of neighboring neurons have a large distance).

A given feature emerges in this transformation as a set of input signals projecting onto the same neuron or its immediate neighbors on the map. Their distance from the pertinent codebook vector in the signal space (called "quantization error") is small. Any "nonmember" of the feature set will become projected either on neurons distant (on the map) from those of the feature (this occurs when relatives of it have been offered in the training set), or (if unprecedented) will be projected at random on the neuron that happens to be at minimum distance, but with large quantization error.

Throughout this study we have taken the Euclidean distance between signal vector and codebook vector as quantization error.

Recognition performance of a Kohonen map (SOM)

After the properly parametrized passage through the training iteration, the SOM recognizes the feature (projection with small

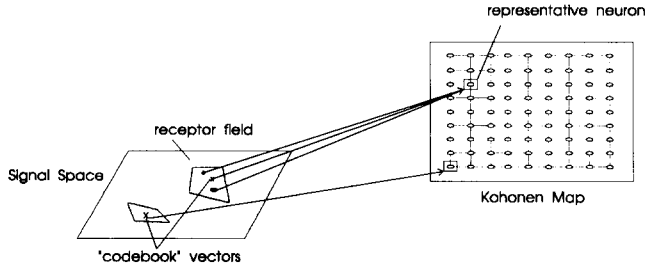

Fig. 3. Schematic illustration of signal space and projection onto the Kohonen map. Sequences are encoded such that they form a "signal"; that is, a vector of numerical values unique for that sequence. For any pair of sequences, a "distance" (or some analogon of it, such as a dissimilarity score) is defined. Kohonen mapping is the projection of the signal of any sequence onto a unique neuron of a two-dimensional layer of neurons (the "map"; see right-hand lattice). After training, a unique "codebook vector" is assigned to each neuron of the map. The "receptor field" of a neuron is the set of all sequence signals whose distance from the pertinent codebook vector is smaller than their distance from any other codebook vector. The receptor fields form polygonal patterns in the signal space (two such "honeycombs" are shown). During training, the position of codebook vectors and hence of the polygonal receptor fields are systematically changed in accordance with the features to be learned (see text).

error into the domain of the pertinent neuron) and distinguishes it from nonfeatures (projection with large error and/or onto a distant neuron).

Kohonen's mapping is essentially a well-selected projection from a multidimensional signal space onto a two-dimensional lattice. It cannot "completely" preserve the topology of the input space, in the sense that orderings in the distance remain intact. All that can be ensured is that clusters of signals remain clusters on the map. A codebook vector of a cluster is in some sense its "center"; hence, previously unknown relatives of the cluster will be recognized as well as the training specimens themselves.

The number of neurons on the map is of importance for the performance. Too few neurons (compared to the set of sequences) will allow only a very coarse classification. Too many neurons will create a tendency to "learn by heart"; that is, to memorize only one individual sequence per trained neuron and to miss close relatives.

The iteration procedure has a number of parameters (adaption step size, learning radius on the map). They have to be suitably varied during the training (see Ritter et al., 1992).

Coarse classification vs fine tuning of feature recognition

By a suitable choice and change of the iteration parameters as well as of the training sample, one may construct a Kohonen map on which more than one feature is distinguishable by projection into different domains of the lattice ("feature distinction"), or, alternatively, a different type of map that recognizes only one feature (against the universe of nonfeatures) but with great precision of detail and ability to distinguish "subfeatures" ("fine tuning of the feature").

Encoding of input sequences

The basic item presented as the input signal is the contents of a "window" of width W, that is, a sequence of W amino acids. The amino acid appearing in a certain position is encoded as a vector of A elements, either of chemical properties (in binary notation, as in Taylor's evaluation scheme [1986]), or of numerical values. These values may be property coefficients (expressing, for example, hydrophobicity or bulkiness) or scores taken from a weight matrix (like PAM or BLOSUM) (Dayhoff et al., 1978; Henikoff & Henikoff, 1992). In our application of such matrices, the column vector of substitution scores belonging to one amino acid is taken as the signal element of that position, so that the whole W-word is represented by an image consisting of $(W \times A)$ elements arrayed in matrix form, and introduced into the computation as a vector of concatenated columns of that matrix. 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. Such a coding procedure defines a convenient distance metric as a Euclidean distance between two signals.

The hierarchy of feature maps

Kohonen maps are arrayed in parallel, each one for a different feature, and then assembled in hierarchic order to perform different recognition stages (see Figure 4 for a schematic view). The processing of the sequence is thus divided into several conceptual stages, which are described in the following sections.

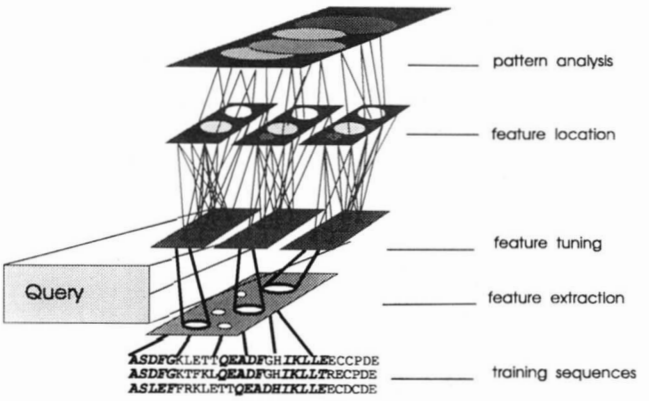

Fig. 4. The feature-map model. The flow of information through a set of Kohonen maps arrayed in hierarchical order is shown. Each Kohonen map is displayed as a two-dimensional area where individual neurons aggregate into neighborhood regions (ellipsoids) representing "features." Lines between the layers symbolize the connection weights that perform the projection from the respective input onto the map. During training, these connections obtain numerical weights. The input is illustrated by a set of unaligned training sequences. The features are set in bold for clarity, but their character is unknown during the training phase. The first (bottom) network learns to recognize typical n-words in the chopped (partial) sequences. In the second "tuning" layer, each Kohonen map deals exclusively with the specimens assigned to one selected feature to find the optimal substitution matrix leading to the most compact cluster. The third layer studies the location and repetition pattern of the features. The uppermost layer is trained to integrate information on one pattern composed of several features in a specified number and location. It is possible to instruct it to recognize more than one pattern, provided the number of neurons is sufficient. Recognition of a family means projection of all signals into one region of the Kohonen map. Application (after training) means that query sequences enter into the character-recognizing (second) layer, go through the upper layers, and become projected into the corresponding family locus.

Step 1: Feature extraction method

In the first step, features need to be extracted from a collection of candidate sequences. The criterion for the definition of a feature is that certain specimens form a cluster in the space of sequences; that is, that their mutual distance is small (or, equivalently, their mutual similarity score large) when compared with samples of nonfeatures.

Finding features in a collection of sequences is the task to be solved by Kohonen maps. We present all continuous subsequences of a specified length (or range of lengths) to the map. After training, the projection algorithm makes it possible to detect typical features that occur frequently in the set. As the chopping produces overlapping words, and as the word length is allowed to vary, a selection procedure follows that defines the most compact feature (in terms of quantization error). Figure 5 shows an example where features are allowed to have different lengths (optimized by trial). Table 1, by contrast, gives an example where the sequences have been cut into nonoverlapping pieces of specified (likewise optimized) length. This approach may be preferable in cases where the alignment problem does not exist or has been solved previously.

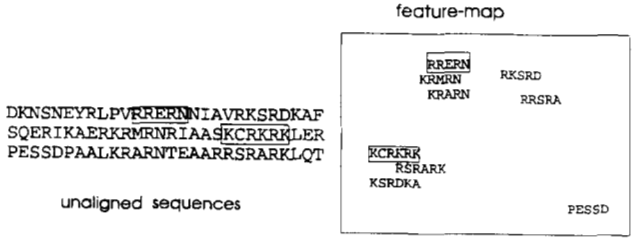

Fig. 5. Feature extraction. The manner in which the subsequences of each training sequence are projected onto corresponding vicinities of the feature map (where individual neurons have not been indicated) is illustrated. The procedure isolates two clusters whose optimal compactness is reached when they are presented as 5-words or 6-words, respectively. The former is the RRERN-cluster, the latter the KCRKRK-cluster (see boxes). They have very similar relatives in each of the three sequences in similar position. Each cluster is projected into one neuron. The quantization error for the RRERN-cluster is 0.67, that of the KCRKRK-cluster 0.78, indicating a tight cluster. There are two examples of 5-words, RKSRD and RRSRA, which are somewhat less related to the cluster and are in different locations in their sequences, and which become projected at some distance from the RRERN-cluster. The quantization errors of RKSRD and RRSRA are 1.3 and 1.55, respectively, indicating that they do not form a tight cluster. The 5-word PESSD is an example of a word that is very distant from all the features so far considered and is consequently projected into the corner. Its quantization error is 2.25, indicating a "maverick." The quantization error is a mean distance between sets of sequences. It stems from the representation of amino acids as vectors. An evaluation with similar conclusions might be done with, for example, BLOSUM45 scores. According to this method, the RRERN-group has a mean intracluster score of 22; the cluster members toward RKSRD have a mean score of 16; and those toward RRSRA have a mean score of 17; whereas the mean score to the distant PESSD is −4. Note that this projection is independent of any alignment. It recognizes small "blocks" and analyzes their occurrence and location.

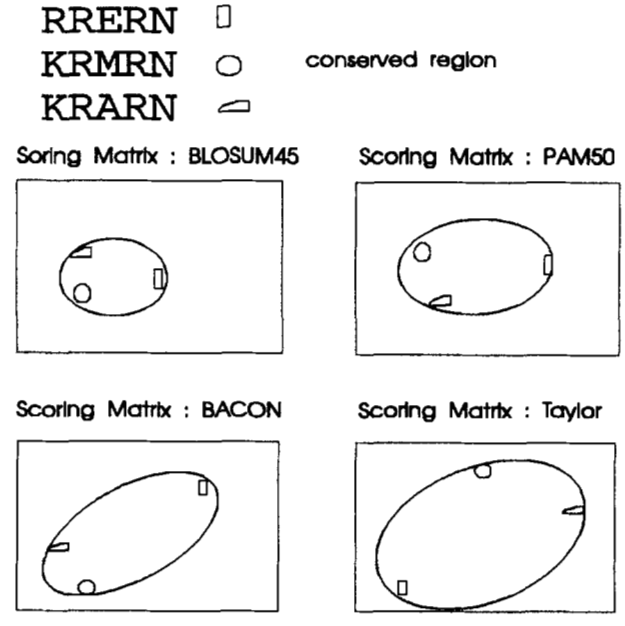

Fig. 6. Feature tuning. One of the features from the previous figure is isolated and processed in detail by subjecting its members to a comparative retraining with different weight matrices (the results of four of them are shown). It is seen that the set may be represented in the most compact and homogenously distributed form by retraining with the BLOSUM45 matrix. We have found that such fine tuning of the feature improves the performance of the later diagnostic process.

Step 2: Feature tuning

Once characteristic features have been identified in the set of chopped sequences, each of them is presented to one SOM of the second stage that specializes in it (Fig. 6). In detailed studies, it has turned out that "feature clouds" of related words become more homogenous and "compact" and, hence, easier to recognize when the similarity criterion is optimized. This effect depends on the biological character of the motif or even of the individual feature of the motif: sometimes mutability may be the appropriate criterion; in other cases, similarity of physicochemical milieu, structural features of residues, or some other feature is decisive. These different principles are encoded in the published weight matrices (e.g., Dayhoff et al., 1978; Bacon & Anderson, 1986; Taylor, 1986; Niefind & Schomburg, 1991; Henikoff & Henikoff, 1992). The SOMs are trained with each in turn, and the "best" criterion is selected by the smallest average distance (quantization error) of the feature set. Furthermore, an iterative procedure tests whether a shortened section of the feature (down to 4-words) yields a more compact representation (i.e., smaller quantization error). If all members of the feature are very similar, then all of them will be projected to the same "winner" or "center of gravity" neuron, but the refined training may likewise yield several centers of gravity; that is, several "subfeatures" located as "images" in different neurons.

At the end of this stage, the "best" weight matrix and the "best" feature length has been determined, and there is a trained subnetwork that is able to report that an input word belongs to that feature. This report indicates the "winner" neuron to which the candidate is most similar and states the distance to it. A dissimilar candidate is at once recognized by a high quantization error or by projection into a distant neuron domain.

Step 3: Feature location

To this point, features have been studied without regard to their location in the learning set. In the third stage, SOMs are again specialized to one feature and are trained to memorize occurrence and position in the mother sequences. In simpler cases, features occur once in a sequence, and in prescribed order. But the network is also able to learn that different combinations of features may occur in different compositions and orders.

Step 4: Pattern analysis

At this stage, the information has been integrated. Now the final, "master" SOM looks at whether the number and position range of features reported for a query sequence are correct. The signal from the query is projected to the closest neighbor on the final map and may therefore be used for diagnostic purposes. Decisions may alternatively be taken after inspection of the Kohonen map on a screen graph. If the capacity of the network (number of neurons) is sufficient, then more than one domain or pattern may be discriminated by the same hierarchic array.

Selection of learning sets

The learning set for HTH detection was adopted from previous studies (see text). In the case of ribokinase and CUB families,

we selected the learning sequences at random (three and five specimens, respectively) from the set of family members as present in SWISSPROT.

Any learning sequence was chopped into all (ungapped, neighbors overlapping) segments of length W. The value of W was started at a reasonable upper boundary (13 residues), then reduced stepwise. The "best" value for a feature is obtained when the mean quantization error in the neurons is minimal (mostly around 4–8 residues).

This optimization was repeated for different evaluation tables of amino acids. This resulted in a "best" W and a "best" similarity table for each feature.

Application and selection of diagnostic criteria

During application, both criteria (W and best table) were fixed for each feature. Whether or not a feature is present in a query window is decided by its individual q.e.: the appropriate cut-off value is a matter of choice by the user.

A default criterion was constructed in the following way. The longest individual sequence from the database that certainly did not contain the feature was selected, and a set of "nonmembers" was obtained from all possible (overlapping) windows of that sequence. Then the individual quantization errors after training was recorded for all "members" of the training sets as well as from all nonmembers, and was arranged by magnitude in a list. If this array separated members from nonmembers without mixing, then the average between the "worst" q.e. of members and the "best" q.e. of nonmembers served as default cut-off.

If both q.e. sets would overlap, then the user has to decide which type of error (number of false positives or number of false negatives) is to be minimized, or, otherwise, which compromise is to be adopted. This selection depends on the specific goal of the search.

Evaluation of the training result

The ratio of the mean distance of training vectors from their codebook vector after training to the mean distance of training vectors from their center-of-gravity vector before training serves as an evaluation measure. For good learning, this measure should be considerably smaller than unity (e.g., 0.1). However, approaching zero means "learning by rote" and is to be avoided, either by reducing the network dimension or by reducing the final "learning radius" of the algorithm.

Programs

The hierarchic network system (called an SOM) is a set of programs and subroutines. They are implemented on a cluster of 5 Sparc 10 stations (i.e., in parallel). The resulting program is quick for both stages—training and execution—the time consumption being proportional to the number of networks and to the dimension of the database. A search through SWISSPROT and PIR (combined) for the HTH motif as described in the paper takes 1.5 h.

The programs are written in ANSI-C language and may be run on UNIX and MS-DOS machines. We are preparing a version accessible via ftp-server (ftp ftp.mdc-berlin.de) and WWW-server (www.mdc-berlin.de).

Very extensive computations were done on the MasPar computer of the university of Stuttgart (IPVR).

Acknowledgments

J.H. was supported by the research program "Neurogen" of the Federal Ministry of Research and Technology (FRG). We are grateful to Prof. P. Levi, Prof. A. Zell, and to Dr. H. Bayer, for making the simulator core available and for continuous help and discussion.

References

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

Anderson WF. 1992. The helix-turn-helix motif and the cro repressor. In: Taylor WR, ed. *Patterns in protein sequence and structure. Springer Series Biophysics (7).* Berlin, Heidelberg, New York: Springer Verlag. pp 85–96.

Arrigo P, Giuliano F, Scalia F, Rapallo A, Damiani G. 1991. Identification of a new motif on nucleic acid sequence data using Kohonen's self-organizing map. *Comput Appl Biosci 7*:353–356.

Bacon DJ, Anderson WF. 1986. Multiple sequence comparisons. *J Mol Biol 191*:153–157.

Bairoch A. 1993. The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucleic Acids Res 21*:3097–3103.

Bairoch A, Boeckmann B. 1993. SWISSPROT protein sequence data bank, recent developments. *Nucleic Acids Res 21*:3093–3096.

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

Bohr H, Bohr J, Brunak J, Cotterill R, Lautrup B. 1988. A novel approach to prediction of three-dimensional structures of protein backbones by neural networks. *FEBS Lett 241*:223–228.

Bork P, Beckmann G. 1993. The CUB domain. A widespread module in developmentally regulated proteins. *J Mol Biol 231*:539–545.

Bork P, Sander C, Valencia A. 1993. Convergent evolution of similar enzymatic function on different protein folds: The hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Sci 2*:31–40.

Branden C, Tooze JT. 1991. *Introduction to protein structure.* New York: Garland Publishing.

Brennan RG, Matthews BW. 1989. The helix-turn-helix DNA binding motif. *J Biol Chem 264*:1903–1906.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, ed. *Atlas of protein sequence and structure. vol V (suppl 3).* pp 345–352. Washington, DC: National Biomedical Research Foundation.

Dodd IB, Egan JB. 1988. The prediction of helix-turn-helix DNA-binding regions in proteins. A reply to Yudkin. *Protein Eng 2*:174–176.

Dodd IB, Egan JB. 1990. Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic Acids Res 18*:5019–5026.

Ferrán EA, Ferrara P. 1991. Topological maps of protein sequences. *Biol Cybern 65*:451–458.

Frishman D, Argos P. 1992. Recognition of distantly related protein sequences using conserved motifs and neural networks. *J Mol Biol 228*:951–962.

Gribskov M, Lüthy R, Eisenberg D. 1990. Profile analysis. *Methods Enzymol 183*:146–159.

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

Henikoff S, Henikoff JG. 1991. Automated assembly of protein blocks for data base searching. *Nucleic Acids Res 19*:6565–6572.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci 89*:10915–10919.

Hertz J, Krogh A, Palmer RG. 1991. *Introduction to the theory of neural computation.* Redwood City: Addison Wesley.

Higgins DG, Bleasby AJ, Fuchs R. 1992. Improved software for multiple sequence alignment. *Comput Appl Biosci 8*:189–191.

Hirst JD, Sternberg MJE. 1991. Prediction of ATP-binding motifs: A comparison of a perceptron-type neural network and a consensus sequence method. *Protein Eng 4*:615–623.

Kohonen T. 1989. *Self-organization and associative memory.* Berlin: Springer Verlag.

Kohonen T. 1990. The self-organizing map. *Proc IEEE 78*:1464–1480.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science 262*:208–213.

Lüthy R, Xenarios I, Bucher P. 1994. Improving the sensitivity of the sequence profile method. *Protein Sci 3*:139–146.

McGregor MJ, Flores TP, Sternberg M. 1989. Prediction of β-turns in proteins using neural networks. *Protein Eng 2*:521–526.

Neuwald AF, Green P. 1994. Detecting patterns in protein sequences. *J Mol Biol 239*:698–712.

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

Pabo CO, Sauer RT. 1992. Transcription factors: Structural families and principles of DNA recognition. *Annu Rev Biochem 61*:1053–1095.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence analysis. *Proc Natl Acad Sci USA 85*:2444–2448.

Qian N, Sejnowski TJ. 1988. Prediction of the secondary structure of globular proteins using neural network models. *J Mol Biol 202*:865–884.

Ritter H, Martinetz Th, Schulten K. 1992. *Neural networks: An introduction to the neural informations process of self-organized networks.* Bonn: Addison-Wesley.

Rost B, Sander C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins Struct Funct Genet 19*:55–72.

Rumelhart DE, McClelland JL. 1986. *Parallel distributed processing: Exploration in the microstructure of cognition, vol I: Foundations.* Cambridge, Massachusetts: MIT Press.

Smith HO, Annau TM, Chandrasegaran S. 1990. Finding sequence motifs in groups of functionally related proteins. *Proc Natl Acad Sci USA 87*:826–830.

Tatusov RL, Altschul SF, Koonin EV. 1994. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci USA 91*:12091–12095.

Taylor WR. 1986. Identification of protein sequence homology by consensus template alignment. *J Mol Biol 188*:233–258.

Thompson JD, Higgins DG, Gibson TJ. 1994. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput Appl Biosci 10*:19–29.

Treisman J, Harris E, Wilson D, Desplan C. 1992. The homeodomain: A new face for the helix-turn-helix? *BioEssays 14*:145–150.

Wallace JC, Henikoff S. 1992. PATMAT: A searching program for sequence, pattern and block queries and databases. *Comput Appl Biosci 6*:249–254.

Yudkin MD. 1987. The prediction of helix-turn-helix DNA-binding regions in proteins. *Protein Eng 1*:371–372.

Yudkin MD. 1988. The prediction of helix-turn-helix DNA-binding regions in proteins: Concluding comment. *Protein Eng 2*:175–176.