Protein Engineering vol.9 no.10 pp.833–842, 1996

Local structural motifs of protein backbones are classified by self-organizing neural networks

Johannes Schuchhardt, Gisbert Schneider, Joachim Reichelt[1], Dietmar Schomburg[1] and Paul Wrede[2]

Freie Universität Berlin, Universitätsklinikum Benjamin Franklin, Institut für Medizinische/Technische Physik und Lasermedizin, AG Molekulare Bioinformatik, Krahmerstraße 6–10, D-12207 Berlin and [1]Gesellschaft für Biotechnologische Forschung mbH, Molekulare Strukturforschung, Mascheroder Weg 1, D-38124 Braunschweig, Germany

[2]To whom correspondence should be addressed

Important and relevant information is expected to be encoded in local structural elements of proteins. An unsupervised learning algorithm (Kohonen algorithm) was applied to the representation and unbiased classification of local backbone structures contained in a set of proteins. Training yielded a two-dimensional Kohonen feature map with 100 different structural motifs including certain helical and strand structures. All motifs were represented in a ϕ–ψ-plot and some of them as a three-dimensional model. The course of structural motifs along the backbone of four selected proteins (cytochrome b_5, cytochrome b_{562}, lysozyme, γ crystallin) was investigated in detail. Trajectories and histograms visualizing the abundance of characteristic motifs allowed for the distinction between different types of protein overall folds. It is demonstrated how the histograms may be used to construct a structural similarity matrix for proteins. The Kohonen algorithm provides a simple procedure for classification of local protein structures independent of any *a priori* knowledge of leading structural motifs. Training of the Kohonen network leads to the generation of 'consensus structures' serving for the task of classification.

Keywords: feature map/Kohonen network/protein similarity/protein structure/structural universe

Introduction

Proteins can be regarded as being built up from modular structural motifs such as helices, β-sheets, turn- or loop-structures. A number of characteristic structural elements found by automatic template-based classification procedures have been described (Levitt and Chothia, 1976; Levitt and Greer, 1977; Richardson, 1981; Kabsch and Sander, 1983; Leszczynski and Rose, 1986; Richards and Kundrot, 1988; Prestrelski *et al.*, 1992; Efimov, 1994; Zhu, 1995). In general, however, structural similarities or dissimilarities between these motifs are not evident *per se*, and a stringent order is not necessarily given. We have applied self-organizing Kohonen networks (Kohonen, 1977, 1990; Ritter and Kohonen, 1989) to feature extraction and classification of local protein backbone elements. The aim was to construct a modular system of small building blocks ordered according to the individual markedness of predominant features. Identification of the different structural motifs may lead to a deeper understanding of the modular architecture of proteins.

The Kohonen algorithm generates a feature map, i.e. an array of formal neurons, each representing a feature derived from the data used for training (Schuchhardt *et al.*, 1992). Kohonen feature maps provide an elegant approach to projecting high-dimensional data on to a low-dimensional map. Most common and easy to visualize are two-dimensional maps. In contrast to many statistical methods, e.g. principal component analysis, the projection is non-linear and allows mapping of a rather complex space on to a plane (Ritter and Kohonen, 1989; Zupan and Gasteiger, 1993; White, 1994). Inherent to the algorithm is the preservation of the topology of the input space, i.e. relations between the high-dimensional data are preserved in their low-dimensional projection. In this work each neuron of the feature map can be considered as representing a consensus structural backbone motif. In contrast to supervised learning strategies, the Kohonen algorithm is able to extract characteristic features without any *a priori* knowledge of leading structures, i.e. it is not template-based.

Materials and methods

Sequence data

From the Brookhaven Database, PDB (Bernstein *et al.*, 1977), 136 non-redundant protein sequences were selected to form a representative set (Hobohm *et al.*, 1992). The pairwise identity within the set is less than 30% and each sequence represents a structural family. The PDB identifiers of the proteins used in this analysis are 1acx, 1ak3, 1azu, 1bbp, 1bds, 1bmv, 1cbh, 1cc5, 1cd4, 1cdt, 1crn, 1cse, 1ctf, 1dhf, 1eca, 1etu, 1fc2, 1fdl, 1fdx, 1fkf, 1fx1, 1fxi, 1gd1, 1gox, 1gp1, 1hds, 1hip, 1hne, 1il8, 1l58, 1lap, 1lrd, 1mcp, 1mrt, 1ovo, 1paz, 1pmb, 1ppt, 1prc, 1pyp, 1r09, 1rbp, 1rhd, 1rnh, 1s01, 1sdh, 1sgt, 1sh1, 1tgs, 1tnf, 1ubq, 1wsy, 256b, 2aat, 2cab, 2ccy, 2cyp, 2fnr, 2fxb, 2gbp, 2gcr, 2gls, 2gn5, 2hla, 2hmz, 2i1b, 2lbp, 2lh4, 2lhb, 2ltn, 2mev, 2mhu, 2or1, 2pab, 2pcy, 2phh, 2pka, 2rsp, 2sns, 2sod, 2stv, 2taa, 2tbv, 2tgp, 2tmv, 2tsc, 2utg, 2wrp, 3ait, 3b5c, 3blm, 3cla, 3cln, 3dfr, 3ebx, 3gap, 3hla, 3hmg, 3icb, 3pgm, 3rnt, 3tim, 4bp2, 4cms, 4cpa, 4cpv, 4fxn, 4gr1, 4mdh, 4pfk, 4rhv, 4rxn, 4sbv, 4sgb, 4ts1, 4xia, 5cyt, 5er2, 5hvp, 5ldh, 5rub, 6cpa, 6cpp, 6cts, 6dfr, 6hir, 6tmn, 7cat, 7icd, 7rsa, 8adh, 8atc, 9api, 9ins, 9pap, 9wga.

Data representation and pattern preparation

A reduced representation model of protein main-chains by their dihedral Ψ and Φ angles was used to describe the set of protein structures. To keep the system as clear as possible, information on the omega angle was not considered and each protein was represented by a sequence of pairs of real numbers between −180 and +180° starting with the N-terminal Φ angle and ending with the C-terminal Ψ angle. The amino acid sequences were scanned by overlapping windows encompassing nine residues each. According to the definition of main-chain dihedral angles a sequence segment of nine amino acid residues corresponds to eight pairs of Ψ and Φ angular values. Training patterns were generated by the sliding-window

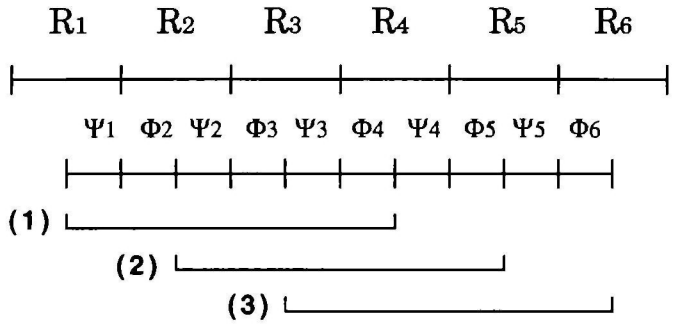

Fig. 1. Generation of training patterns from a small polypeptide. Below the amino acid sequence the Ψ and Φ angles are aligned. In this example the window covers only six angular values while in the experiments a window covering 16 values was applied. Scanning the peptide in steps of one residue generates three different training patterns.

technique employing a window size of 16 angular values (Figure 1) and moving in steps of one residue (i.e. two angular values). A training pattern reached from the Φ angle of the first residue ψ_1 to the Ψ angle of the last residue φ_9.

The width of nine amino acids for the the window was somewhat arbitrarily chosen. We performed a number of experiments with different window sizes all leading to meaningful feature maps. Generally a small window size will be favoured taking into account database completeness (Rooman and Wodak, 1988; Fidelis *et al.*, 1994). A nine-residue window was selected since it is large enough to cover characteristic structural elements at all, e.g. about three turns of a helix. On the other hand it is still small enough to contain uniform segments of periodical structures, e.g. only parts of a long helix or a β-strand. The training set generated from the 136 proteins contained 23 151 patterns.

Angular distance measure for local structural elements

For network training and for the investigation of the feature map a measure for the distance of two structural elements needed to be defined. Given the representation of two structural elements s and s′ in terms of their dihedral angles the distance ε was defined as the root mean square deviation (r.m.s.d.) on angular values:

$$\varepsilon(s,s') = \sqrt{\frac{1}{16}\sum_{k=1}^{8}\left[(\varphi_k - \varphi_k')^2 + (\psi_k - \psi_k')^2\right]}$$

The angular values may stem from either a training pattern or from a neural weight vector (see below).

Network architecture and training algorithm

The neural network used to generate the feature map consisted of 10×10 neurons arranged in a plane. For the network a flat topology was chosen rather than a toroidal or spherical one, since the space of structural elements cannot be expected to have a closed topology. Each neuron was characterized by its position $\mathbf{r} = (r_x,r_y)$ and its weight-vector $\mathbf{w}$ (Figure 2). The weight vectors were 16-dimensional ($N = 16$), just as the training patterns. Starting from a random initialization the network was trained to perform a nonlinear mapping. During training a pattern $\mathbf{v}$ was randomly selected from the pattern set and presented to the network. Then the neuron closest to the pattern $\mathbf{v}$ was determined by finding the weight vector $\mathbf{w}^{next}$ being closest to the training pattern $\mathbf{v}$ presented:

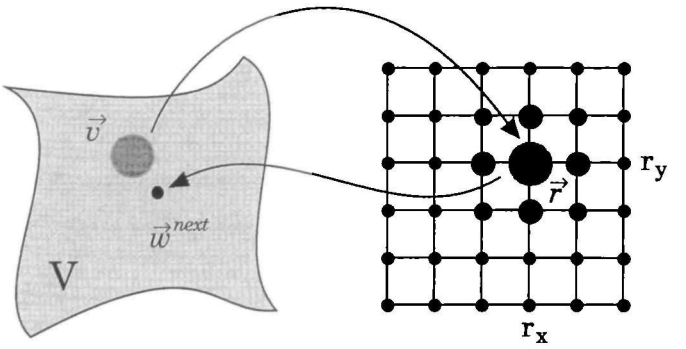

Fig. 2. Schematic representation of the nonlinear mapping from the complex input-space V the two-dimensional plane formed by the network. The input pattern $\mathbf{v}$ is represented by the neuron located in $\mathbf{r} = (r_x,r_y)$ with the corresponding weight vector $\mathbf{w}^{next}$.

$$\varepsilon(\mathbf{w}^{next},\mathbf{v}) = \text{minimum}$$

A distance measure ε was required to determine the next neuron. Here we chose the r.m.s.d. on angles as explained above. In terms of the weight vectors, the distance measure was

$$\varepsilon(\mathbf{w},\mathbf{v}) = \sqrt{\frac{1}{N}(\mathbf{w} - \mathbf{v})^2}$$

Once the neuron closest to the pattern $\mathbf{v}$ had been determined, the weight vectors were changed according to the prescription

$$\mathbf{w}^k(t+1) = \mathbf{w}^k(t) + \left[\mathbf{v} - \mathbf{w}^k(t)\right]\eta\, e^{-\frac{1}{2r^2}(\mathbf{r}^k - \mathbf{r}^{next})^2}$$

The learning rate η and the update radius r were continuously decreased during the training procedure:

$$\eta(t) = \frac{\eta_0}{1 + \frac{t}{\tau}}, \quad r(t) = \frac{r_0}{1 + \frac{t}{\tau}}$$

where τ is the number of training-patterns. On average each pattern was presented 10 times. A schematic description of the nonlinear mapping performed by the network is shown in Figure 2 (Schuchhardt *et al.*, 1992): any input pattern $\mathbf{v}$ presented to the network is mapped to the neuron with its weight vector $\mathbf{w}^{next}$ closest to the pattern.

Construction of a structural similarity measure

Based on the topological map, a structural measure of protein similarity was defined by assigning a number between 0 and 1 to each pair of proteins of known three-dimensional structure. First a histogram was constructed by scanning the protein and counting how often each neuron is active. This histogram was normalized yielding a positive vector $\mathbf{p}$ with absolute norm equal to one. The distance between two histograms A and B was defined:

$$\text{dist}(A,B) = \frac{1}{2}\sum_{i=1}^{100}|p_A^i - p_B^i|$$

The value of dist varies between 0 and 1 and the similarity measure was defined:

$$\text{sim}(A,B) = 1 - \text{dist}(A,B)$$

For proteins with a comparable proportion of similar structural elements the distance will be small and the similarity

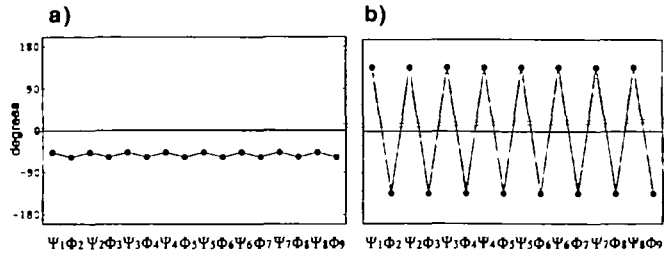

Fig. 3. Ψ–Φ plots of two ideal structural motifs, α-helix with Ψ = –57° and Φ = –47° (a) and β-strand with Ψ = –139° and Φ = +135° (b). The dihedral angles Ψ and Φ indicated by black dots are drawn along a sequence of nine amino acid residues

will be large. For proteins with a different structural spectrum the distance will be large and their similarity will be small.

Results

Protein backbones are represented by their dihedral angles

Protein structures (136) (Hobohm *et al.*, 1992) from the Brookhaven Database (Bernstein *et al.*, 1977) served for the generation of training patterns employing the 'sliding window' technique (see Materials and methods). Each training pattern corresponds to a structural element encompassing nine residues and it can be visualized by plotting the sequence of its main-chain Ψ and Φ dihedral angles versus the sequence position. This representation is termed a ϕ–ψ plot and is illustrated for two idealized structural motifs in Figure 3.

The two idealized motifs correspond to an α-helix with periodical dihedral angle values ϕ = –57° and ψ = –47°, and a β-strand with periodical values ϕ = –139° and ψ +135° (Ramachandran and Sassiekharan, 1968). The 23 151 structural elements generated from the protein data served for training of a Kohonen feature map (see Materials and methods).

A structural feature map results from network training

The trained network consisting of 10×10 neurons is given in Figure 4(a). Each of the neurons is characterized by its position in the network and its weight vector. The position is a pair of coordinate numbers (*x*/*y*) with the neuron (0/0) in the lower left corner and neuron (9/9) in the upper right corner of the feature map. A neuron's weight vector is an array of 16 real values corresponding to the Ψ and Φ angles of the structural elements used for training. The weight vector itself can be interpreted as a structural motif and visualized in a ϕ–ψ-plot (Figure 3). During network training the weights were changed adaptively to generate a set of 100 motifs which are representative for all structural elements in the training data. The motifs were automatically ordered in a topological map assembling similar structural elements in close vicinity. The resulting feature map generated by the Kohonen algorithm is given in Figure 4(a).

Each of the 100 squares in the map is a ϕ–ψ plot of an adapted weight vector and can be interpreted as a consensus 3-D-backbone structure. A helical motif similar to that given in Figure 3 was found by neuron (0/8) localized in the upper left area of the map. Some distorted helical motifs are assembled in close vicinity, e.g. neuron (0/7) and neuron (1/8). β-Strand-like structures are located in the opposite corner of the feature map, where the most regular strand is motif (6/2) surrounded by a number of more distorted strand structures.

There are homogeneous and complex structural features

An overview of the Kohonen feature map in terms of structural similarities shows a clear separation of dominant structural elements as α-helix in the upper left corner and β-strand in the lower right corner (Figure 4b). Calculation of the similarity value ε (see Materials and methods) is based on the r.m.s.d. on dihedral angles between the structural element under investigation and the ideal helix or strand respectively (Figure 3a and b).

The motifs most similar to the idealized α-helix or β-strand extracted by the Kohonen network are neuron (0/8) for the helix, and neuron (6/2) for the strand. This may be seen from Figure 4(b) giving the proximity to the ideal helix in yellow and to the β-strand in red. Dark regions correspond to loop motifs with only minor helical or strand contributions. Very dark-coloured squares located mainly at the right edge of the map indicate a low content of both helix and strand. Topology conservation implies that a small step in the map should result in a comparably small alteration of the local structure. This is confirmed by the observation that abrupt changes of shading rarely occur in the map: the path from a helix-prone structure to a sheet-like structure usually includes a number of intermediate motifs. The results of a numerical calculation of the distances to the next neighbours are listed in Table I.

For an interpretation of the distance values a simplifying two-state model (α–β-model) of the structural elements is useful: the dihedral angles may assume only two configurations, the helical (ϕ, ψ) = (–57, –47°) or the strand configuration (ϕ, ψ) = (–139, +135°). If a single position of a structural element with eight pairs of dihedral angles is changed (single-site alteration) the resulting r.m.s. angular distance (see Materials and methods) is approximately 50°. The average distance from a random ensemble of such two-state structures is approximately 100°, corresponding to half of the positions (four) altered. These values should be compared with the data contained in Table I. The first two columns show for the α-helical motif (0/8) that the average distance of the next three (23.5°) and the next five structural motifs (24.8°) is considerably smaller than the value expected for a single-site alteration (50°). Indeed only two of the five neighbouring structures in Figure 4(a) show considerable deviation in one position. For the strand motif (6/2) the average distance of the neighbours approximately corresponds to a single-site alteration. This interpretation is confirmed in Figure 4(a), where most of the motifs around neuron (6/2) have one position changed from strand to helix. The mean distance from one neuron to an adjacent neuron is comparable to a single-site alteration (58.4°) (Table I). The average distance for all motifs in the feature map is 88° which is comparable to the alteration of three positions. Of course, a discussion in terms of a two-state system cannot yield more than a rough idea of what the network actually learned since in principle the whole range of possible angular values is accessible to the network. Indeed, a number of important motifs found by the network (e.g. turn or loop structures) are not included in the simple two-state model. However, the two-state model makes possible an easy interpretation of the mean distance values and leads to the conclusion that the Kohonen algorithm was well suited for clustering structural motifs according to the similarity of their dihedral angles.

Abundant and rare backbone motifs

The consensus motifs defined by the neurons' weight vectors differ in their abundance in the training data. A histogram

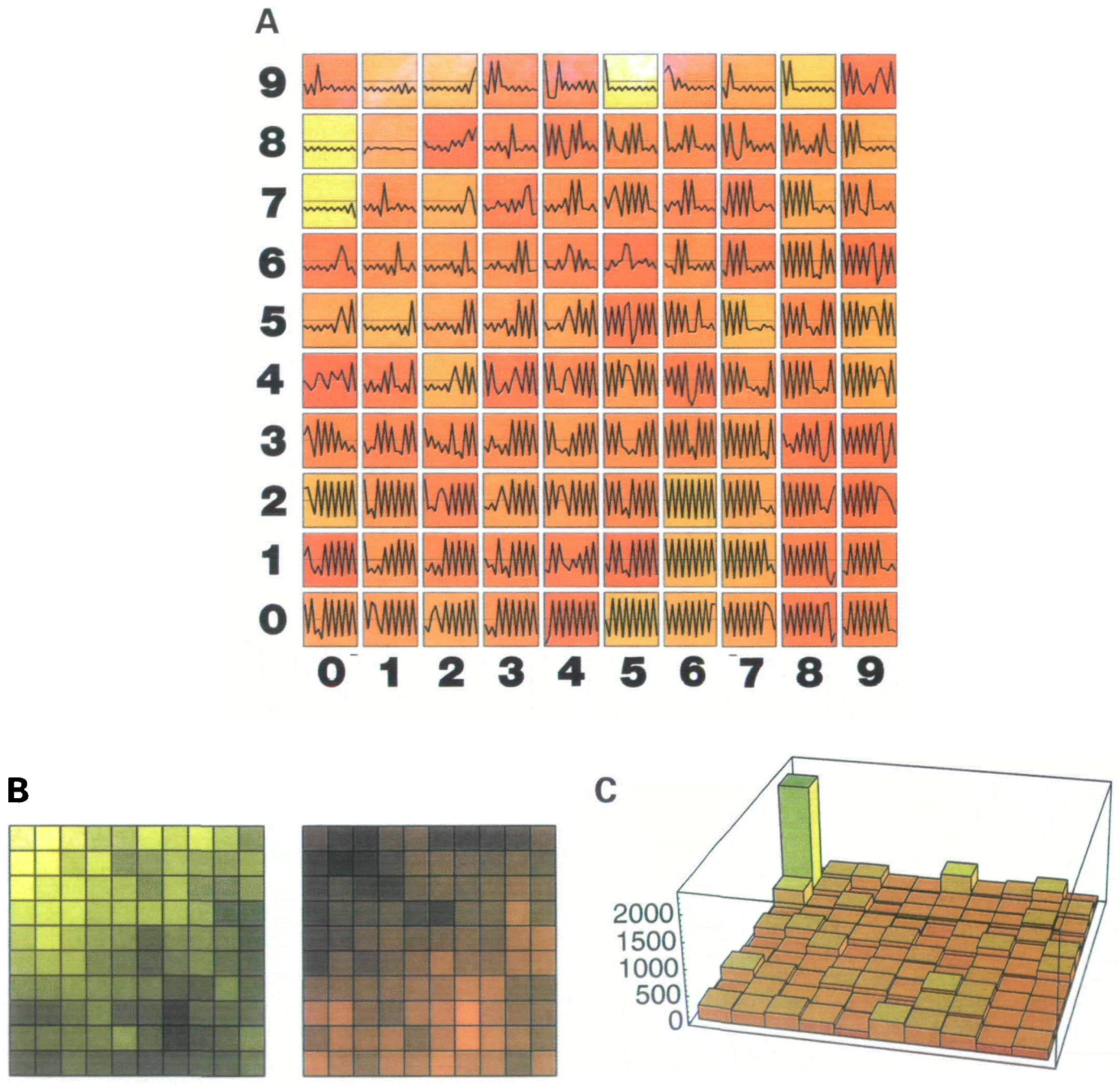

Fig. 4. (a) Kohonen feature map. Local structural motifs found by the Kohonen algorithm are shown by 100 Ψ–Φ plots. The following striking motifs can be identified: an α-helix in neuron (0/8), a β-strand in neuron (6/2) and a β–turn–β–motif in neuron (5/4) The colours of the squares indicate the frequency of occurrence of the particular structural element. Light yellow, many examples; dark red, few examples. (b) Similarity map of consensus structural motifs. Left: each square shows the similarity of the corresponding structural motif to the ideal α-helical structure given in Figure 3(a). Right: similarity of each structural motif to the ideal β-strand given in Figure 3(b). Intense yellow stands for a helical motif, intense orange for a β-strand motif. (c) Histogram with the number of examples from the training set assigned to each consensus structure. The large bar corresponds to neuron (0/8) with a helical motif.

gives the frequency of the consensus structures in the training set (Figure 4c). It was constructed by assigning the set of most similar training patterns to each neuron. The regular helical motif of neuron (0/8) is very common (2349 examples) whereas for the strand motif of neuron (6/2) only 382 examples were found. This is still, however, above the average of 232 patterns per neuron. The strand-type elements are distributed among more neurons than those with a helical motif. This reflects the remarkable range of average length and structural variability among β-strands (Orengo and Thornton, 1993; Flower, 1994).

Structural elements are compared with neural motifs

What is the 3-D shape of the structural motifs found by the network and how similar are the structural elements used for training? In Figure 5(a–h) the 10 most similar examples found in the training data are fitted to the consensus structures (blue ribbons) of neurons (0/8), (6/2), (9/8), (0/6) and (5/4). All 3-D representations were calculated from the dihedral Ψ and Φ angles. The Ω angle was kept fixed to 180° and examples were checked not to contain a *cis* X-Pro group. The best fits are shown by the red lines representing the backbone structure

Table I. R.m.s.d. on angle of some structural motifs contained in the feature map from surrounding motifs

	Next three motifs	Next five motifs	All motifs
Motif (0/8)	23.5°	24.8°	91.9°
	Next four motifs	Next eight motifs	All motifs
Motif (6/2)	44.8°	52.8°	86.0°
All motifs	58.4°	62.4°	88.0°

First column: the average distance of the adjacent neighbours to the helical motif (0/8), the average distance of the adjacent neurons to the strand motif (6/2) and the average distance value calculated over all motifs. Second and third columns: average distances considering the five motifs surrounding motif (0/8) and the eight motifs surrounding motif (6/2), and considering all motifs of the network.

of the corresponding training pattern. For convenience the r.m.s.d. on distance δ was calculated for the best fitting example, but it should be kept in mind that inclusion of the Ω angle may lead to further distortions in 3-D space (see Discussion).

Neuron (0/8) clusters helical elements (Figure 5a). The best fit is a fragment of D-galactose binding protein (2GBP) covering positions 49–58. The r.m.s.d. on Ψ and Φ dihedral angles ε (see Materials and methods) from the consensus motif is 2.4°. The corresponding r.m.s.d. on distance is $\delta = 0.09$ Å. To gain insight into the divergence within the helical cluster the average r.m.s.d. on angles of all helical elements from their consensus motif was calculated yielding $\bar{\varepsilon} = 10.9°$.

Neuron (6/2) represents regular β-strand elements (Figure 5b). The closest fit stems from human class I histocompatibility antigen AW 68.1 (2HLA, residues 4–12) with $\varepsilon = 10.3°$ ($\delta = 0.48$ Å). The structural element belongs to one of the central β-strands forming the 'floor' of the α_1 domain containing the antigenic binding site (Branden and Tooze, 1991). Obviously the Kohonen algorithm was able to recognize α-helical structures very precisely. Strand structures are less well defined and the algorithm reproduced this type of structure with a larger error. This observation is confirmed by calculating the divergence within the strand cluster yielding $\bar{\varepsilon} = 22.10$. This value is about twice as large as for the helical cluster, which is consistent with the mean distance of next neighbours shown in Table I.

A transition from strand to helix is observed in neuron (9/8) (Figure 5c). This structural feature is located, e.g. at the N-terminus of subtilisin (1S01, residues 3–11). Subtilisin is a typical α/β protein. Despite its complex shape the (9/8)-structure is represented in the training patterns with a comparable small deviation ($\varepsilon = 9.9°$, $\delta = 0.58$ Å) to the more uniform single-strand motif of neuron (6/2).

According to the Ψ–Φ plot, the motif (0/6) begins with a fragment of α-helical structure and ends with an irregular loop (Figure 5d). The best fit to this structure was found in the region 28–36 of the phosphate-free ribonuclease A (7RSA). While the deviation remains small along the helical part, it increases in the loop region leading to a total fit error of $\varepsilon = 15.8°$ ($\delta = 1.7$ Å).

Another complex helix–loop combination is described by neuron (3/8) (Figure 5e). The shape of the Ψ–Φ plot (cf. Figure 4a) suggests an interpretation as a helix–loop–helix motif. The best example fitting to this motif was found in deoxyhemoglobin (1SDH, residues 24–32) from *Scapharca* with $\varepsilon = 12.80$ ($\delta = 0.88$ Å).

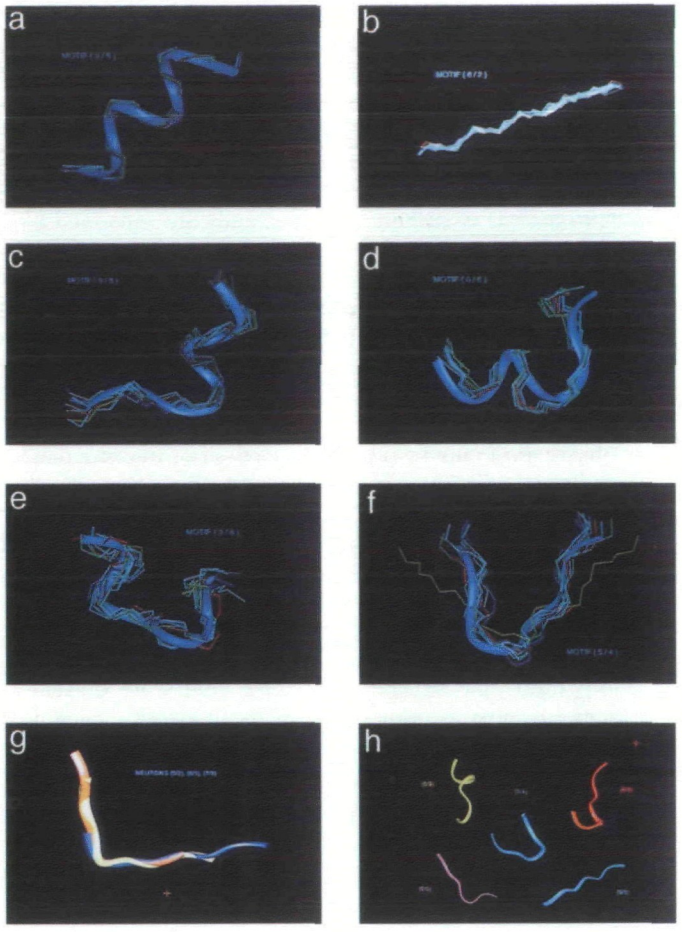

Fig. 5. Illustration of some striking structural consensus motifs found by the network together with the ten closest representatives in the training data. The blue ribbon always represents the consensus structural motif found by the neuron; the red line gives the main-chain of the most similar training pattern. (**a**) The helical motif represented by neuron (0/8). Corresponding to the large bar in the histogram this helical motif is the one most frequently found in the training data. Closest match found in D-galactose binding protein (2GBP, residues 49–58). (**b**) Motif of neuron (6/2) representing a β-strand. The most similar training pattern was found in human class I histocompatibility antigen AW 68.1 (2HLA, residues 4–12). (**c**) Strand–helix transition of neuron (9/8). The closest training pattern was found in subtilisin (1S01, residues 3–11). (**d**) Helix–loop transition of neuron (0/6). The most similar training pattern was found in phosphate free ribonuclease A (7RSA, residues 28–36). (**e**) Helix–loop–helix combination of neuron (3/8). The closest representative was found in dimeric deoxyhemoglobin (1SDH, residues 24–32). (**f**) Strand–loop–strand motif of neuron (5/4), with the closest example stemming from pea lectin (2LTN, residues 53–61). One of the fitted structures (green line) shows considerable deviation in 3-D space while the r.m.s.d. in terms of Ψ–Φ angles is comparably small. (**g**) Ribbon representation of the motifs found by neurons (5/2), (6/3) and (7/3). A corner of 90° seems to move along the sequence. The appearance of several shifted motifs may be a consequence of the moving window technique for training pattern generation. (**h**) Assembly of the consensus loop motifs from the corners (0/0), (0/9), (9/0) and (9/9) of the feature map and motif (5/4) from the centre

Motif (5/4) consists of two small β-strands connected by a loop (Figure 5f). The ribbon display shows that the two strands are bent by almost 180° strongly resembling a 'hairpin motif'. The closest fit stems from pea lectin (2LTN, residues 53–61) with $\varepsilon = 19.8°$ ($\delta = 1.0$ Å). The consensus motif is very similar to a β–turn–β structure. Remarkably, one of the fitted examples is rather close to the neuron structure according to

the Ψ–Φ angles but deviates considerably in 3-D space (green backbone structure: endothia aspartic proteinase 5ER2 residues 97–102). This results from a marked difference of a single dihedral angle in the vicinity of the turn, while the two β-strands are fitted fairly well according to their Ψ–Φ main-chain angles.

The similarity of the motifs of neuron (5/2), neuron (6/3) and neuron (7/3) is striking: two β-strands are separated by a loop, but the loop is shifted to different positions within the motifs represented by the three neurons. To decide whether their corresponding 3-D structures are similar ribbon models were superimposed as shown in Figure 5(g). The orange ribbon belongs to neuron (5/2) with a 90°-corner motif. The corner located close to the central residue in the orange ribbon is shifted more to one end in the yellow and the blue ribbons. This observation may be due to the method of training pattern generation: a window was slid along the sequence thus reproducing the motif in several different positions. The close localization of similar motifs is, however, a consequence of the topology conservation within the map.

An assembly of several unique structural elements derived from the neurons in the corners and the middle part of the feature map is given in Figure 5(h). A distorted β-strand, a β-loop-β motif and several other loop motifs are presented. These motifs give an impression of the network's capability to extract stranger patterns from the data that may not be described in simple terms of helix and strand.

Local sequence patterns are compared by structural alignment

Analysis of the amino acid sequences clustered by a neuron sometimes reveals patterns of conserved amino acid residues. Figure 6 gives an example for 30 sequences forming a β–turn–β motif.

Most of the nine-residue windows of neuron (5/4) have a glycine in the central position (24 out of 30). This sequence position is characterized by two subsequent positive dihedral angles (see Figure 4a). Glycines are known to be favoured in some tight turn structures with one or two positive Ψ angles, e.g. type II′ turns, since they do not cause steric hindrance (Richardson *et al.*, 1992). This observation is clearly reflected by the sequences clustered in neuron (5/4). Sequences 14 and 15 and also 17 and 21, stem from the human rhinovirus and from trypsinogen, respectively. This is a consequence of a partial redundancy in the data (see Discussion).

A similar analysis was carried out for neuron (2/4) representing a helix–strand transition: again most of the sequences (26 out of 30) have a glycine in position 6. This is in remarkable agreement with the helix-C-cap rule (Schellman, 1980; Richardson *et al.*, 1989). Clustering by a neuron can be regarded as similar to local structural alignment of amino acid sequences without gaps (Bowie *et al.*, 1991), with the difference that the leading structure (the 'consensus motif') is of artificial nature rather than representing a 'real' structural motif.

Structural classes are identified by tracing protein backbones

To illustrate the classification of local structural elements for a single protein, the complete Ψ–Φ plot of cytochrome b_5 (3B5C) is shown in Figure 7. This protein belongs to the α/β class and contains a number of different local structures. The 3-D structure of cytochrome b_5 is given in Figure 8(IIIa) together with three other examples. To classify the structural elements along the main chain the sequence of dihedral Ψ–Φ angles was scanned by a sliding window of 16 angular values.

```
 1: 2LTN (153-161): KLQNGEEAN 19.84
 2: 1PAZ ( 57- 65): KSKINENYV 21.54
 3: 1FKF ( 15- 23): FPKRGQTCV 21.70
 4: 6CPP (311-319): QLKKGDQIL 22.81
 5: 5ER2 ( 98-106): LTVTGQAVE 23.51
 6: 3EBX ( 45- 53): TVKPGIKLS 24.59
 7: 6DFR ( 52- 60): RPLPGRKNI 24.89
 8: 3AIT ( 47- 55): AVAPGQITT 24.92
 9: 2PCY ( 20- 28): SISPGEKIV 24.94
10: 4CMS (163-171): MLTLGAIDP 25.08
11: 1CD4 (  5- 13): LGKKGDTVE 25.12
12: 6CPP (382-390): SIAPGAQIQ 25.18
13: 3DFR ( 68- 76): YQAQGAVVV 25.44
14: 4RHV (152-160): YVPPGAPNP 25.56
15: 1R09 (152-160): YVPPGAPNP 25.60
16: 4CMS ( 94-102): IVDIQQTVG 26.53
17: 2TGP (124-132): CASAGTQCL 26.65
18: 1DHF ( 65- 73): RPLKGRINL 26.96
19: 5ER2 (167-175): TYNFGFIDT 27.11
20: 1PAZ ( 24- 32): KANPGDTVT 27.94
21: 1TGS (124-132): CASAGTQCL 28.05
22: 5HVP ( 57- 65): RQYDQILIE 28.15
23: 2I1B ( 60- 68): LGLKEKNLY 28.24
24: 1HNE (119-127): LPAQGRRLG 28.51
25: 1TNF (125-133): QLEKGDRLS 28.91
26: 1UBQ ( 60- 68): NIQKESTLH 28.95
27: 4GR1 (244-252): EVLKFSQVK 29.24
28: 1LAC (311-319): ANKPGDVVR 29.35
29: 1MCP ( 43- 51): QQKPGQPPK 29.56
30: 2FNR ( 69- 77): PYREGQSVG 29.70
```

Fig. 6. Sequences of the 30 structural elements most similar to neuron (5/4). From left to right the sequence numbers, the PDB identifiers, the sequence positions (in parentheses), the amino acid sequences, and the r.m.s.d. on angles from the consensus neuron motif are listed. The similarity to the consensus motif decreases from the top to the bottom. The grey box marks the central window position which is occupied by a Gly in most cases.

The structure of each segment covered by the sliding window was compared with the motifs of the feature map. The most similar motif of the feature map according to the r.m.s.d. on angles was selected and the neuron's label was aligned below the amino acid sequence (Figure 7).

The neurons (8/7), (7/5), (9/8) and (8/9), e.g. represent strand–helix transitions (cf. Figures 4a and 5c), thus correctly classifying the initial two structures of cytochrome b_5: a small piece of β-strand and the small helix α_1. A ribbon representation of cytochrome b_5 is given in Figure 8(IIIa). Helix α_1 is followed by a larger loop extending over the residues 13–19 and a β–turn–β motif. The loop region is mapped onto neurons (4/1) and (4/3) representing loop structures. Here, one limitation

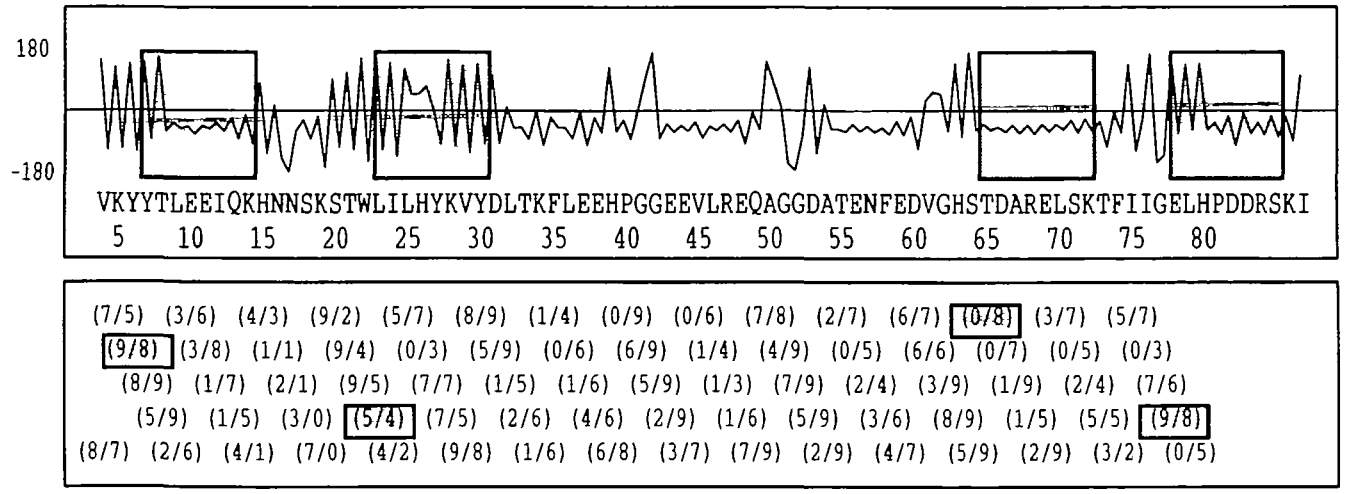

Fig. 7. Active neurons tracing the main-chain of cytochrome b_5. Above: Ψ–Φ plot showing the main-chain dihedral Ψ and Φ angles aligned with the sequence of amino acids. Below: list of structural motifs called up tracing the backbone with a nine-residue window.

Table 2. Structural matrix for the five proteins myohaemerythrin, cytochrome b_{562}, T4 lysozyme, cytochrome b_5 and γ IV crystallin

		2MHR	265B	1L58	3B5C	2GCR
Myohaemerythrin	2MHR	1.00	0.69	0.59	0.36	0.10
Cytochrome b_{562}	265B		1.00	0 54	0.35	0.12
Lysozyme	1L58			1.00	0.54	0.25
Cytochrome b_5	3B5C				1.00	0.26
γ IV crystallin	2GCR					1.00

of the network containing only 100 neurons appears: the representation is not very close to the crystallized structure and a larger network may be able to perform a better fit. The β-turn-β motif between positions 24 and 32 is mapped to the turn motif of neuron (5/4) shown in Figure 5(f). Following the list of motifs neuron (9/8) appears two more times for sequence positions 29–37 and 79–87, indicating strand–helix transitions to the helices α_2 and the C-terminal helix (Figure 7). This is in accordance with the X-ray structure (Figure 8(IIIa). Also, the helical segments α_3 and α_4 are correctly classified, e.g. by neurons (7/9) and (2/7). The only helical segment long enough to be mapped to the all-helical motif (0/8) is the helix α_5 extending over almost three helical turns (residues 65–73). The following two motifs (0/7) and (1/9) are both variations on ideal helical motifs (Figure 7).

Tracing the backbone of cytochrome b_5 demonstrates that the small 10×10 Kohonen network is sufficient for a sensible classification of local structural elements. It is, however, not precise enough to describe a protein in great structural detail. This is especially valid for loop regions.

Classification in terms of local structural elements was also used for a structural comparison of entire proteins. Four different proteins were analyzed, cytochrome b_{562} (256B), T4-lysozyme (1L58), cytochrome b_5 (3B5C) and γ IV crystallin (2GCR). The four proteins were arranged from an all-helical type to an all-β type (Figure 8, rows I–IV). Cytochrome b_{562}, a typical four helical bundle protein, and γ IV crystallin built up from two β-sheets with four antiparallel β-strands each are most different in their tertiary structure. From a comparison of the histograms, the structural difference between these two proteins becomes even more obvious (Figure 8Ib and IVb). Most local structural elements in cytochrome b_{562} are mapped to the all-helical motif (0/8) whereas for γ IV crystallin the variety of local structural elements is indicated by a multi-peaked histogram. The trajectory of cytochrome b_{562} (Figure 8Ic) almost completely avoids neurons representing β-strand type motifs while in γ IV crystallin (Figure 8IVc) this is true for the upper left neurons.

In the histograms lysozyme and cytochrome b_{562} appear to be very similar owing to the high α-helical content (Figure 8Ib and IIb). However, the corresponding trajectories look different owing to the loops present in the lysozyme structure (Figure 8IIc). In the trajectory of cytochrome b_5 the central β-sheet region can easily be recognized (Figure 8IIIc).

A structure-based protein similarity matrix

The histograms derived from backbone tracing may be used to define a structural similarity measure for comparison of proteins. The similarity measure is based on the relative content of secondary structural elements found in two proteins and calculated by subtracting the normalized values of their histograms (cf. Materials and methods). Proteins with a similar relative content of secondary structural elements will have a simililarity value close to one, while those with a very different content of structural elements have a value close to zero. Five proteins were compared by this method (Table II). Myohemerythrin (2MHR) and cytochrome b_{562} are four-helical bundle proteins, lysozyme is an α + β protein, cytochrome b_5 belongs to the α/β class and γ IV crystallin represents a characteristic Greek-key motif composed of β-sheets (Richardson, 1981; Hutchinson and Thornton, 1993). Myohemerythrin and cytochrome b_{562} turned out to be the most similar pair (similarity value of 0.69). γ IV crystallin reveals the smallest similarity to the two helical proteins, cytochrome b_{562} and hemerythrin, with similarity values of 0.1 and 0.12. This is in accordance with the overall structure of γ IV crystallin.

Discussion

Kohonen networks already have been successfully applied to protein sequence classification (Ferrán *et al.*, 1994), spectrum analysis, e.g. evaluation of secondary structure of proteins

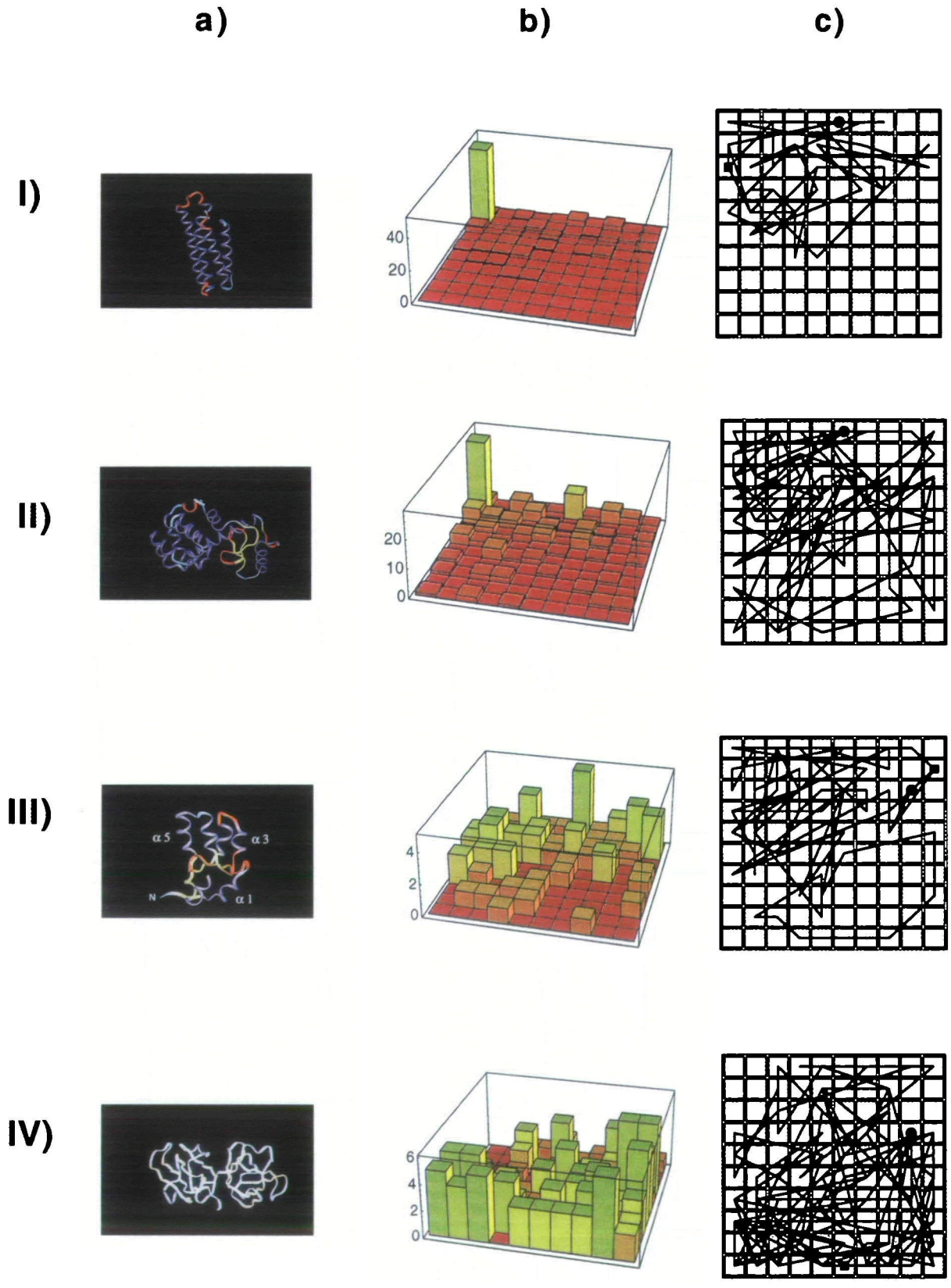

Fig. 8. (**a**) Tertiary structure models, (**b**) histograms and (**c**) trajectories of cytochrome b_{562} (I), T4 lysozyme (II), cytochrome b_5 (III) and γ IV crystallin (IV). A circle marks the begin of a trajectory and a square its end. Note the different z-scaling of the histograms. Trajectories and histograms reveal clearly the difference between the α-helical cytochrome b_{562} and γ IV crystallin with a large content of β-strand

from circular dichroism spectra (Andrade *et al.*, 1993), and mapping and classification in organic chemistry (Gasteiger and Zupan, 1993). Here, the Kohonen algorithm was applied to find consensus structural motifs from a set of 136 proteins. All structures were encoded by their main-chain Ψ and Φ dihedral angles (Figure 3), and the consensus structures were

assembled in a feature map (Figure 4a). Two main regions corresponding to more helical and more β-strand-like structures are striking. The topological ordering of the motifs shows that the algorithm was able to perform a non-linear projection of the complex space of structural elements onto a two dimensional map (Figure 4b).

Predominant structural motifs such as α-helix, β-strand and coils are localized in separate regions in the two-dimensional map. In particular, the neurons representing the helical and β-strand motifs are clearly separated as defining two essential structural 'superclasses'.

Erroneously, three identical structures (1ECA and 1SDH, 1TGP and 1TGS, 1RHV and 1R09) were included in the version of the PDB-Select data set used for network training (U.Hobohm, personal communication). This explains the observation of some identical sequence segments made in the section. Local sequence patterns are compared by structural alignment. However, this over-representation is expected to have only vanishing influence on the statistics of local structural elements. More severely, the data set might be a biased selection of all proteins, namely towards those which are crystallizable. This may be one reason for the pronounced appearance of helical and β-strand motifs in the consensus structures as well as in the training data since the space of possible consensus structures is limited by the structures appearing in the database. In addition to the two main classes, a variety of loop motifs appear in the feature map, leading to a rather differentiated classification whilst at the same time giving an impression of the manifold local structural elements within the 136 proteins.

How should a distance measure for comparing the structural elements be chosen? Any clustering algorithm requires a sensible distance measure for effective performance. Here the comparison of two structures is based on the r.m.s.d. of the dihedral angles ϕ and ψ (see Materials and methods). This is sensible in that a vanishing deviation in the angular values will result in an identical 3-D structure. A more refined choice would account for the periodic nature of the angular values replacing $\varphi-\varphi'$ by $|\varphi-\varphi|\mathrm{mod}(180°)$ and $\psi-\psi'$ correspondingly. This substitution leads to new results only in the case where the two angles ϕ and ϕ' to be compared are separated by the 180° border. This case is assumed to be of minor influence, since most of the calculated distances concern neighbouring structures and most of the angular values are located in some distance from the border considering the Ramachandran plot. The only amino acid for which some effect is to be expected is glycine, where a statistically significant part of angular values crosses the 180° border. A characteristic property of the angular distance measure is that even a small r.m.s.d. in the ϕ and ψ values can lead to substantial distortion in 3-D space. This will happen if the deviation is concentrated in one angle of the structural element. A deeper understanding of the relation of r.m.s.d. in angular and in 3-D space would require an extensive statistical analysis and the inclusion of ω angle information which lies beyond the scope of this investigation. Experiments including Ω angle information are currently being performed.

The network size of 100 neurons was selected for the following reasons: in the Ramachandran plot only three regions of angular values are significantly populated, the strand, the α_R and the α_L region. Since eight pairs of dihedral angles define a nine-residue structural element, the expected number of different structures is of the order of 3^8 (≈6600). Since the α_L helices are rare, in a rough estimation only $2^8 = 256$ different structural elements need to be considered. Concentrating on the most important structural elements a network size of 100 neurons is a reasonable choice with the advantage that

(i) the algorithm focuses on the most common sequence motifs thereby rendering insight into the principal organization of the space of structural elements;
(ii) the overview over all the consensus structures may be kept owing to an easy visualization of the feature map;
(iii) a reasonably compact description makes possible a graphical representation of single proteins by the histogram and the trajectory plot.

An appropriate selection of network size depends on the task to be performed: the rather small network shown here obviously is too small for a precise representation of all local structures appearing in the training data. In particular in loop regions only a qualitative representation was achieved. On the other hand the example of cytochrome b_5 shows that the description of the backbone conformation in terms of the structural motifs though not exact in all regions is very reasonable. This clearly demonstrates the potential of the Kohonen algorithm to extract relevant information from the training data. Experiments with larger networks and window sizes revealed additional and more complex loop structures, but the overall shape of the map remained unchanged. Systematic investigations altering network size and dimension have been performed and will be discussed elsewhere (manuscript in preparation).

The ϕ–ψ plots of the helix and the β-strand motifs of the neurons (0/8) and (6/2) are regular, which probably is a consequence of the algorithm averaging over a whole cluster of similar patterns to extract the consensus structural motif. As given by the histogram the consensus helical motif occurs almost 2000 times and the consensus β-strand almost 400 times (Figure 4c). As expected, the regular α-helix is far more frequent than any other helical motif. In contrast neuron (6/2) representing the most regular β-strand is surrounded by several other β-type motifs of comparable abundance. This is due to the fact that the helical segments in the data are larger and less frequently interrupted by loops than are the β-strands. This holds at least comparing the histograms of cytochrome b_{562} and γ crystallin: the four long helical segments of cytochrome appear in one pronounced peak of the histogram (Figure 8IIIb) while the eight β-strands forming the Greek-key motif of γ-crystallin are accompanied by a number of loop motifs leading to a broader distributed histogram (Figure 8IVb). A protein composed of a few uniform elements will thus result in a histogram completely different from that of a protein composed of many small segments such as γ crystallin and cytochrome b_5.

An important question is whether the clusters formed by the network can yield deeper insight into the sequence–structure relation and, even more interesting, the structure–function relation. The analysis carried out for two of the loop-containing neurons led to the rediscovery of two important rules for the role of glycine in protein structures: the 'C-cap rule' and the 'glycine motif' in tight turns. Notably these observations could be made without any sophisticated data analysis. A more refined analysis of the clustering done by the network is expected to yield further insight into the sequence–structure relation.

The combination of structural and sequence information

provided by a trained Kohonen network may also be useful for the design of new functional proteins. Design work employing self-organizing networks is currently being performed.

Acknowledgements

The authors are grateful to Arnaud Ducruix, Wolfram Saenger and Gerhard Müller for helpful discussions and encouragement. This work was financially supported by the BMBF (DETHEMO project), the DFG (Graduiertenkolleg 'Signalketten lebende Systeme'), Freie Universität Berlin (Sondermittel) and the FCI (to G.S.).

References

Andrade,A., Chacón,P., Merelo,J.J. and Morán,F. (1993) *Protein Engng*, **6**, 383–390.
Bernstein,F.C. *et al.* (1977) *J. Mol. Biol.*, **112**, 535–542.
Bowie,J.U.R , Luthy,R. and Eisenberg, D. (1991) *Science*, **253**, 164–170.
Branden,C. and Tooze,J. (1991) *Introduction to Protein Structure*. Garland Publishing, New York.
Efimov,A.V. (1994) *Structure*, **2**, 999–1002.
Ferrán,E.A., Pflugfelder,B. and Ferrara,P. (1994) *Protein Sci.*, **3**, 507–521.
Fidelis,K., Stern,P.S., Bacon,D. and Moult,J. (1994) *Protein Engng*, **7**, 953–960.
Flower,D.R. (1994) *Protein Engng*, **7**, 1305–1310.
Gasteiger,J. and Zupan,J. (1993) *Angew. Chem.*, **105**, 510–536.
Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.
Hutchinson,E.G. and Thornton,J.M. (1993) *Protein Engng.*, **6**, 233–245.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Kohonen,T. (1977) *Self-Organization and Associative Memory*. Springer, Berlin
Kohonen,T. (1990) *IEEE*, **78**, 1464–1480.
Leszczynski,J.F. and Rose,G.D. (1986) *Science*, **234**, 849–855.
Levitt,M. and Chothia,C. (1976) *Nature*, 261, 552–542.
Levitt,M. and Greer,J. (1977) *J. Mol. Biol.*, **114**, 181–293.
Orengo,C.A. and Thornton,J.M. (1993) *Structure*, **1**, 105–120.
Prestrelski,S.J., Williams,A.L. and Liebman,M.N. (1992) *Proteins*, **14**, 430–439.
Ramachandran,G.N. and Sassiekharan,V. (1968) *Adv Protein Chem.*, **28**, 283–437.
Richards,F.M. and Kundrot,C.E. (1988) *Proteins*, **3**, 71–84.
Richardson,J.S. (1981) *Adv. Protein Chem.*, **34**, 167–399
Richardson,J.S. and Richardson,D.C. (1989) In Fasman,G.D. (ed.), *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum Press, New York, pp. 1–98.
Richardson,J.S. *et al.* (1992) *Biophys. J.*, **63**, 1186–1209.
Ritter,H. and Kohonen,T. (1989) *Biol. Cybernet.*, **61**, 241–254.
Rooman,M.J. and Wodak,S.J. (1988) *Nature*, **335**, 45–49.
Schellman,C. (1980) In Jaenicke,R. (ed.), *Protein Folding*. Elsevier, Amsterdam, pp. 53–61.
Schuchhardt,J., Gruel,J.C., Lüthje,N., Molgedey,L., Radons,G. and Schuster,H.G. (1992). In Schuster,H.G. (ed.), *Applications of Neural Networks*. VCH, Weinheim, pp. 239–249.
White,S.H. (1994) *Annual Rev. Biophys. Biomol. Struct*, **23**, 407–439.
Zhu,Z.Y. (1995) *Protein Engng.*, **8**, 103–108.
Zupan,J. and Gasteiger,J. (1993) *Neural Networks for Chemists. An Introduction*. VCH, Weinheim.

Received September 28, 1995; revised March 21, 1996; accepted March 26, 1996