Protein Science (1994), *3*:507-521. Cambridge University Press. Printed in the USA.

Self-organized neural maps of human protein sequences

EDGARDO A. FERRÁN,[1] BERNARD PFLUGFELDER,[2] AND PASCUAL FERRARA[1]

[1] Sanofi Elf Bio Recherches, Labège Innopole, BP 137—31676 Labège Cedex, France
[2] Societé Nationale Elf Aquitaine, 26, Avenue des Lilas, 64018 Pau, France

(RECEIVED May 18, 1993; ACCEPTED November 22, 1993)

Abstract

We have recently described a method based on artificial neural networks to cluster protein sequences into families. The network was trained with Kohonen's unsupervised learning algorithm using, as inputs, the matrix patterns derived from the dipeptide composition of the proteins. We present here a large-scale application of that method to classify the 1,758 human protein sequences stored in the SwissProt database (release 19.0), whose lengths are greater than 50 amino acids. In the final 2-dimensional topologically ordered map of 15 × 15 neurons, proteins belonging to known families were associated with the same neuron or with neighboring ones. Also, as an attempt to reduce the time-consuming learning procedure, we compared 2 learning protocols: one of 500 epochs (100 SUN CPU-hours [CPU-h]), and another one of 30 epochs (6.7 CPU-h). A further reduction of learning-computing time, by a factor of about 3.3, with similar protein clustering results, was achieved using a matrix of 11 × 11 components to represent the sequences. Although network training is time consuming, the classification of a new protein in the final ordered map is very fast (14.6 CPU-seconds). We also show a comparison between the artificial neural network approach and conventional methods of biosequence analysis.

Keywords: clustering algorithms; neural networks; protein classification; self-organized maps; sequence representations

During the last several years, the databases of macromolecular sequences have been continuously increasing. This growth has been accompanied by a sustained development of computer hardware and software. In particular, advanced computational tools to search for sequence similarities in macromolecular databases have been developed. There are powerful algorithms for comparing 2 (Needleman & Wunsch, 1970) or more sequences (Gribskov et al., 1987; Corpet, 1988). In general, these comparisons involve sequence alignments, allowing for the existence of gaps in each sequence. Although these methods are sensitive, they are extremely time consuming. Faster but less sensitive algorithms to identify related proteins have also been proposed (Lipman & Pearson, 1985; Altschul & Lipman, 1990; Altschul et al., 1990). In spite of these developments, the search for sequence similarities in macromolecular databases is still a subject of major concern, because sequencing data keep increasing at high speed as a consequence of many genome sequencing projects (Watson, 1990; Maddox, 1992; Sulston et al., 1992), and searching time, in standard algorithms, is usually proportional to the database size. One possible strategy for dealing with this problem is to cluster macromolecular sequences into families and then compare new sequences only with consensus patterns representing each family. This approach should not be limited by database size, because the number of macromolecular families is expected to grow more slowly than the number of sequences. Recently, 2 different neural-network-based methods following this approach have been proposed (Ferrán & Ferrara, 1991; Wu et al., 1992).

Artificial neural networks (ANNs) are simplified models inspired by the nervous system, in which neurons are considered as simple processing units linked with weighted connections called synaptic efficacies. These weights are gradually adjusted according to a learning algorithm.

ANNs have been applied as a computational tool to a large number of different fields. In most cases, a feed-forward architecture of the network is used to predict new correspondences of a relationship between inputs and outputs of the network, after "learning" some known examples of that relationship. The corresponding final set of synaptic connections is determined using a supervised learning algorithm: usually, the delta rule algorithm (Rosenblatt, 1962) for networks having only 1 layer of adaptable synaptic efficacies, and the backpropagation algorithm (Le Cun, 1985; Rumelhart et al., 1986) for multilayered networks. In particular, feed-forward ANNs have been applied to the analysis of biological sequences (Petersen et al., 1990;

Reprint requests to: Edgardo A. Ferrán, Sanofi Elf Bio Recherches, Labège Innopole voie 1—BP 137, 31676 Labège Cedex, France.

von Heijne, 1991; Hirst & Sternberg, 1992), considering some representation of the sequence as input to the network. For protein sequences, this approach has been used to predict immunoglobulin domains (Bengio & Pouliot, 1990), surface exposure of amino acids (Holbrook et al., 1990), disulfide-bonding states of cysteines (Muskal et al., 1990), signal peptides (Ladunga et al., 1991), ATP-binding motifs (Hirst & Sternberg, 1991), water-binding sites (Wade et al., 1992), and 3-dimensional (Bohr et al., 1990) and secondary structures of proteins (Bohr et al., 1988; Qian & Sejnowski, 1988; Holley & Karplus, 1989; McGregor et al., 1989; Andreassen et al., 1990; Kneller et al., 1990; Vieth & Kolinski, 1991; Muskal & Kim, 1992; Stolorz et al., 1992; Zhang et al., 1992; Rost & Sander, 1993a, 1993b). It has also been used to recognize distantly related protein sequences (Frishman & Argos, 1992). Concerning nucleic acid sequences, this approach has been used to predict DNA-binding sites or promoters (Stormo et al., 1982; Lukashin et al., 1989; Demeler & Zhou, 1991; O'Neill, 1991, 1992; Horton & Kanehisa, 1992), mRNA splice sites (Brunak et al., 1990, 1991; Engelbrecht et al., 1992), and coding regions in DNA (Lapedes et al., 1990; Uberbacher & Mural, 1991; Farber et al., 1992; Snyder & Stormo, 1993).

An alternative method to the supervised learning algorithm is the unsupervised one proposed by Kohonen (1982), where the neural network self-organizes its activation states into topologically ordered maps. These maps result from an information compression that only retains the most relevant common features of the set of input signals. This approach has been applied to detect signal peptide coding regions (Arrigo et al., 1991) and to cluster small organic molecules of analogue structure into families of similar activity (Rose et al., 1991). We have proposed a method based on Kohonen's algorithm to cluster protein sequences into families according to their degree of sequence similarity (Ferrán & Ferrara, 1991, 1992a). The network was trained using, as inputs, matrix patterns of 20 × 20 components derived from the dipeptide composition of the protein sequences. This naive representation of the whole sequence information has also been successfully applied to classify proteins with statistical techniques (Nakayama et al., 1988; Van Heel, 1991). Such representation allowed us to feed the network with a constant number of inputs, regardless of the protein length. Although network training is time consuming, once the topological map is obtained, the classification of a new protein is very fast. We have tested the method by considering both small (≈ 10 sequences) and large (≈ 450 sequences) learning sets of well-defined protein families (Ferrán & Ferrara, 1991, 1992a). For small learning sets, we have also shown that the trained network is able to classify correctly mutated or incomplete sequences of the learned proteins (Ferrán & Ferrara, 1991). We have also found, using a learning set of 76 cytochrome *c* sequences belonging to different species, that the time evolution of the map during learning roughly resembles the phylogenetic classification of the involved species (Ferrán & Ferrara, 1992b).

Wu et al. (1992) have recently proposed another neural-network-based method to classify protein sequences into families. The main difference between these 2 ANN approaches resides in the learning procedure: Wu et al. have used a supervised learning algorithm, whereas we have used an unsupervised one. They have trained several modules of a multilayered network using the backpropagation algorithm. Each module was trained with known examples taken from those sequences of the PIR database that have been previously identified as belonging to a given protein superfamily, using as inputs 1 or more "*n*-gram" encodings of the sequence. During learning, the synaptic efficacies between the neurons were changed in order to reduce a cost function. This function is a measure of the difference between the actual outputs provided by the network to each entry and the corresponding desired outputs, that is, the outputs encoding the correct superfamily classification. Because the desired output values to each entry must be known, this kind of learning algorithm is called supervised. On the contrary, in our method, it is not necessary to know a priori the number and composition of the protein families; thus, it could be used, for instance, to classify automatically the nonannotated entries of the PIR database.

The computing training time for the neural network with the unsupervised algorithm can be reduced by taking a smaller number of components in the input and synaptic vectors. In the present paper, we explore other matrix representations of the protein sequences in which all amino acids having similar physicochemical properties are considered as a same type of residue. We use both small and large learning sets to compare the classifications resulting from different sequence representations. In particular, we show a large-scale application of the method in which a learning set, probably as complex as the whole protein database as regards the clustering of the patterns, is considered: the set of all human protein sequences stored into the SwissProt database (release 19.0, 8/91). We also compare the results of the ANN approach with a standard statistical classification of the dipeptide matrices of this set of human protein sequences. In addition, we use a small set of protein sequences (the chemokine family) to compare the ANN approach with other conventional methods of biosequence analysis.

Results

Classification of small sets of well-defined protein families

Analysis of different matrix representations of the protein sequences

We investigated the influence of 4 different dipeptide matrix representations of the sequences on the protein classification: (1) 20 × 20 matrix. Each amino acid was taken as a different residue (representation that we have used in our previous works). (2) 11 × 11 matrix. Eleven groups of residues were considered: {V,L,I}, {T,S}, {N,Q}, {E,D}, {K,R,H}, {Y,F,W}, {M}, {P}, {C}, {A}, and {G}. The matrix representation of the sequence was built taking into account an alphabet of 11 symbols instead of 20. (3) 6 × 6 matrix. Six groups of residues were considered to build the matrix: {V,L,I,M} (hydrophobic), {Y,F,W} (hydrophobic, aromatic), {P,A,G,S,T} (neutral, weakly hydrophobic), {N,Q,E,D} (hydrophilic, acid), {K,R,H} (hydrophilic, basic), and {C} (crosslink forming). This grouping is the one considered by the GCG software package (Devereux et al., 1984) to determine the percentage of sequence similarity between 2 protein sequences according to the Needleman–Wunsch method. (4) 3 × 3. Three groups of residues were considered to build the matrix: {V,L,I,W,A} (hydrophobic), {Y,F,P,G,C,M}, and {N,Q,E, D,K,R,H,T,S} (hydrophilic).

The above protein representations were tested on a 7 × 7 network with 2 different learning sets of well-defined protein families. The first set consisted of 50 sequences belonging to the following 10 families (5 sequences per family): cytochromes *c*, hemoglobin α-chains, hemoglobin β-chains, insulins, growth hormones, prolactins, interferon-α precursors, *env* gene products (human immunodeficiency virus [HIV]-I subfamily), *pol* gene products (HIV-I subfamily), and *gag* gene products (HIV-I subfamily). The second learning set included 76 cytochrome *c* sequences that we used in a previous work (Ferrán & Ferrara, 1992a).

The topological map for each learning set and each protein representation was generated and then the number of proteins that could be considered wrongly classified was estimated (Table 1). For the cytochrome *c* learning set, we have considered that a pattern was wrongly classified if it did not fit the phylogenetic classification. This consideration gives an upper limit for the estimation of the number of wrong classifications because the degree of homology between protein sequences may be greater than the one expected by phylogenetics (Ferrán & Ferrara, 1992a). In Table 1 we have also indicated the CPU time required for the learning procedure (on a VAX 6310 computer, 2.67 MIPS) and the number of neurons assigned as winners in the final classification of the learned patterns. The CPU time decreases and the number of winner neurons increases when proteins are represented by matrices with a smaller number of components. Interestingly, the 11 × 11 representation led to a classification that was similar to the one obtained with the 20 × 20 representation, though the required computing time was reduced, as expected, by a factor of about 3.3.

Comparison of the ANN approach with conventional methods

In order to evaluate the ANN approach, we have compared the classification given by a topological mapping with the results obtained with other conventional methods of biosequence analysis. For this, we have used a set of inflammatory chemotactic cytokines known as the chemokine family (Table 2). This set can be divided into 2 subfamilies on the basis of 2 conserved cysteine residues that are either adjacent (CC or intercrine β family) or separated by 1 amino acid (CXC or intercrine α family). These cytokines are synthesized by a variety of cell types, usually in response to an inflammatory stimulus (Oppenheim et al., 1991), and their characteristic activity is leukocyte chemoattraction and activation, with a cellular selectivity for neutrophils (CXC members) or for monocytes (CC members) (Minty et al., 1993). The 2 subfamilies correspond to the well-characterized PROSITE groups SMALL_CYTOKINES_CC and SMALL_CYTOKINES_CXC. We have only considered complete sequences, except for the case of MCP-2 whose fragment was quite long (77 amino acids, whereas the lengths of the other 35 sequences were about 100 amino acids). All sequences but 1 (MCP-3) have been taken from the release 25.0 (4/93) of Swiss-Prot. The complete sequence of MCP-3 has been taken from Minty et al. (1993).

Table 1. *Number of wrong classifications (Wr), winner neurons (Nr), and computing time (Time, in CPU-minutes) required by the learning procedure on a VAX 6310 when different protein representations are used*[a]

	10 Families			Cytochromes *c*		
Representation	Wr	Nr	Time	Wr	Nr	Time
20 × 20	0	14	101.60	6	28	148.00
11 × 11	0	16	29.97	7	28	45.18
6 × 6	1	19	9.98	11	34	14.99
3 × 3	5	26	3.33	14	38	4.68

[a] In all cases, a network of 7 × 7 neurons was trained during 500 epochs, with an exponential decrease of α [$\alpha(0) = 0.90$, $a = 0.90$, $\Delta t_\alpha = 5$] and shrinking the winner neighborhood every $\Delta t_\upsilon = 50$ epochs. Both a learning set of 10 different families of proteins and another one composed of cytochromes *c* belonging to 76 different species were considered.

Both a 20 × 20 and an 11 × 11 matrix representation of the above-mentioned set of 36 protein sequences have been used to train a 5 × 5 neural network, during 30 learning cycles (or epochs). Figure 1 shows the corresponding final topological maps. In both cases, the neural network classified the 36 sequences in about 6 main groups of proteins having similar biological functions. In the map of Figure 1A there is only 1 case in which sequences belonging to the CC and CXC subfamilies were merged into the same neuron (sisd-h and the INIG proteins). Other unexpected classifications of this map, according to the description of the biological function given in SwissProt, were (1) sisf-m was placed far apart from the remaining SIS (or MIP-1) proteins; (2) emfi-ch was clustered with the IL-8 sequences; (3) mig-m was gathered together with the PLF-4 sequences; and (4) pf4l-h was not classified within the PLF-4 group. On the contrary, in the map of Figure 1B there was no merging of CC and CXC subfamilies, the sequence sisd-h was correctly classified within the SIS group, and the sequence pf4l-h was gathered together with the PLF-4 group. On the other hand, the sequence sisf-m was clustered within the MCP group, emfi-ch was placed in a neuron neighboring that of the IL-8 sequences, and mig-m was still classified in the PFL-4 group. Interestingly, all the sequences belonging to the CC subfamily were placed in the right side of the map, whereas those belonging to the CXC subfamily occupied the left part. Note that this remarkable separation of the CC and CXC subfamilies, which was initially defined on the basis of only 1 pair of amino acids, has been achieved using a dipeptide composition matrix to represent the protein sequence.

We have also classified the 36 protein sequences with several conventional methods of biosequence analysis. Figure 2 shows a dendrogram obtained from a multiple sequence alignment of those proteins, using a simplification of the progressive alignment method of Feng and Doolittle (1987), similar to the method described by Higgins and Sharp (1989) (PileUp command of the GCG software package). The dendrogram can be partitioned in nearly the same 6 main groups that were found with the ANN approach. Note also that most of the controversial points of the ANN approach correspond to cases of sequences having a big distance to the closer group of proteins (sisf-m, mig-m, emfi-ch, and pf4l-h). Table 3 shows the first 5 better scores obtained using the FastA algorithm proposed by Pearson and Lipman (1988), when each sequence was compared with the remaining 35 sequences (FastA command of the GCG package). Note that almost all of the results obtained with FastA and the ANN approach are consistent: the sequences are classified mainly in 6 groups; emfi-ch is in the IL-8 group; sisf-m is in the MCP group; and the closest sequence to mig-m is pf4l-h (though it is not clear to which of the 6 mains groups it belongs). In addition, though

Table 2. *List of 36 protein sequences belonging to the chemokine subfamilies CC or CXC and used to compare the ANN approach with other conventional methods*

Code	SwissProt name	Subfamily	SwissProt description
mcp1-b	mcpi_bovin	CC	Bovine monocyte chemotactic protein 1 precursor
mcp1-h	mcpi_human	CC	Human monocyte chemotactic protein 1 precursor
mcp1-m	mcpi_mouse	CC	Mouse monocyte chemotactic protein 1 precursor
mcp1-rb	mcpi_rabit	CC	Rabbit monocyte chemotactic protein 1 precursor
mcp1-r	mcpi_rat	CC	Rat monocyte chemotactic protein 1 precursor
mcp2-h	mcp2_human	CC	Human monocyte chemotactic protein 2 precursor (fragment)
mcp3-h	mcp3_human[a]	CC	Human monocyte chemotactic protein 3 precursor
sisa-m	mi1a_mouse	CC	Mouse macrophage inflammatory protein 1-α precursor
sisb-h	mi1a_human	CC	Human tonsillar lymphocyte LD78 α protein precursor
mi10-h	mi10_human	CC	Human tonsillar lymphocyte LD78 β protein precursor
sisc-h	mi1b_human	CC	Human T-cell activation protein 2 precursor
sisc-m	mi1b_mouse	CC	Mouse macrophage inflammatory protein 1-β precursor
sisd-h	sisd_human	CC	Human T-cell-specific RANTES protein precursor
sisf-m	sisf_mouse	CC	Mouse T-cell activation protein TCA3 precursor
inig-h	inig_human	CXC	Human interferon-γ-induced protein precursor
inig-m	inig_mouse	CXC	Mouse interferon-γ-induced protein CRG-2 precursor
mig-m	mig_mouse	CXC	Mouse γ-interferon-induced monokine
il8-h	il8_human	CXC	Human interleukin-8 precursor
il8-p	il8_pig	CXC	Pig interleukin-8 precursor
il8-rb	il8_rabit	CXC	Rabbit interleukin-8 precursor
emfi-ch	emfi_chick	CXC	Chicken embryo fibroblast protein 9E3 precursor
gro-c	gro_crigr	CXC	Chinese hamster growth regulated protein precursor
gro-h	gro_human	CXC	Human growth regulated protein precursor
gro-m	gro_mouse	CXC	Mouse platelet-derived growth factor-inducible protein KC precursor
gro-r	gro_rat	CXC	Rat cytokine-induced neutrophil chemoattractant
mip2a-h	mi2a_human	CXC	Human macrophage inflammatory protein-2-α precursor
mip2b-h	mi2b_human	CXC	Human macrophage inflammatory protein-2-β precursor
mip2-m	mip2_mouse	CXC	Mouse macrophage inflammatory protein 2 precursor
mip2-r	mip2_rat	CXC	Rat macrophage inflammatory protein 2 precursor
plf4-b	plf4_bovin	CXC	Bovine platelet factor 4
plf4-h	plf4_human	CXC	Human platelet factor 4 precursor
plfv-h	plfv_human	CXC	Human platelet factor 4 variant precursor
pf4l-h	pf4l_human	CXC	Human platelet basic protein precursor
plf4-p	plf4_pig	CXC	Pig platelet factor 4
plf4-r	plf4_rat	CXC	Rat platelet factor 4 precursor
plf4-s	plf4_sheep	CXC	Sheep platelet factor 4

[a] The mcp-3 protein sequence has been taken from Minty et al. (1993).

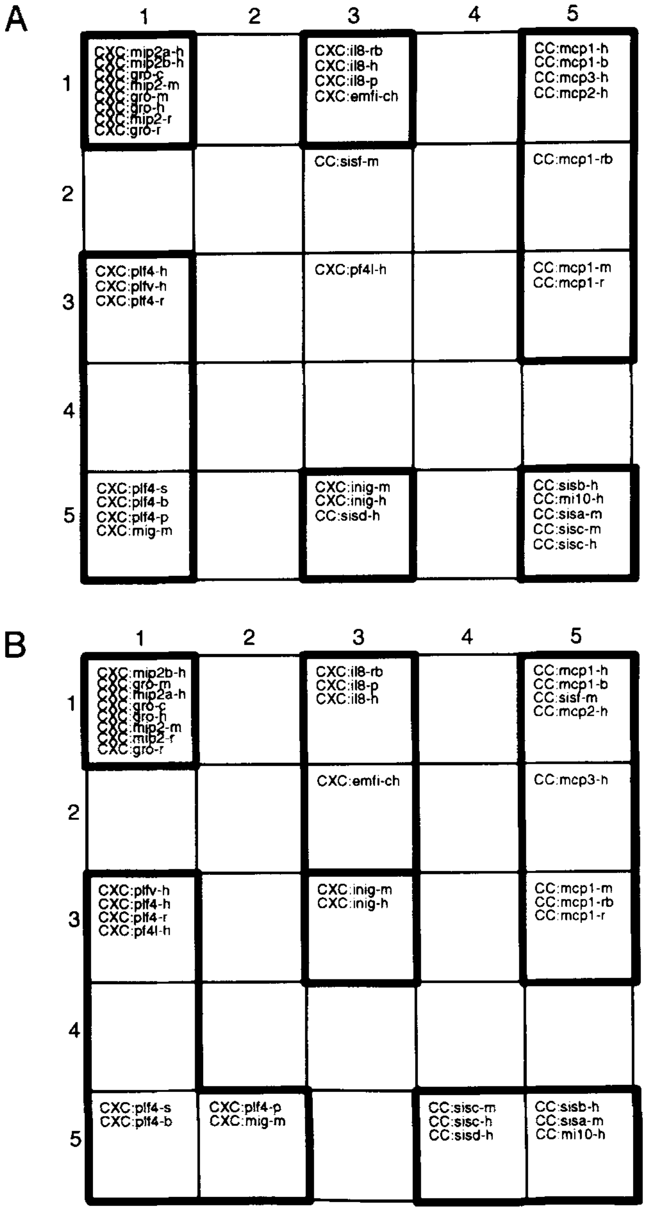
the 5 better scores for pf4l-h correspond to sequences of the GRO group, slightly lower scores were also found with the remaining sequences of the PFL-4 group (not shown). Finally, we have performed the alignments between each pair of the 36 sequences, using the method proposed by Needleman and Wunsch (Gap command of the GCG package). Table 4 shows the 5 better scores for each sequence. Note that, again, sisf-m and emfi-ch are classified within the MCP and IL-8 groups, respectively, and that mig-m and pf4l-h are not clearly classified in 1 group. In conclusion, the ANN approach is at least as effective as the above-mentioned conventional methods of biosequence analysis.

Classification of all known human proteins

Neural network classification—20 × 20 sequence representation

Slow learning protocol. The 1,758 human protein sequences stored into the SwissProt database (release 19.0, 8/91) were represented by the 20 × 20 dipeptide matrices and classified in a network of 225 neurons, $N_x = N_y = 15$ (see Fig. 3). A slow learning protocol, 500 epochs, that took about 100 CPU-hours (CPU-h) on a SUN 4/360 computer (16 MIPS, 2.6 MFLOPS), led to the topological map represented in Figure 4. Because the resulting classification is too long to be shown here, we only indicate the total number of sequences associated with each neuron and the location of some known families of proteins on the final map. The whole learning set was associated with 214 out of the 225 available neurons. There were 7.8 proteins per neuron (SD = 5.8). The average distance between protein patterns and the synaptic vector of the corresponding winner neuron was 0.6137. This value corresponds to 3.07% of the maximum possible distance d_m ($d_m = n$, where n^2 is the dimensionality of the synaptic space). In the case of the 20 × 20 protein representation, d_m is equal to 20 (for the 11 × 11 case, $d_m = 11$).

In most of the cases, sequences belonging to a given known family were placed into neighboring neurons. Table 5 shows that this is the case for actins, α-amylases, collagens, enolases, hap-



A

	1	2	3	4	5
1	CXC:mip2a-h CXC:mip2b-h CXC:gro-c CXC:mip2-m CXC:gro-m CXC:gro-h CXC:mip2-r CXC:gro-r		CXC:il8-rb CXC:il8-h CXC:il8-p CXC:emfi-ch		CC:mcp1-h CC:mcp1-b CC:mcp3-h CC:mcp2-h
2			CC:sisf-m		CC:mcp1-rb
3	CXC:plf4-h CXC:plfv-h CXC:plf4-r		CXC:pf4l-h		CC:mcp1-m CC:mcp1-r
4					
5	CXC:plf4-s CXC:plf4-b CXC:plf4-p CXC:mig-m		CXC:inig-m CXC:inig-h CC:sisd-h		CC:sisb-h CC:mi10-h CC:sisa-m CC:sisc-m CC:sisc-h

B

	1	2	3	4	5
1	CXC:mip2b-h CXC:gro-m CXC:mip2a-h CXC:gro-c CXC:gro-h CXC:mip2-m CXC:mip2-r CXC:gro-r		CXC:il8-rb CXC:il8-p CXC:il8-h		CC:mcp1-h CC:mcp1-b CC:sisf-m CC:mcp2-h
2			CXC:emfi-ch		CC:mcp3-h
3	CXC:plfv-h CXC:plf4-h CXC:plf4-r CXC:pf4l-h		CXC:inig-m CXC:inig-h		CC:mcp1-m CC:mcp1-rb CC:mcp1-r
4					
5	CXC:plf4-s CXC:plf4-b	CXC:plf4-p CXC:mig-m		CC:sisc-m CC:sisc-h CC:sisd-h	CC:sisb-h CC:sisa-m CC:mi10-h

Fig. 1. Topological maps for the **(A)** 20 × 20 and **(B)** 11 × 11 representations of 36 protein sequences belonging to the chemokine family (CC and CXC subfamilies). Thick line boxes separate sequences belonging to the 6 following main groups: monocyte chemotactic proteins (MCP); macrophage inflammatory protein 1 (or SIS proteins); interleukin-8 precursors (IL-8), interferon-γ-induced proteins (INIG); macrophage inflammatory protein 2 or growth regulated proteins (MIP-2/GRO); and platelet factor 4 (PLF4). Learning proceeded during 30 epochs, linearly decreasing α at each epoch ($\Delta t_\alpha = 1$), from a value of 0.9 ($a_1 = 0.08$ in the first 10 epochs and $a_2 = 0.0047619$ in the last 20) and decreasing the winner neighborhood every 6 epochs ($\Delta t_v = 6$).

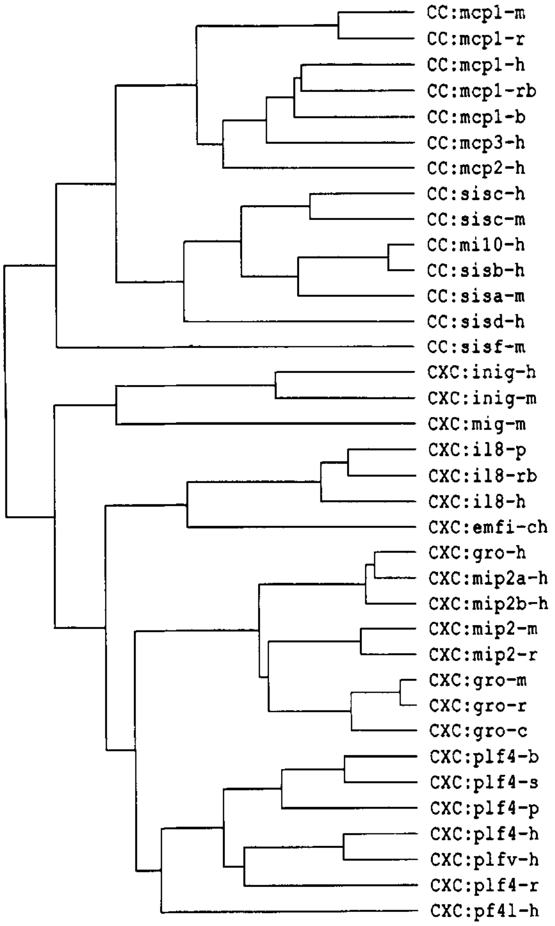


Fig. 2. Dendrogram of the 36 protein sequences belonging to the chemokine family (CC and CXC subfamilies).

toglobins, human leukocyte antigen (HLA) histocompatibility antigens (class I, class II α-chain, and class II β-chain subfamilies), all immunoglobulins but 1 (κ, λ, and heavy chains, variable and constant regions), interferon-α precursors, lamins, metallothioneins, myosins (light and heavy chains, tropomyosins), proline-rich proteins, tubulin β-chains, and zinc-finger proteins (note that there is only 1 line in Table 5 for each of these protein families or subfamilies). For example, zinc-finger proteins were grouped together into 5 neighboring neurons. Interestingly, inside of the "immunoglobulin-dominated" zone, placed in the lower left corner of the map, immunoglobulins were subclassified according to their type of chain or region (Table 6). Furthermore, in many of the above cases, all sequences of a same family were placed into a single neuron, e.g., all HLA class I histocompatibility antigens were clustered into neuron (1,1).

In a few cases, a family was split into subfamilies that were associated with neurons positioned far apart in the map. Table 5 shows that this is the case for keratins, hemoglobins, and G-protein coupled receptors. For example, hemoglobins were split into 2 subfamilies: 3 sequences were placed in neuron (3,1) and 4 in neuron (11,15).

In many cases, 2 or more different families of proteins were clustered together into the same neuron. For example, all actins and tubulin β-chains were associated with neuron (9,11). The percentage of sequence similarity between sequences of these 2 families, determined by the Needleman–Wunsch method, is about 39%. In fact, in most of the cases, sequences belonging to many different families are placed in the same neuron. This may indicate that there is a high percentage of sequence similarity among them. For example, desmin, desmoplakin, and vimen-

Table 3. *Five best matchings obtained with the FastA algorithm (Pearson & Lipman, 1988), when each protein belonging to the chemokine family is compared with the remaining 35 sequences*[a]

Protein	First match	Second match	Third match	Fourth match	Fifth match
mcp1-m (693)	mcp1-r (579)	mcp1-rb (315)	mcp3-h (305)	mcp1-h (296)	mcp1-b (293)
mcp1-r (688)	mcp1-m (579)	mcp1-rb (313)	mcp3-h (296)	mcp1-b (292)	mcp1-h (291)
mcp1-h (489)	mcp1-b (406)	mcp1-rb (406)	mcp3-h (395)	mcp1-m (296)	mcp1-r (291)
mcp1-rb (586)	mcp1-h (406)	mcp1-b (400)	mcp3-h (362)	mcp1-m (315)	mcp1-r (313)
mcp1-b (490)	mcp1-h (406)	mcp1-rb (400)	mcp3-h (371)	mcp1-m (293)	mcp1-r (292)
mcp3-h (506)	mcp1-h (395)	mcp1-b (371)	mcp1-rb (362)	mcp1-m (305)	mcp1-r (296)
mcp2-h (400)	mcp1-h (280)	mcp3-h (265)	mcp1-b (255)	mcp1-rb (250)	mcp1-m (246)
sisc-h (483)	sisc-m (413)	sisa-m (333)	mi10-h (296)	sisb-h (286)	sisd-h (262)
sisc-m (496)	sisc-h (413)	mi10-h (326)	sisa-m (323)	sisb-h (293)	sisd-h (264)
mi10-h (461)	sisb-h (421)	sisa-m (368)	sisc-m (326)	sisc-h (296)	sisd-h (258)
sisb-h (456)	mi10-h (421)	sisa-m (303)	sisc-m (293)	sisc-h (286)	sisd-h (267)
sisa-m (479)	mi10-h (368)	sisc-h (333)	sisc-m (323)	sisb-h (303)	sisd-h (252)
sisd-h (459)	sisb-h (267)	sisc-m (264)	sisc-h (262)	mi10-h (258)	sisa-m (252)
sisf-m (487)	mcp3-h (135)	mcp1-h (128)	mcp1-b (124)	mcp2-h (118)	mcp1-rb (118)
inig-h (481)	inig-m (341)	mig-m (126)	emfi-ch (110)	plf4-b (95)	il8-h (88)
inig-m (480)	inig-h (341)	mig-m (138)	emfi-ch (97)	il8-rb (81)	il8-p (81)
mig-m (621)	pf4l-h (146)	emfi-ch (146)	il8-p (139)	inig-m (138)	il8-h (134)
il8-p (514)	il8-rb (445)	il8-h (417)	emfi-ch (251)	pf4l-h (165)	gro-m (156)
il8-rb (514)	il8-p (445)	il8-h (429)	emfi-ch (207)	pf4l-h (173)	gro-m (172)
il8-h (503)	il8-rb (429)	il8-p (417)	emfi-ch (252)	gro-m (163)	gro-h (157)
emfi-ch (496)	il8-h (252)	il8-p (251)	il8-rb (207)	gro-c (195)	gro-m (188)
gro-h (479)	mip2a-h (448)	mip2b-h (428)	mip2-m (321)	gro-c (315)	mip2-r (306)
mip2a-h (489)	gro-h (448)	mip2b-h (428)	gro-c (344)	mip2-m (341)	mip2-r (325)
mip2b-h (480)	mip2a-h (428)	gro-h (428)	mip2-m (344)	mip2-r (327)	gro-c (306)
mip2-m (478)	mip2-r (430)	mip2b-h (344)	mip2a-h (341)	gro-h (321)	gro-c (307)
mip2-r (474)	mip2-m (430)	mip2b-h (327)	mip2a-h (325)	gro-h (306)	gro-c (305)
gro-m (463)	gro-c (364)	gro-r (338)	mip2a-h (315)	gro-h (296)	mip2-m (293)
gro-r (356)	gro-m (338)	gro-c (311)	mip2a-h (283)	gro-h (262)	mip2-r (262)
gro-c (481)	gro-m (364)	mip2a-h (344)	gro-h (315)	gro-r (311)	mip2-m (307)
plf4-b (434)	plf4-s (355)	plf4-h (265)	plf4-p (245)	plfv-h (232)	plf4-r (197)
plf4-s (406)	plf4-b (355)	plf4-h (271)	plf4-p (255)	plfv-h (247)	plf4-r (203)
plf4-p (354)	plf4-s (255)	plf4-b (245)	plf4-h (229)	plfv-h (221)	pf4l-h (196)
plf4-h (504)	plfv-h (405)	plf4-s (271)	plf4-b (265)	plf4-r (229)	plf4-p (229)
plfv-h (503)	plf4-h (405)	plf4-r (264)	plf4-s (247)	plf4-b (232)	plf4-p (221)
plf4-r (506)	plfv-h (264)	plf4-h (229)	plf4-s (203)	plf4-b (197)	plf4-p (195)
pf4l-h (599)	gro-c (230)	gro-h (227)	gro-r (220)	gro-m (219)	mip2-m (219)

[a] The corresponding *initn* scores are indicated in parentheses. The score shown in the first column corresponds to a matching between a sequence and itself.

tin, which have about 55% of sequence similarity with keratins and lamins, were clustered together with those sequences in neuron (1,13) (Table 7). Proteins placed in neighboring neurons may also have a high degree of sequence similarity. For example, 1 dystrophin sequence (≈46% of sequence similarity with lamins) was placed in neuron (2,13). In the same way, haptoglobins were clustered into neuron (11,14), close to α-amylases [neuron (11,13)] and to a cluster involving 3 enolases and 4 hemoglobins (≈42% of sequence similarity between haptoglobins and the other mentioned sequences).

Fast learning protocol. In a previous work (Ferrán & Ferrara, 1991), we have shown that, when the number of neurons of the network is greater than the number of families to be classified, fast learning protocols give more compact classifications (the network does not use all the available neurons to further subdivide protein families into subfamilies). To extend that study to cases in which the number of families is greater than the number of neurons, we have trained a 15 × 15 network with the same learning set of 1,758 human protein sequences as before, but using a fast learning protocol (30 epochs, instead of 500). The required computing time for this learning protocol was only 6.7 CPU-h (SUN 4/360 computer). Figure 5 shows that, as before, there were 11 empty neurons in the final map. The standard deviation of the number of proteins per neuron was similar to the value obtained before (SD = 5.593). The average distance between protein patterns and the synaptic vector of the corresponding winner neuron was 0.6775, which corresponds to 3.37% of d_m. This value shows that the final synaptic vectors were only slightly less representative of the associated protein patterns than before.

Figure 5 also shows the location on the map of the same known families of proteins previously analyzed. The overall aspect of the fast-learning classification is very similar to the one obtained with the slower protocol (as regards the way sequences are grouped on the map). For example, we also found regions

Table 4. *Five best matchings obtained with the Needleman-Wunsch algorithm (1970), when each protein belonging to the chemokine family is compared with the remaining 35 sequences*[a]

Protein	First match	Second match	Third match	Fourth match	Fifth match
mcp1-m (222)	mcp1-r (192)	mcp1-rb (111)	mcp1-b (99)	mcp1-h (95)	mcp3-h (94)
mcp1-r (222)	mcp1-m (192)	mcp1-rb (108)	mcp1-b (95)	mcp3-h (93)	mcp1-h (92)
mcp1-h (149)	mcp1-rb (119)	mcp1-b (119)	mcp3-h (116)	mcp1-m (95)	mcp1-r (92)
mcp1-rb (188)	mcp1-h (119)	mcp1-b (115)	mcp1-m (111)	mcp1-r (108)	mcp3-h (105)
mcp1-b (149)	mcp1-h (119)	mcp1-rb (115)	mcp3-h (108)	mcp1-m (99)	mcp1-r (95)
mcp3-h (149)	mcp1-h (116)	mcp1-b (108)	mcp1-rb (105)	mcp1-m (94)	mcp1-r (93)
mcp2-h (116)	mcp1-h (83)	mcp3-h (77)	mcp1-rb (74)	mcp1-b (73)	mcp1-m (70)
sisc-h (138)	sisc-m (113)	sisa-m (99)	mi10-h (99)	sisb-h (95)	sisd-h (81)
sisc-m (138)	sisc-h (113)	mi10-h (98)	sisb-h (93)	sisa-m (92)	sisd-h (80)
mi10-h (140)	sisb-h (132)	sisa-m (112)	sisc-h (99)	sisc-m (98)	sisd-h (81)
sisb-h (138)	mi10-h (132)	sisa-m (107)	sisc-h (95)	sisc-m (93)	sisd-h (83)
sisa-m (138)	mi10-h (112)	sisb-h (107)	sisc-h (99)	sisc-m (92)	sisd-h (79)
sisd-h (137)	sisb-h (83)	mi10-h (81)	sisc-h (81)	sisc-m (80)	sisa-m (79)
sisf-m (138)	mcp3-h (59)	mcp1-r (55)	mcp1-h (55)	mcp1-m (53)	mcp1-b (52)
inig-h (147)	inig-m (111)	mig-m (70)	emfi-ch (57)	gro-m (53)	gro-c (53)
inig-m (147)	inig-h (111)	mig-m (70)	gro-c (57)	emfi-ch (57)	gro-m (56)
mig-m (189)	inig-m (70)	inig-h (70)	gro-c (65)	il8-p (63)	mip2a-h (63)
il8-p (155)	il8-rb (133)	il8-h (122)	emfi-ch (90)	gro-c (73)	gro-m (71)
il8-rb (152)	il8-p (133)	il8-h (125)	emfi-ch (90)	gro-c (71)	pf4l-h (70)
il8-h (149)	il8-rb (125)	il8-p (122)	emfi-ch (91)	mip2-r (72)	gro-m (72)
emfi-ch (155)	il8-h (91)	il8-rb (91)	il8-p (90)	mip2-r (82)	mip2-m (77)
gro-h (161)	mip2a-h (149)	mip2b-h (146)	mip2-m (107)	gro-c (107)	mip2-r (101)
mip2a-h (161)	gro-h (149)	mip2b-h (147)	gro-c (114)	mip2-m (110)	gro-m (103)
mip2b-h (161)	mip2a-h (147)	gro-h (146)	mip2-m (111)	gro-c (108)	mip2-r (104)
mip2-m (150)	mip2-r (135)	gro-c (112)	mip2b-h (111)	mip2a-h (110)	gro-h (107)
mip2-r (150)	mip2-m (135)	gro-c (113)	mip2b-h (104)	mip2a-h (103)	gro-h (101)
gro-m (144)	gro-c (125)	gro-r (105)	mip2-m (104)	mip2a-h (103)	mip2-r (101)
gro-r (108)	gro-m (105)	gro-c (97)	mip2a-h (87)	mip2b-h (83)	mip2-m (82)
gro-c (152)	gro-m (125)	mip2a-h (114)	mip2-r (113)	mip2-m (112)	mip2b-h (108)
plf4-b (132)	plf4-s (111)	plf4-h (88)	plfv-h (82)	plf4-p (79)	plf4-r (77)
plf4-s (126)	plf4-b (111)	plf4-h (88)	plfv-h (83)	plf4-p (81)	plf4-r (79)
plf4-p (104)	plf4-s (81)	plf4-b (79)	plf4-h (76)	plfv-h (74)	pf4l-r (70)
plf4-h (152)	plfv-h (132)	plf4-r (109)	plf4-s (88)	plf4-b (88)	pf4l-h (83)
plfv-h (156)	plf4-h (132)	plf4-r (104)	plf4-s (83)	plf4-b (82)	pf4l-h (79)
plf4-r (158)	plf4-h (109)	plfv-h (104)	gro-h (81)	mip2a-h (81)	mip2b-h (80)
pf4l-h (192)	gro-h (86)	mip2a-h (83)	plf4-h (83)	gro-c (81)	mip2b-h (80)

[a] The corresponding scores are indicated in parentheses. The score shown in the first column corresponds to a matching between a sequence and itself.

of the map dominated by immunoglobulins (upper left corner) and zinc-finger proteins (upper right corner); all HLA class I histocompatibility antigens were again placed in only 1 neuron (lower left corner); and enolases were clustered together with the same 4 hemoglobins as before, but into 3 neighboring neurons [(6,10), (7,9), and (7,10)], instead of 1.

The main differences in Figure 5 with respect to Figure 4 are (1) 5 keratins are close to lamins (instead of 7); (2) G-protein coupled receptors are classified into 3 main clusters instead of 2; (3) the hemoglobin δ-chain (SwissProt name hbaz$human) is classified into neuron (7,7), far from its previous group of 3 sequences and closer to the group of 4 hemoglobins clustered with enolases; and (4) haptoglobins are placed apart [neuron (4,11)] from enolases and hemoglobins, but still close to α-amylases.

Neural network classification — 11 × 11 sequence representation

The classification of all the human proteins represented by the 11 × 11 matrix was performed, as before, in an $N_x = N_y = 15$ network, with either a slow (500 epochs, 27.4 CPU-h on the SUN 4/360) or a fast (30 epochs, 1.8 CPU-h on the same computer) learning protocol. In the final maps (Figs. 6, 7), the average distances between protein patterns and synaptic vectors of the corresponding winner neurons were 0.3637 and 0.4117, respectively, for the slow and fast protocols. These values correspond, respectively, to 3.31% and 3.74% of $d_m = 11$. In both maps there were only 5 empty neurons. The SDs of the number of proteins per neuron were, respectively, 4.423 and 5.077. Figures 6 and 7 show, respectively, the number of proteins associated with each neuron and the position on the maps of the same known families of proteins that we have analyzed for the 20 × 20 representation.

Slow learning protocol. The main differences in the map of Figure 6 (11 × 11 matrix), with respect to that of Figure 4 (20 × 20 matrix) are (1) all keratins are grouped together with lamins [neurons (11,15), (12,15)] and placed far apart from a group of 4 heterogeneous ribonucleoproteins [neurons (13,3),

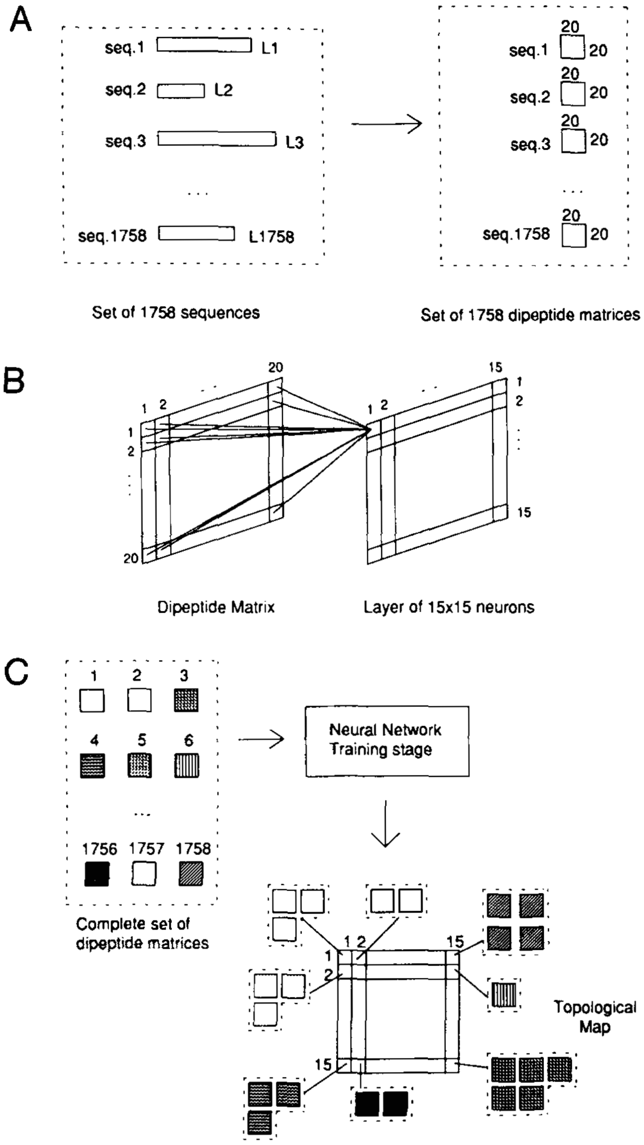
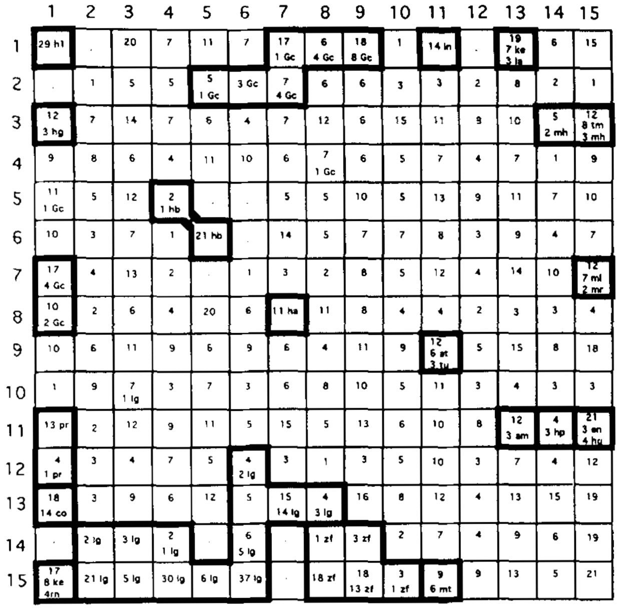


Fig. 3. Scheme of the neural network approach to classify the human protein sequences into families. **A:** The set of 1,758 sequences is transformed in a set of 1,758 dipeptide matrices. Note that, though sequences usually have different lengths Li (i = 1, 2, ..., 1,758), the dipeptide matrices always have 20 x 20 components. **B:** Each neuron of a 15 x 15 layer takes its inputs from the 400 components of a dipeptide matrix [only some of the connections to neuron (1,1) are shown]. **C:** As a result of Kohonen's algorithm, the complete set of dipeptide matrices is partitioned into several protein families, each one associated with a neuron of the 15 x 15 layer. The synaptic vector of a neuron may be considered as the cluster's centroid of the dipeptide matrices associated with that neuron. Related families of proteins are placed in neighboring spatial positions of the layer.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	29 hl	.	20	7	11	7	17 1 Gc	6 4 Gc	18 8 Gc	1	14 In	.	19 7 ke 3 la	6	15
2	.	1	5	5	5 1 Gc	3 Gc	7 4 Gc	6	6	3	3	2	8	2	1
3	12 3 hg	7	14	7	6	4	7	12	6	15	11	9	10	5 2 mh	12 8 tm 3 mh
4	9	8	6	4	11	10	6	7 1 Gc	6	5	7	4	7	1	9
5	11 1 Gc	5	12	2 1 hb	.	.	5	5	10	5	13	9	11	7	10
6	10	3	7	1	21 hb	.	14	5	7	7	8	3	9	4	7
7	17 4 Gc	4	13	2	.	1	3	2	8	5	12	4	14	10	12 7 ml 2 mr
8	10 2 Gc	2	6	4	20	6	11 ha	11	8	4	4	2	3	3	4
9	10	6	11	9	6	9	6	4	11	9	12 6 at 3 tu	5	15	8	18
10	1	9	7 1 Ig	3	7	3	6	8	10	5	11	3	4	3	3
11	13 pr	2	12	9	11	5	15	5	13	6	10	8	12 3 am	4 3 hp	21 3 en 4 hg
12	4 1 pr	3	4	7	5	4 2 Ig	3	1	3	5	10	3	7	4	12
13	18 14 co	3	9	6	12	5	15 14 Ig	4 3 Ig	16	8	12	4	13	15	19
14	.	2 Ig	3 Ig	2 1 Ig	.	6 5 Ig	.	1 zf	3 zf	2	7	4	9	6	19
15	17 8 ke 4rn	21 Ig	5 Ig	30 Ig	6 Ig	37 Ig	.	18 zf	18 13 zf	3 1 zf	9 6 mt	9	13	5	21

Fig. 4. Topological map for the 20 × 20 representation of 1,758 human protein sequences. We only indicate the number of sequences having each neuron as winner and the location of the following families of proteins: 6 actins (at); 3 α-amylases (am); 14 collagens/procollagens (co); 3 enolases (en); 29 G-protein coupled receptors (Gc); 29 HLA class I histocompatibility antigens (hl); 11 α-chains of HLA class II histocompatibility antigens (ha); 22 β-chains of HLA class II histocompatibility antigens (hb); 7 hemoglobins (hg); 3 haptoglobins (hp); 130 immunoglobulins (Ig); 14 interferon-α precursors (In); 15 keratins (ke); 3 lamins (la); 6 metallothioneins (mt); 7 myosin light chains (ml); 2 myosin regulatory light chains (mr); 5 myosin heavy chains (mh); 8 tropomyosins (tm); 14 proline-rich proteins (pr); 4 (out of 7) heterogeneous ribonucleoproteins (rn); 3 tubulin β-chains (tu), and 36 zinc-finger proteins (zf). Thick line boxes indicate main groups of sequences belonging to the same protein family. Learning proceeded during 500 epochs, linearly decreasing α at each epoch (Δt_{α} = 1), from a value of 0.9 (a_1 = 0.008 in the first 100 epochs and a_2 = 0.0002499 in the last 400) and decreasing the winner neighborhood every 30 epochs (Δt_{v} = 30).

(9,2)]; (2) all hemoglobins are clustered together into 2 neighboring neurons [6 sequences into neuron (6,15) and hbaz$human into neuron (5,15)], far apart from haptoglobins [neuron (10,12)] and enolases [neuron (10,13)]; (3) there are 2 separated clusters of α-chains of HLA class II histocompatibility antigens [7 sequences in neuron (1,6) and 4 in neurons (3,11), (4,11) and (5,12)]; (4) actins [neuron (5,13)] and tubulin β-chains [neuron (9,7)] are clustered far apart from each other; (5) haptoglobins [neuron (10,12)] are placed far apart from α-amylases [neuron (9,7)] and hemoglobins, but still close to enolases; (6) myosin light and heavy chains and tropomyosins are clustered in neighboring neurons [neurons (13,12), (13,13), (13,14), (13,15), (14,15), and (15,14)]; and (7) the 5 small proline-rich proteins are clustered apart from the other types of proline-rich proteins, but placed in a neighboring neuron [neurons (10,1) and (11,1)].

Fast learning protocol. The map obtained with the 11 x 11 matrix representation and the fast learning procedure (Fig. 7) does not essentially differ from that obtained with the slower one (Fig. 6). Differences between both of them are: the keratin family, which, in the map of Figure 7 was again split into 2 separated clusters [neurons (4,5) and (12,1)]; the α-chain sequences of HLA class II histocompatibility antigens that were placed in neighboring neurons [(7,13), (8,13), (9,13), and (10,14)]; and the G-protein coupled receptors that were clustered together (right border of the map).

Statistical classification

The set of 1,758 human proteins (20 x 20 protein representation) was classified using standard statistical methods. First,

Table 5. *Neural network clustering of known protein families (20 × 20 protein representation, 500 training epochs)*

	Number of neighboring		
Protein family	Proteins	Winner neurons	List of neighboring neurons
Actins	6	1	(9,11)
α-Amylases	3	1	(11,13)
Collagens	14	1	(13,1)
Enolases	3	1	(11,15)
G-coupled receptors	21	6	(1,7),(1,8),(1,9),(2,5),(2,6),(2,7)
	6	2	(7,1),(8,1)
	1	1	(5,1)
	1	1	(4,8)
Haptoglobins	3	1	(11,14)
Hemoglobins	4	1	(11,15)
	3	1	(3,1)
HLA histocompatibility antigens			
Class I	29	1	(1,1)
α-Chain, class II	11	1	(8,7)
β-Chain, class II	22	2	(5,4),(6,5)
Immunoglobulins			
λ-Chain, V-region	34	3	(14,4),(15,3),(15,4)
κ-Chain, V-region	48	3	(14,6),(15,6),(15,5)
Heavy chain, V-region	28	4	(14,2),(14,3),(15,2),(15,3)
	8	1	(13,7)
C-region	11	3	(12,6),(13,7),(13,8)
	1	1	(10,3)
Interferon-α precursors	14	1	(1,11)
Keratins	7	1	(1,13)
	8	1	(15,1)
Lamins	3	1	(1,13)
Metallothioneins	6	1	(15,11)
Myosins			
Light chains	9	1	(7,15)
Heavy chains	5	2	(3,14),(3,15)
Tropomyosins	8	1	(3,15)
Proline-rich proteins	14	2	(11,1),(12,1)
Tubulin, β-chains	3	1	(9,11)
Zinc-finger proteins	36	5	(14,8),(14,9),(15,8),(15,9),(15,10)

Table 6. *Classification of immunoglobulins into subfamilies on the map of Figure 4 (20 × 20 protein representation, 500 training epochs)*

Neuron	Associated immunoglobulin sequences
(15,6)	37 κ-Chains (24 V-I, 13 V-III)
(15,5)	6 κ-Chains (V-II)
(14,6)	5 κ-Chains (4 V-IV, 1 V-V)
(14,4)	1 λ-Chain (V)
(15,4)	30 λ-Chains (9 V-I, 11 V-II, 5 V-IV, 4 V-VI, 1 V-VII)
(15,3)	3 λ-Chains (V-III, V-V, V-VI); 2 heavy chains (V-I)
(15,2)	21 Heavy chains (V-III)
(14,2)	2 Heavy chains (V-I, V-II)
(14,3)	3 Heavy chains (V-I)
(13,7)	8 Heavy chains (V-II); 3 γ-chains (C); 2 α-chains (C); 1 ε-chain (C)
(13,8)	1 μ-Chain (C); 1 μ-heavy chain disease protein; 1 γ-chain (C)
(12,6)	1 κ-Chain (C); 1 λ-chain (C)
(10,3)	1 δ-Chain (C)

a principal component analysis (PCA) was performed, using the PRINCOMP procedure of the SAS package (SAS Institute, 1985, chapter 28). We have found that the first 60 principal components gave account of 70% of the inertia of the whole cloud of 1,758 "points" (corresponding to the representation of each protein as a point in the 400-dimensional vector space). Therefore, a reduced number of independent variables seems to be enough to provide a suitable sequence representation. Then we classified the set of 1,758 reduced patterns (built with the first 60 principal components of the dipeptide matrices) into 225 clusters, using the FASTCLUS procedure of the SAS package (SAS Institute, 1985, chapter 18). This procedure uses the nearest centroid sorting method (Anderberg, 1973), which is directly inspired by the leader (Hartigan, 1975) and k-means (MacQueen, 1967) algorithms. After 30 iterations, each 1 of the 1,758 human proteins was assigned to 1 of the 225 clusters. The numbers of iterations and clusters of the statistical method were chosen to coincide, respectively, with the numbers of epochs and neurons of the ANN approach in order to render both algorithms as similar as possible. In the resulting classification there was a great number of "clusters" (116) composed of only 1 protein and some

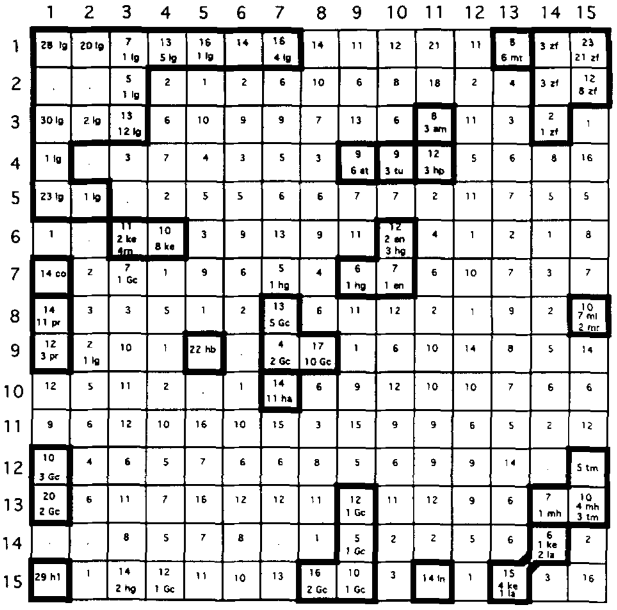
Table 7. *Protein sequences having neuron (1,13) of the map of Figure 4 as winner (20 × 20 protein representation, 500 training epochs)*[a]

SwissProt Name	*d*	Protein
lama$human	0.4364	Lamin a
lamc$human	0.4612	Lamin c
k2c7$human	0.4757	Keratin, type II cytoskeletal 7
desm$human	0.4830	Desmin
desp$human	0.4911	Desmoplakin I and II
lamb$human	0.5032	Lamin b
k1cr$human	0.5140	Keratin, type I cytoskeletal 18
vime$human	0.5361	Vimentin
k2c8$human	0.5415	Keratin, type II cytoskeletal 8
gfap$human	0.5563	Glial fibrillary acidic protein
k1cs$human	0.5636	Keratin, type I cytoskeletal 19
k1cm$human	0.6232	Keratin, type I cytoskeletal 13
k2c4$human	0.6386	Keratin, type II cytoskeletal 4
k2ca$human	0.6444	Keratin, type II cytoskeletal 56
ape$human	0.7420	Apolipoprotein E precursor
pdgb$human	0.7888	Platelet-derived growth factor, B chain
brn2$human	0.8073	Brain-specific homeobox/pou domain protein 2
pdga$human	0.8102	Platelet-derived growth factor, A chain
brn1$human	0.8163	Brain-specific homeobox/pou domain protein 1

[a] Sequences are sorted according to increasing distances between dipeptidic patterns and the synaptic vector of the neuron.

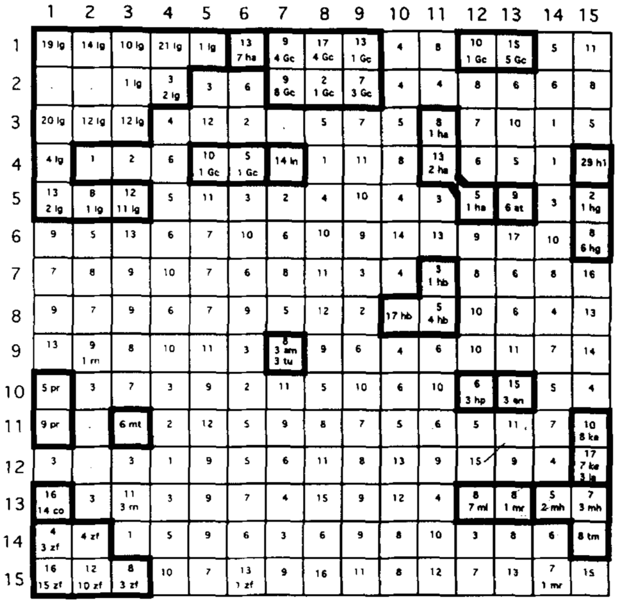
clusters with too many proteins (for instance, the biggest cluster was composed of 374 proteins). This distribution of clusters population ($\theta = \infty$ in Table 8, see below for the meaning of θ) corresponds to a poorly defined protein classification (from a biological point of view). Similar results (not shown) were also obtained for reduced patterns including more principal components and for a direct classification of the dipeptide matrices (i.e., without reducing the number of independent variables via the PCA step).

The extremely heterogeneous distribution of clusters population found with the statistical clustering method reflects the nonuniform distribution of protein sequences as points in the 400-dimensional vector space (or in a subspace of lower dimensionality, for the reduced representation) but clearly differs from the somehow even number of protein sequences associated with each winner neuron obtained with the ANN approach. It also reveals the existence of a great number of "singular" points or outliers. To obtain a cleaner classification with the statistical approach, these outliers should be discarded before performing the clustering stage. To this end, after the PCA step, we have considered as outliers all the protein patterns having at least 1 principal component with module greater than a threshold value θ. For $\theta = 1.5$, we have found 1,126 outliers. We have then classified the reduced patterns (60 components) of the remaining set of 632 nonsingular patterns into 225 clusters, using the FASTCLUS procedure (30 iterations). At the end, we have assigned the outliers to the nearest cluster centroid (center of gravity of the nonsingular patterns associated with that cluster). As it can be seen in Table 8, the new distribution of clusters population is clearly more uniform. There were only 13 clusters with 1 protein, and the biggest cluster only included 39 proteins. The resulting classification is also better than the previous one, from a biological point of view. For instance, the biggest cluster included 35 of 36 zinc-finger proteins (and other related proteins) and the cluster of 37 proteins included 30 immunoglobulin



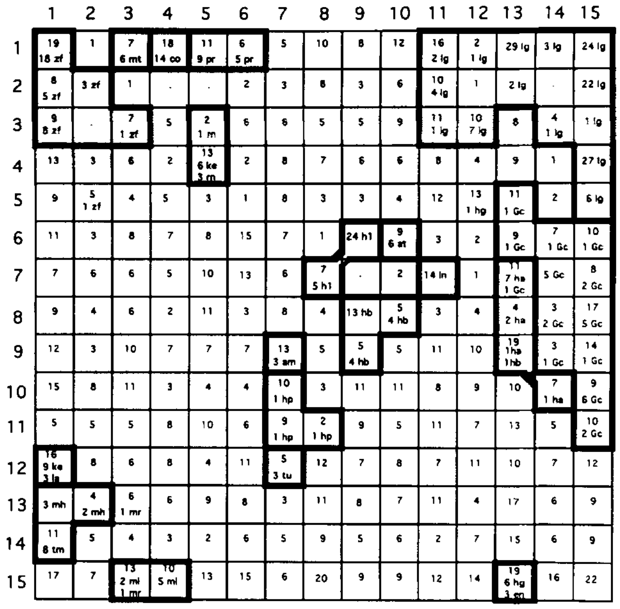
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	28 Ig	20 Ig	7 1 Ig	13 5 Ig	16 1 Ig	14	16 4 Ig	14	11	12	21	11	8 6 mt	3 zf	23 21 zf
2	.	.	5 1 Ig	2	1	2	6	10	6	8	18	2	4	3 zf	12 8 zf
3	30 Ig	2 Ig	13 12 Ig	6	10	9	9	7	13	6	8 3 am	11	3	2 1 zf	1
4	1 Ig	.	3	7	4	3	5	3	9 6 at	9 3 tu	12 3 hp	5	6	8	16
5	23 Ig	1 Ig	.	2	5	5	6	6	7	7	2	11	7	5	5
6	1	.	11 2 ke 4m	10 8 ke	3	9	13	9	11	12 2 en 3 hg	4	1	2	1	8
7	14 co	2	7 1 Gc	1	9	6	5 1 hg	4	6 1 hg	7 1 en	6	10	7	3	7
8	14 11 pr	3	3	5	1	2	13 5 Gc	6	11	12	2	1	9	2	10 7 ml 2 mr
9	12 3 pr	2 1 Ig	10	1	22 hb	.	4 2 Gc	17 10 Gc	1	6	10	14	8	5	14
10	12	5	11	2	.	1	14 11 ha	6	9	12	10	10	7	6	6
11	9	6	12	10	16	10	15	3	15	9	9	6	5	2	12
12	10 3 Gc	4	6	5	7	6	6	8	5	6	9	9	14	.	5 tm
13	20 2 Gc	6	11	7	16	12	12	11	12 1 Gc	11	12	9	6	7 1 mh	10 4 mh 3 tm
14	.	.	8	5	7	8	.	1	5 1 Gc	2	2	5	6	6 1 ke 2 la	2
15	29 h1	1	14 2 hg	12 1 Gc	11	10	13	16 2 Gc	10 1 Gc	3	14 In	1	15 4 ke 1 la	3	16

Fig. 5. Topological map for the 20 × 20 representation of 1,758 human proteins. Learning proceeded during 30 epochs, linearly decreasing α at each epoch ($\Delta t_\alpha = 1$), from an initial value of 0.9 ($a_1 = 0.08$ in the first 10 epochs and $a_2 = 0.00047619$ in the last 20) and decreasing the winner neighborhood every $\Delta t_v = 2$ epochs.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	19 Ig	14 Ig	10 Ig	21 Ig	1 Ig	13 7 ha	9 4 Gc	17 4 Gc	13 1 Gc	4	8	10 1 Gc	15 5 Gc	5	11
2	.	.	1 Ig	3 2 Ig	3	6	9 8 Gc	2 1 Gc	7 3 Gc	4	4	8	6	6	8
3	20 Ig	12 Ig	12 Ig	4	12	2	.	5	7	5	8 1 ha	7	10	1	5
4	4 Ig	1	2	6	10 1 Gc	5 1 Gc	14 In	1	11	8	13 2 ha	6	5	1	29 h1
5	13 2 Ig	8 1 Ig	12 11 Ig	5	11	3	2	4	10	4	3	5 1 ha	9 6 at	3	2 1 hg
6	9	5	13	6	7	10	6	10	9	14	13	9	17	10	8 6 hg
7	7	8	9	10	7	6	8	11	3	4	3 1 hb	8	6	8	16
8	9	7	9	6	7	9	5	12	2	17 hb	5 4 hb	10	6	4	13
9	13	9 1 rn	8	10	11	3	8 3 am 3 tu	9	6	4	6	10	11	7	14
10	5 pr	3	7	3	9	2	11	5	10	6	10	6 3 hp	15 3 en	5	4
11	9 pr	.	6 mt	2	12	5	9	8	7	5	6	5	11	7	10 8 ke
12	3	.	3	1	9	5	6	11	8	13	9	15	9	4	17 7 ke 3 la
13	16 14 co	3	11 3 rn	3	9	7	4	15	9	12	4	8 7 ml	8 1 mr	5 2 mh	7 3 mh
14	4 3 zf	4 zf	1	5	9	6	3	6	9	8	10	3	8	6	8 tm
15	16 15 zf	12 10 zf	8 3 zf	10	7	13 1 zf	9	16	11	8	12	7	13	7 1 mr	15

Fig. 6. Topological map for the 11 × 11 representation of 1,758 human proteins. Learning proceeded during 500 epochs (same learning parameters as in Fig. 4).



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	19 18 zf	1	7 6 mt	18 14 co	11 9 pr	6 5 pr	5	10	8	12	16 2 lg	2 1 lg	29 lg	3 lg	24 lg
2	8 5 zf	3 zf	1	.	.	2	3	8	3	6	10 4 lg	1	2 lg	.	22 lg
3	9 8 zf	.	7 1 zf	5	2 1 m	6	6	5	5	9	11 1 lg	10 7 lg	8	4 1 lg	1 lg
4	13	3	6	2	13 6 ke 3 m	2	8	7	6	6	8	4	9	1	27 lg
5	9	5 1 zf	4	5	3	1	8	3	3	4	12	13 1 hg	11 1 Gc	2	6 lg
6	11	3	8	7	8	15	7	1	24 h1	9 6 at	3	2	9 1 Gc	7 1 Gc	10 1 Gc
7	7	6	6	5	10	13	6	7 5 h1	.	2	14 ln	1	11 7 ha 1 Gc	5 Gc	8 2 Gc
8	9	4	6	2	11	3	8	4	13 hb	5 4 hb	3	4	4 2 ha	3 2 Gc	17 5 Gc
9	12	3	10	7	7	7	13 3 am	5	5 4 hb	5	11	10	19 1ha 1hb	3 1 Gc	14 1 Gc
10	15	8	11	3	4	4	10 1 hp	3	11	11	8	9	10	7 1 ha	9 6 Gc
11	5	5	5	8	10	6	9 1 hp	2 1 hp	9	5	11	7	13	5	10 2 Gc
12	16 9 ke 3 ln	8	6	8	4	11	5 3 tu	12	7	8	7	11	10	7	12
13	3 mh	4 2 mh	6 1 mr	6	9	8	3	11	8	7	11	4	17	6	9
14	11 8 tm	5	4	3	2	6	5	9	5	6	2	7	15	6	9
15	17	7	13 2 ml 1 mr	10 5 ml	13	15	6	20	9	9	12	14	19 6 hg 3 en	16	22

Fig. 7. Topological map for the 11 × 11 representation of 1,758 human proteins. Learning proceeded during 30 epochs (same learning parameters as in Fig. 5).

λ-chains (of 34). However, there were still cases, as in the second biggest cluster, in which too many proteins belonging to different families were clustered together (only 6 of the 38 proteins of that cluster belonged to the same family, α-chains of the HLA histocompatibility antigen).

To study how the sequences belonging to a same known family were distributed on different but neighboring clusters, we have built a tree of the 225 cluster centroids, using a standard hierarchical clustering algorithm (CLUSTER and TREE procedures of the SAS package). Table 9 shows how the sequences of each protein family were distributed among neighboring or distant clusters in the hierarchical tree. In a few cases, all the sequences of a given family were assigned to the same cluster (enolases, β-chains of HLA class II histocompatibility antigens, interferon-α precursors, myosin heavy chains, and tropomyosins). In some cases, the protein family was split in subfamilies but assigned to a group of neighboring clusters in the tree (actins, α-amylases, haptoglobins, hemoglobins, HLA class I histocompatibility antigens, immunoglobulin κ-chains, metallothioneins, and myosin light chains). In other cases, this subdivision was done involving several distant groups of neighboring clusters (collagens and proline-rich proteins). In the most common case, the bulk of the protein family was classified into 1 or more neighboring clusters and the remaining sequences were placed in several distant clusters (α-chains of HLA class II histocompatibility antigens; immunoglobulin λ-, heavy and C-region chains; keratins and zinc-finger proteins). In three of the analyzed cases (lamins, G-protein coupled receptors, and tubulin β-chains), the proteins were scattered on several distant clusters.

In Table 9 we have also indicated (last column) whether the sequences were considered or not as outliers. Note that in many cases, though all the sequences belonging to a same protein family were considered as outliers, they were still grouped together into neighboring clusters of the tree (actins, HLA class I histocompatibility antigens, metallothioneins, etc.).

Table 8. *Statistical classification of 1,758 human protein sequences: distribution of cluster population*

	Number of clusters	
Population	$\theta = \infty$[a]	$\theta = 1.5$[b]
1	116	13
2	34	16
3	14	19
4	7	26
5	11	22
6	5	23
7	3	21
8	1	23
9	2	9
10	4	4
11	3	5
12		6
13	1	3
14	2	5
15		5
16		7
17	1	2
18		4
19		2
20		1
21	1	1
22	2	
23	1	2
24		1
25	1	
26	2	
27	2	
28		1
29	1	
32		1
33	1	
37		1
38		1
39		1
46	1	
47	1	
53	1	
62	2	
74	1	
78	1	
88	1	
374	1	

[a] $\theta = \infty$ (no outlier elimination).
[b] $\theta = 1.5$ (1,126 outliers).

Discussion

We have described here a large-scale application of the ANN method to classify all the 1,758 known human protein sequences. We have also investigated the influence on the protein clustering of 4 different dipeptide matrix representations of the sequences as well as that of slow and fast learning protocols. For the 20 × 20 or the 11 × 11 matrices, similarity relationships between protein sequences were found to be mapped into neighborhood relationships of neural activity on the 2-dimensional layer of neurons, regardless of the speed of the learning proce-

Table 9. *Statistical classification of known protein families*[a]

Protein family	Number of proteins	Number of clusters	Outliers
Actins	6	2	Yes
α-Amylases	3	2	Yes
Collagens	6	4	Yes
	4	4	Yes
	4	1	Yes
Enolases	3	1	No
G-coupled receptors	29	(15)	Yes
Haptoglobins	3	2	Yes
Hemoglobins	7	3	No
HLA histocompatibility antigens			
Class I	12	5	Yes
α-Chain, class II	10	3	Yes
	1	1	Yes
β-Chain, class II	22	1	No
Immunoglobulins			
λ-Chain, V-region	30	1	No
	4	(4)	No
κ-Chain, V-region	48	3	No/yes
Heavy chain, V-region	29	4	No/yes
	7	(4)	No
C-region	7	3	No/yes
	5	(3)	No/yes
Interferon-α precursors	14	1	No
Keratins	9	4	Yes
	2	2	Yes
	2	2	Yes
	2	(2)	Yes
Lamins	3	(2)	Yes
Metallothioneins	6	2	Yes
Myosins			
Light chains	9	2	No/yes
Heavy chains/tropomyosins	13	1	No
Proline-rich proteins	5	1	Yes
	3	1	No/yes
	6	1	No/yes
Tubulin, β-chains	3	(2)	Yes
Zinc-finger proteins	35	1	No
	1	1	No

[a] Each line corresponds to a set of neighboring clusters of related proteins, except for lines where the number of clusters is indicated between parentheses. The last column indicates whether the dipeptide matrices of a given protein family have been considered as outliers in the final classification (No/yes: mixture of outliers and nonsingular individuals).

dure. The set of neighborhood relationships given by 1 particular map should be considered as 1 of many possible suitable ways to classify the set of proteins into a multiple number of clusters. Because the network actually classifies dipeptide representations of the protein sequences, some of these neighborhood relationships may indicate only a resemblance between the corresponding dipeptide matrices. This shortcoming, due to the simplified way in which we have encoded sequence information, is compensated by the fact that such encoding allows for comparison of sequences without having to align them.

We have found very similar neural maps for the 20 × 20 representation of the 1,758 human proteins when both slow and fast learning procedures were used. In addition, the map for the 11 × 11 representation, though somehow different from the map obtained with the 20 × 20 representation, also seems to lead to a suitable classification for that learning set. The 11 × 11 matrix representation, based on considering amino acids of similar physicochemical properties as a same kind of residue, reduces the required computing time for the learning procedure by lowering the number of inputs to the network, that is, by reducing the number of floating point operations necessary to compute distances between vectors. This approach is an example of a typical strategy to design neural networks that consists of including all of the available knowledge of the task to be learned, either in the network architecture or in a pre-processing of input signals. Using both an 11 × 11 protein representation and a fast learning procedure, the computing time required for the training stage is reduced from 100 to 1.8 CPU-h, on a SUN 4/360. Because the algorithm scales up nearly linearly with the number of protein sequences of the learning set, this reduction in computing time renders the classification of the whole database feasible. Work is in progress to increase further the speed of the learning stage by using parallel-processing machines based on neural network algorithms (Gamrat et al., 1991).

Our 20 × 20 protein representation is a particular case of n-gram, previously called bi-gram or a2 by Wu et al. (1992). The n-gram encoding of a given protein sequence is an n-dimensional matrix that gives the number of occurrences of all possible patterns of n consecutive residues (Wu et al., 1992). Interestingly, Wu et al. reported that the highest predictive accuracy and fastest convergence rate are obtained when this particular encoding is concatenated with the amino acid compositions (i.e., the a1 n-gram) and some of the 2 or 3 lowest exchange group n-grams (the e1 and e2 or the e1, e2, and e3 n-grams). This suggests that our results may be further improved by considering this kind of concatenated sequence representation.

The learning procedure, which is the most time-consuming step of the ANN approach, needs to be performed only once. Furthermore, once the network has self-organized itself, it can be used to classify unknown sequences rapidly. As an example of the retrieval stage, we fed the network trained with the 20 × 20 representation of the human proteins (map of Fig. 4) with a sequence that was not included in the learning set, the mouse vimentin. First, the dipeptide matrix corresponding to this protein sequence was obtained. Then, the input pattern was compared with the whole set of synaptic vectors to determine the winner neuron. As a result of the retrieval stage, we obtained the position of the winner neuron [(1,13) in our example]; the Euclidean distance d between the input pattern and the synaptic vector of the winner neuron ($d_{new} = 0.5226$ in our case); and the list of learned proteins having that neuron as winner, with their corresponding distances d (that is, the list shown in Table 7).

The comparison of the value d_{new} with those of Table 7 shows that d_{new} is between the distances for human vimentin and keratin (type I cytoskeletal 18). This suggests that these are the closest human sequences to the sequence that we have fed as input. The whole retrieval stage for our example took 14.6 CPU-seconds (CPU-s) on a VAX 6310 computer. This computing time is smaller than those corresponding to the BLAST (29 CPU-s) and FASTA algorithms (57 CPU-s), for an equivalent searching procedure. In addition, a vimentin fragment (the partial sequence stored for pig vimentin) was also classified into neuron (1,13), but between a keratin (k2ca$human) and an apolipoprotein (ape$human) ($d_{pig} = 0.7317$). This fragment has a length of only 275 amino acids (the human-vimentin sequence length is 465 amino acids). It should be noted that the retrieval stage may also be performed quickly in less powerful personal computers, with a simple software that makes use of the final computed synaptic vector values of the topological map.

Although protein sequence databases may be updated daily, we presently perform the learning stage quarterly (i.e., for each new release of the SwissProt database). However, it should be noted that new sequences can always be immediately compared with the proteins already classified in the topological map to find the family to which they belong. We are also exploring several strategies to cope with a more frequent update of the database, based mainly on additional short learning stages with low values of adaptation gain and small sizes of winner neighborhood.

In the computational experiences described above, the number of neurons of the network has been chosen by experience. This is a typical constraint of neural network approaches. In our case, this is somehow analogous to the choice of the number of cluster seeds in standard nonhierarchical clustering statistical methods. The number of neurons in the network introduces an upper limit on the number of clusters in which patterns may be classified. In the ANN approach, the network makes use of all the patterns of the learning set to determine the synaptic vectors (which play the role of cluster seeds). However, in Kohonen's algorithm, this determination is mainly guided by those input patterns that appear most frequently (in our case, this corresponds to those families having the highest number of homologous sequences). After training, those input patterns that do not belong to the biggest families are finally assigned to the closest synaptic vector. This is somewhat reminiscent of the strategy that we have used to cope with the great number of outliers found in the statistical classification of the human protein sequences. In fact, Kohonen's algorithm handles the outliers in a natural way, without any need of further action to detect and classify them. The existence of clusters including "outliers" can be inferred from "jumps" that appear in the list of d values, when the set of proteins associated with each neuron is sorted according to those values. For example, in the list of Table 7, the jump in the values of d from 0.6444 to 0.7420 is bigger than for any other pair of successive values of that list. This jump may be used to classify further the proteins of the list into 2 subfamilies. In general, the existence of such jumps may also be regarded as an indication that the number of neurons is still not enough to achieve a complete classification of the whole set of sequences. However, a bigger network might lead to an excessive subclassification of the outliers, taking each of them as a particular cluster.

The evaluation of a final topological map is a challenging task. In the present work, that evaluation has been performed using biological knowledge (that is, looking whether proteins with similar biological functions were classified together), in the case of human proteins. Van Heel (1991) has used a statistical method to classify dipeptide matrices and Gonnet et al. (1992) have organized an entire protein sequence database by indexing on a patricia tree. Although the comparison of classifications obtained with alternative methods may help to judge their quality, this comparison of different algorithms becomes quite difficult when a great number of protein sequences and families is considered. For that reason, we have used a small set of protein sequences (the chemokine family) to show that the classification obtained with the ANN approach is very similar to the one obtained with other conventional methods of biosequence analysis. Recently, we have also compared the ANN approach with a statistical nonhierarchical clustering method (similar to the one described above; see Pflugfelder & Ferrán [1992] for further details) and we found that both classifications were quite similar (Ferrán & Pflugfelder, 1993). However, we have shown here that the ANN method seems to provide a better classification than standard nonhierarchical clustering methods, when more complex sets of protein sequences are considered.

In addition, we have recently shown that the results issued from the statistical methods can be used not only to validate the classifications obtained with the ANN approach, but also to choose, in a more reasonable way, the number of neurons of the network (Ferrán & Pflugfelder, 1993). This can be done by relating a statistical determination of the optimal number of clusters with the number of neurons of the network. Note, however, that the great number of outliers found in the statistical classification of the set of human proteins may hinder this determination, thus the optimal number of clusters should be obtained by taking into account only the set of nonsingular patterns.

In conclusion, we have shown that the proposed unsupervised method is a helpful computational tool for clustering proteins into families without having previous knowledge of the number and composition of the final clusters. The clustering of a biosequence database may reduce searching time in large protein databases. However, the simplification in the protein representation also implies a degradation in sensitivity. Therefore, this alternative approach should be seen as complementary and cooperative to the developments in pattern-matching algorithms, RISC technology, and massively parallel computing under way to cope with the volumes of data expected from the genome sequencing projects.

Methods

In this section we summarize the standard formalism of the method that we have previously proposed (see Ferrán & Ferrara [1991, 1992a] for a detailed description).

In general, we consider a 2-dimensional network, that is, 1 layer of $N_x \times N_y$ neurons. Each neuron receives, as input signals, a pattern of $n \times n$ components ξ_{kl}, obtained from the dipeptide composition of the protein to be learned (where n is the number of different symbols in the residues' alphabet). The n^2 values of the corresponding synaptic efficacies that weight the input signals are the components of a synaptic vector associated with each neuron. We denote by $\boldsymbol{m}_{ij}$ the synaptic vector of the neuron placed in the (i,j) site of the output layer. We will usually identify each neuron directly by its position. At the beginning, all synaptic vector components $\mu_{ij,kl}$ are real numbers randomly taken from the interval [0,1]. Both input patterns and synaptic vectors are normalized to unitary vectors. Each protein pattern is presented as input to the network and the neuron having the closest synaptic vector to the protein pattern (the winner neuron) is selected. Then, the synaptic vectors of all neurons belonging to a winner neighborhood N_w are changed in order to bring them closer to the vector of input signals:

$$\mu_{ij,kl}(t+1) = \mu_{ij,kl}(t) + \alpha(t)[\xi_{kl}(t) - \mu_{ij,kl}(t)],$$

$$\forall \text{ neuron } (i,j) \in N_w,$$

where $0 < \alpha(t) < 1$. All protein patterns of the learning set are repeatedly processed by the network, in the same sequential order. Each processing cycle of the whole learning set is called an epoch. As learning proceeds, α is linearly or exponentially decreased every Δt_α epochs ($0 < a < 1$):

$$\alpha(t + \Delta t_\alpha) = \alpha(t) - a,$$

$$\alpha(t + \Delta t_\alpha) = a\alpha(t),$$

and the winner neighborhood is shrunk, from the whole network to the winner neuron, every Δt_v epochs. Usually, the number of training epochs is initially fixed, but taking care to end the learning stage with $\alpha \approx 0$ and N_w equal to the winner neuron.

Once learning has been accomplished, each sequence of the learning set is finally associated with the neuron having the closest synaptic vector. Thus, each synaptic vector of the trained network may be considered as a "consensus pattern" for the set of dipeptide matrices of protein sequences associated with the corresponding neuron.

Acknowledgments

We thank N. Vita, A. Minty, D. Verrier, and G. Hammonds for useful comments and suggestions on our work and L. Hernández and M. Magazin for their valuable corrections of the manuscript.

References

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

Altschul SF, Lipman DJ. 1990. Protein database searches for multiple alignments. *Proc Natl Acad Sci USA 87*:5509–5513.

Anderberg MR. 1973. *Cluster analysis for applications*. New York: Academic Press.

Andreassen H, Bohr H, Bohr J, Brunak S, Bugge T, Cotterill RMJ, Jacobsen C, Kusk P, Lautrup B, Petersen SB, Særmark T, Ulrich K. 1990. Analysis of the secondary structure of the human immunodeficiency virus (HIV) proteins p17, gp120, and gp41 by computer modeling based on neural network methods. *J Acquired Immun Defic Syndrome 3*:615–622.

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*:353–357.

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 S, Cotterill RMJ, Fredholm H, Lautrup B, Petersen SB. 1990. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Lett 261*:43–46.

Bohr H, Bohr J, Brunak S, Cotterill RMJ, Lautrup B, Nørskov L, Olsen OH, Petersen SB. 1988. Protein secondary structure and homology by neural networks. The α-helices in rhodopsin. *FEBS Lett 241*:223–228.

Brunak S, Engelbrecht J, Knudsen S. 1990. Neural network detects errors in the assignment of mRNA splice sites. *Nucleic Acids Res 18*:4797–4801.

Brunak S, Engelbrecht J, Knudsen S. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol 220*:49–65.

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

Demeler B, Zhou G. 1991. Neural network optimization for *E. coli* promoter prediction. *Nucleic Acids Res 19*:1593–1599.

Devereux J, Haeberli P, Smithies O. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res 12*:387–395.

Engelbrecht J, Knudsen S, Brunak S. 1992. G+C-rich tract in 5′ end of human introns. *J Mol Biol 227*:108–113.

Farber R, Lapedes A, Sirotkin K. 1992. Determination of eukaryotic protein coding regions using neural networks and information theory. *J Mol Biol 226*:471–479.

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

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

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

Ferrán EA, Ferrara P. 1992b. A neural network dynamics that resembles protein evolution. *Physica A 185*:395–401.

Ferrán EA, Pflugfelder B. 1993. A hybrid method to cluster protein sequences based on statistics and artificial neural networks. *Comput Appl Biosci 8*:671–680.

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

Gamrat C, Mougin A, Peretto P, Ulrich O. 1991. The architecture of MIND neurocomputers. *Proc. Second Conference on Microelectronics for Neural Networks*. München. pp 463–469.

Gonnet GH, Cohen MA, Benner SA. 1992. Exhaustive matching of the entire protein sequence database. *Science 256*:1443–1445.

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

Hartigan JA. 1975. *Clustering algorithms*. New York: John Wiley & Sons.

Higgins DG, Sharp PM. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. *Comput Appl Biosci 5*:151–153.

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.

Hirst JD, Sternberg MJE. 1992. Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry 31*:7211–7218.

Holbrook SR, Muskal SM, Kim SH. 1990. Predicting surface exposure of amino acids from protein sequence. *Protein Eng 3*:659–665.

Holley LH, Karplus M. 1989. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA 86*:152–156.

Horton PB, Kanehisa M. 1992. An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. *Nucleic Acids Res 20*:4331–4338.

Kneller DG, Cohen FE, Langridge R. 1990. Improvements in protein secondary structure prediction by an enhanced neural network. *J Mol Biol 214*:171–182.

Kohonen T. 1982. Self-organized formation of topologically correct feature maps. *Biol Cybern 43*:59–69.

Ladunga I, Czakó F, Csabai I, Geszti T. 1991. Improving signal peptide prediction accuracy by simulated neural network. *Comput Appl Biosci 7*:485–487.

Lapedes A, Barnes C, Burks C, Farber R, Sirotkin K. 1990. Application of neural networks and other machine learning algorithms to DNA sequence analysis. In: Bell G, Marr T, eds. *Computers and DNA. SFI studies in the sciences of complexity, vol VII*. Addison-Wesley. pp 157–182.

Le Cun Y. 1985. A learning scheme for asymmetric threshold networks. *Proc Cognitiva (Paris)*:599–604.

Lipman DJ, Pearson WR. 1985. Rapid and sensitive protein similarity searches. *Science 227*:1435–1441.

Lukashin AV, Anshelevich VV, Amirikyan BR, Gragerov AI, Frank-Kamenetskii MD. 1989. Neural network models for promoter recognition. *J Biomol Struct Dyn 6*:1123–1133.

MacQueen JB. 1967. Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability 1*:281–297.

Maddox J. 1992. Ever-longer sequences in prospect. *Nature 357*:13.

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

Minty A, Chalon P, Guillemot JC, Kaghad M, Liauzun P, Magazin M, Miloux B, Minty C, Ramond P, Vita N, Lupker J, Shire D, Ferrara P, Caput D. 1993. Molecular cloning of the MCP-3 chemokine gene and regulation of its expression. *Eur Cytokine Netw 4*:99–110.

Muskal SM, Holbrook SR, Kim SH. 1990. Predicting of the disulfide-bonding state of cystein in proteins. *Protein Eng 3*:667–672.

Muskal SM, Kim SH. 1992. Predicting protein secondary structure content. A tandem neural network approach. *J Mol Biol 225*:713–727.

Nakayama SI, Shigezumi S, Yoshida M. 1988. Method for clustering proteins by use of all possible pairs of amino acids as structural descriptors. *J Chem Inf Comput Sci 28*:72–78.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol 48*:443–453.

O'Neill MC. 1991. Training back-propagation neural networks to define and detect DNA-binding sites. *Nucleic Acids Res 19*:313–318.

O'Neill MC. 1992. *Escherichia coli* promoters: Neural networks develop distinct descriptions in learning to search for promoters of different spacing classes. *Nucleic Acids Res 20*:3471–3477.

Oppenheim JJ, Zachariae COC, Mukaida N, Matsushima K. 1991. Properties of the novel proinflammatory supergene "intercrine" cytokine gene family. *Annu Rev Immunol 9*:617.

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

Petersen SB, Bohr H, Bohr J, Brunak S, Cotterill RMJ, Fredholm H, Lautrup B. 1990. Training neural networks to analyse biological sequences. *Trends Biotechnol 8*:304–308.

Pflugfelder B, Ferrán EA. 1992. Bipeptidic matrix clustering. *Proceedings of SEUGI*. SAS Institute. pp 656–686.

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

Rose VS, Croall IF, MacFie HJH. 1991. An application of unsupervised neural network methodology (Kohonen topology-preserving mapping) to QSAR analysis. *Quant Struct Act Relat 10*:6–15.

Rosenblatt F. 1962. *Principles of neurodynamics*. New York: Spartan Books.

Rost B, Sander C. 1993a. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA 90*:7558–7562.

Rost B, Sander C. 1993b. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning representations by back-propagating errors. *Nature 323*:533–536.

SAS Institute, Inc. 1985. *SAS user's guide: Statistics, version 5*. Cary, North Carolina: SAS Institute, Inc.

Snyder EE, Stormo GD. 1993. Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks. *Nucleic Acids Res 21*:607–613.

Stolorz P, Lapedes A, Xia Y. 1992. Predicting protein secondary structure using neural net and statistical methods. *J Mol Biol 225*:363–377.

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. 1982. Use of the "Perceptron" algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res 10*:2997–3011.

Sulston J, Du Z, Thomas K, Wilson R, Hillier L, Staden R, Halloran N, Green P, Thierry-Mieg J, Qiu L, Dear S, Coulson A, Craxton M, Durbin R, Berks M, Meltztein M, Hawkins T, Ainscough R, Waterston R. 1992. The *C. elegans* genome sequencing project: A beginning. *Nature 356*:37–41.

Uberbacher EC, Mural RJ. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc Natl Acad Sci USA 88*:11261–11265.

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

Vieth M, Kolinski A. 1991. Prediction of protein secondary structure by an enhanced neural network. *Acta Biochim Pol 38*:335–351.

von Heijne G. 1991. Computer analysis of DNA and protein sequences. *Eur J Biochem 199*:253–256.

Wade RC, Bohr H, Wolynes PG. 1992. Prediction of water binding sites on proteins by neural networks. *J Am Chem Soc 114*:8284–8285.

Watson JD. 1990. The human genome project: Past, present and future. *Science 248*:44–49.

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

Zhang X, Mesirov JP, Waltz DL. 1992. Hybrid system for protein secondary structure prediction. *J Mol Biol 225*:1049–1063.