PROTEINS: Structure, Function, and Genetics 42:460–470 (2001)

SOMCD: Method for Evaluating Protein Secondary Structure from UV Circular Dichroism Spectra

Per Unneberg,[1] Juan J. Merelo,[2] Pablo Chacón,[3#] and Federico Morán[4*]
[1]*Department of Biotechnology, Royal Institute of Technology (KTH), Stockholm, Sweden*
[2]*Departamento de Arquitectura y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, Granada, Spain*
[3]*Centro de Astrobiología, INTA-CSIC, Madrid, Spain*
[4]*Departamento de Bioquímica y Biología Molecular I, Facultad de Químicas, Universidad Complutense de Madrid, Madrid, Spain*

ABSTRACT **This article presents SOMCD, an improved method for the evaluation of protein secondary structure from circular dichroism spectra, based on Kohonen's self-organizing maps (SOM). Protein circular dichroism (CD) spectra are used to train a SOM, which arranges the spectra on a two-dimensional map. Location in the map reflects the secondary structure composition of a protein. With SOMCD, the prediction of β-turn has been included. The number of spectra in the training set has been increased, and it now includes 39 protein spectra and 6 reference spectra. Finally, SOM parameters have been chosen to minimize distortion and make the network produce clusters with known properties. Estimation results show improvements compared with the previous version, K2D, which, in addition, estimated only three secondary structure components; the accuracy of the method is more uniform over the different secondary structures. Proteins 2001;42:460–470.** © 2001 Wiley-Liss, Inc.

Key words: neural networks; prediction; secondary structure; unsupervised learning; circular dichroism

INTRODUCTION

A protein's structure is closely related to its function, and it is therefore of great interest to determine the structure of a protein macromolecule. The prevailing and most accurate method to determine the three-dimensional structure of a macromolecule is X-ray crystallography and, to date, some 7,000 proteins have been crystallized and classified structurally. However, not all proteins are readily crystallized, and even when possible, data collection is often tedious and time-consuming. Moreover, with X-ray crystallography, conformational changes in rapid-mixing experiments cannot be studied.

When three-dimensional information is not accessible, other methods that determine the secondary structure are employed. The secondary structure of a protein depends on its amino acid composition, whereby one approach is to estimate protein secondary structure from amino acid sequence. An alternative approach is to use circular dichroism (CD) spectroscopy, which measures the optical activity of a protein in solution. The optical activity of a substance depends on its structure, and indeed, CD spectroscopy does provide information about protein secondary structure. Moreover, data can be collected continuously, whereupon changes in structure due to ligand binding, for instance, can be monitored. These arguments show that CD spectroscopy is a valuable tool for studying the secondary structure of macromolecules and that it is of interest to develop methods that relate CD spectra to structure.

In this article, we have used a self-organizing neural network algorithm that manages to arrange CD spectra on a two-dimensional grid. When a CD spectrum is presented to the network, it is mapped onto a node of the grid. The location of the node in the grid provides information about secondary structure. The algorithm is based on Kohonen's self-organizing map (SOM) model,[1] in which data of high dimension are mapped nonlinearly onto a two-dimensional map. Hereby, intrinsic features in the input data can be extracted from location in the map.

The remainder of the discussion is organized as follows. An attempt is made to place the problem in context, describing Kohonen's self-organizing map, and providing some background on secondary structure prediction from circular dichroism spectra. We then discuss the choice of network parameters, and give details on the reference protein set and the accompanying secondary structure assignments. This is followed by a discussion of the results, which are divided into four parts. The clusters of the network are analyzed, followed by the construction of structure maps, accompanied by an example. All methods

An implementation of the method presented in this article will be available online at http://somcd.geneura.org.

#Current address: Department of Molecular Biology, The Scripps Research Institute, La Jolla, California.

*Correspondence to: Federico Morán, Departamento de Bioquímica y Biología Molecular I, Universidad Complutense de Madrid, 280 40 Madrid, Spain. E-mail: fmoran@solea.quim.ucm.es

Received 3 February 2000; Accepted 6 October 2000

described in this article are then compared, followed by a discussion.

DESCRIPTION OF THE PROBLEM

Before describing our particular solution to the problem of protein secondary structure prediction, a brief introduction to the functionality of neural nets is provided. The methods used so far in secondary structure prediction, including our own, are then reviewed.

Neural Nets

Neural nets (NN) are statistical techniques commonly used for pattern recognition, forecasting and scientific visualization.[2] It is impossible, and probably unfair, to generalize about current neural net algorithms, so we will focus only on the algorithm used in this article: Kohonen's self-organizing map.[3]

Kohonen's SOM can be described as a single-layer, feed-forward, nonsupervised neural net. As happens with all the other NN (and many statistical algorithms), SOMs must be "trained" before being able to perform whatever they were designed to do. Training means presenting a set of vectors to the NN, so that it changes its internal values. In this case, the values by which each sample is classified need not be set in advance, which is why it is called unsupervised.

A SOM is basically a set of *N*-dimensional vectors in which a neighborhood relation has been defined. They are arranged in a two-dimensional grid, in such a way that each vector or unit is the neighbor of six others. That is why each unit is represented as a hexagon (another possibility is each unit being a neighbor of another 8, in which case each unit is represented as a square and the map as a rectangle itself). A SOM must be trained for each task. All vectors are initially set to a small random quantity; training means selecting one vector from the training set, computing the unit closest to it, and changing that unit's vector and all the neighbors to make them closer to the input vector. The neighborhood arrangement makes the map self-organize, so that close units respond to contiguous zones in the input space, and since the neighborhood decreases during training, each unit gets fine-tuned to a particular zone in input space.

SOMs that have been trained with a particular training set perform several tasks at the same time:

Clustering: By analyzing its results, clusters, that is, natural groups, can be discovered in the data.

Nonlinear projection: This capability keeps metric distances.

In this article, both capabilities are used: the first to check that proteins with similar secondary structure are mapped close to each other, and the second to map unknown proteins and assign them secondary structure values.

Circular Dichroism

It has long been known that the ultraviolet (UV) CD spectrum of a protein in solution can be related to the overall secondary structure composition of the protein.[4–7] Given a CD spectrum, it is possible to determine the fractions of, for instance, α-helix (*H*), β-sheet (*E*), and β-turn (*T*) in the protein of interest. The problem has always been how to determine the contribution of each structure component to the final spectrum.

Most methods simply assume that each structure component produces a characteristic basis spectrum, or reference spectrum $b(\lambda)$.[5–7] More specifically, one assumes that a reference spectrum is obtained from a polypeptide consisting only of one secondary structure element (e.g., α-helix). Therefore, the CD spectrum $c(\lambda)$ of a protein can be reconstructed as the linear combination of the base spectra multiplied by the abundance of the respective structure elements:

$$c(\lambda) = \sum_{i=1}^{n} f_i b_i(\lambda) \quad (1)$$

Here, n is the number of secondary structure components, and f_i is the fraction of structure i in the protein.

The main problem with this approach is that the reference spectra are difficult to define.[8] The results for α-helical content are often satisfactory, because the spectra of proteins with high α-helical content are very similar to the α-helix reference spectrum.[9] However, the remaining structures, especially β-turn, pose difficulties.

Moreover, the assumption that a CD spectrum is a linear combination of secondary structure component spectra is not completely correct.[10,11] Aromatic groups may contribute to the overall spectrum, as may the interactions between amino acids far off in the sequence (tertiary interactions). Moreover, the contribution of α-helix to the spectrum depends on chain length. Therefore, various techniques, based on statistics, have been developed over the years that avoid using pure reference spectra[12] and that try to overcome the deficiencies of pure linear methods.

Ridge regression[8] reconstructs a CD spectrum from a basis set of CD spectra from proteins with known secondary structure compositions. A spectrum in the basis set that shows little resemblance to the problem spectrum contributes less to the final reconstruction. This method greatly improved the estimation of β-structures. However, it suffers the drawback that results depend on the proteins in the basis set. Variable selection[13] is based on the same ideas, except that a subset of the entire basis set is used in the reconstruction. The evaluation of protein conformation in solution is excellent. The disadvantage is that the method requires considerable computing time to find the subset that best reconstructs the problem spectrum. By doing some sort of averaging of the contributions from the reference spectra, the assumption of linearity can be overcome.[11] The selcon method[14] is a modification of variable selection. It improves speed by first arranging the spectra in the basis set in RMS distance from the problem spectrum.

An alternative technique, which does not use X-ray crystallographic data, is convex constraint analysis.[15] This algorithm calculates chiral spectra components from a set

of spectra. These chiral components are then used to reconstruct the problem spectrum. The evaluation of β-structures is poor compared with other methods. However, this technique is well suited for examining the spectra of proteins as a function of temperature, pH value, or ligand binding.

Some of the latest improvements include the estimation of six secondary structure categories[16] and the estimation of the number of α-helical and β-strand segments in proteins from CD spectra.[11] Recently, Pancoska et al.[10] have designed a matrix descriptor for secondary structure segments that estimates the connectivity and numbers of segments.

With the introduction of neural networks,[17,18] another nonlinear approach was enabled. With the program K2D, a neural network, based on Kohonen's self-organizing maps, is trained with a set of protein spectra. Training is done in an unsupervised manner, and the algorithm itself organizes the spectra in a meaningful way on a two-dimensional discrete output space. This eliminates the problem that the final reconstruction will depend on the basis set of protein spectra used, as in ridge regression. Moreover, once parameters that give good self-organization have been found, estimation of protein secondary structure from a problem spectrum is straightforward and instantaneous.

The methods described in this section were reviewed by Greenfield.[12] For a detailed discussion of CD spectroscopy, see the review by Woody.[19] The method presented in this article is based on the K2D algorithm,[18,20,21] which estimated the fractions of α-helix and β-sheet of a protein in the wavelength range 200–240 nm. With the SOMCD method, the lower wavelength limit has been extended to 190 nm. In addition, the evaluation of the fraction of β-turn has been included. These improvements have been made possible due to an addition of more proteins to the training set, which now includes 39 protein CD spectra plus an additional six reference spectra.

MATERIALS AND METHODS

The parameters that govern network performance are briefly described. In addition, the definition of distortion, a useful measure that provides information about network performance, is introduced. This discussion is not intended to be a self-contained theoretical description of neural network parameters. Such information can be found elsewhere.[1,3,22]

Network Parameters

The network is based on the self-organizing maps proposed by Kohonen.[1] It consists of an input layer of n neurons, each of them corresponding to one wavelength, and a square output layer of $m \times m$ neurons. A training set of CD spectra from proteins with known three-dimensional structure is presented to the network, and if suitable parameters are chosen, the algorithm organizes the CD spectra topologically onto the two-dimensional output layer.

The previous estimation method, K2D, used a square lattice of 13 × 13 neurons.[18] The training set consisted of 24 CD spectra, taken from Yang et al.[23] The size of the training set determines the output layer size. Consequently, a larger training set would make it possible to increase the number of output neurons, which would make the precision of prediction better. We have increased the training set, which now consists of 39 CD spectra from proteins with known three-dimensional structure as well as six reference spectra, all taken from previous articles.[13,23–26]

A valuable measure to evaluate network performance is distortion,[18,20] which can be defined as

$$D = \sum_{s \in sample} \|x_s - w_i\|^2 \qquad (2)$$

where x_s is the sample vector, and w_i is the winning neuron. Since the sample vector must not be included in the training set, the network is trained with 44 spectra, and the remaining spectrum serves as the sample. By letting all the spectra in the training set serve as the test sample in turn, an average distortion over the entire training set can be calculated. This value is used to estimate network size and other network parameters.

Distortion was calculated for different network architectures, output layer size and network parameters. A low value of the mean distortion indicates that training has proceeded well, but it is also necessary to examine the clusters that have formed (see the section, Results, Clustering). It is desirable for the euclidean distance between weights of neighboring neurons within clusters to be small, and that these weights are closest to spectra from proteins with similar secondary structure compositions.

With these guidelines, optimal network parameters and size of output layer could be set. Distortion did not vary much for networks of sizes from approximately 16 × 16 to 20 × 20 (data not presented here). An output layer size of 16 × 16 was therefore chosen since larger networks give longer calculation times. The remaining parameters were chosen so that the clusters fulfilled the above mentioned criteria. This step can be done more systematically, using, for instance, simulated annealing.

Learning Parameters

As pointed out earlier, during training, or learning, the training set is presented to the network a prescribed number of times. The network changes its internal values, or weight vectors, according to the following learning rule:

$$\delta w_{jk} = \alpha(x_i - w_{jk}) \qquad (3)$$

where w_{jk} is the weight vector located at grid position j,k; x_i is training vector i; and α is the learning rate parameter, a time-decreasing function that imposes convergence on the weights. The weights to be updated are the set of weights that are located in the neighborhood of the weight vector, which is closest to the current training vector. When training starts, the neighborhood function includes all neurons in the map and decreases with time, ultimately including only the closest neurons to the winning neuron. Training of the network was done with som_pak[22] (URL:

cochlea.hut.fi/research), a program package that implements the SOM algorithm.

Reference Protein Set

The proteins used in this article are presented in Table II. Also listed are the Protein Data Bank (PDB) files used to determine secondary structure, the abbreviated names of the spectra as they will appear later on in this work, the sources of the spectra, and the secondary structure values used for each protein.

The PDB files were chosen so that, if possible, the species of the PDB file matched the species of the protein used to collect the CD spectrum. Moreover, if more than one possible PDB structure was available for the same species, the one with the best resolution and the best *R*-value was chosen. Some crystal data have a resolution worse than 2.0 Å and an *R*-value of >0.20 (1a8m, 1eri, 1sbt, 2sbt, 9ldb), which most researchers regard as too poor for the protein to be included in the basis set.[16] In order to provide the network with as large a training set as possible, these were not discarded.

A word of notice is in order. Since the fractions of secondary structures in a protein add up to unity, an extra structure class can be defined by subtracting the fractions of the well-defined structure classes from one. In the K2D method, this meant subtracting the fractions of α-helix and β-sheet from one, and the third class was referred to as "random coil." This denomination is not entirely correct, since the unclassified amides of a protein do have some sort of defined static structure, although not belonging to any designated class of secondary structure. The term random coil refers to a dynamic peptide. In this article, amides not classified as belonging to any of the structures α-helix, β-sheet, or β-turn are referred to as other (*O*). Nevertheless, the random reference spectra are used as references for the "other" class. We use the term random reference spectra, since one of them is a CD spectrum of a free linear polypeptide, which clearly justifies the use of the word random. The other spectrum is a theoretical result from multilinear regression, and should in some sense be referred to as "other," but since the name random was used in the original article,[23] we stick to this denomination here. With this clarification of the use of both terms in the following, we now proceed with the secondary structure assignments.

Secondary Structure Classification

The secondary structure values were calculated following a variant of the DSSP[27] algorithm, which is implemented in the program Promotif.[28] The amino acid residues are labeled *H* and *h* for α-helix, *G* and *g* for 3_{10}-helix, *E* and *e* for β-sheet, and *T* and *t* for β-turn. Several assignments were used and tested in the method. The results presented are for the assignment giving the best performance. Residues assigned *H*, *h*, *G*, and *g* were classified as α-helix, residues assigned *E* were classified as β-sheet, residues assigned *T* were classified as β-turn, and residues assigned *t* and *e*, as well as unassigned residues, were classified as other (*O*). The percentages were then obtained by dividing the number of residues in each class by the total number of residues in the protein.

RESULTS

The results from our analysis of network clustering are presented. With the help of clusters, it is shown that the network has preserved structural information in the map. This allows for the construction of structure maps, which are the key to understanding how this method works. An example is included to shed some light on how the method works. This is followed by a comparison of the performance of other methods.

Clustering

One of the programs included in the som_pak package, umat, allows the visualization of the formed clusters using the Umatrix algorithm.[29] The output of the program umat for a run with the parameters set as indicated in the previous section is shown in Figure 1.

Each neuron is labeled with a three- or four-letter sequence. Neighboring neurons are separated by a hexagon, which is gray color shaded. The color of a hexagon separating neurons indicates the euclidean distance between the weight vectors of the neurons it joins; the darker the hexagon, the larger the distance. The euclidean distance *d* between two neighboring neurons with weight vectors *x* and *y* is calculated as

$$d(x, y) = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{1/2} \quad (4)$$

where *n* is the number of wavelengths considered. Consequently, clusters consist of white or light gray areas surrounded by darker-colored hexagons.

The labels in Figure 1 correspond to training spectra. Each training spectrum is provided with a unique letter sequence. The reference spectra all begin with a capital letter, with *A* for α-helix, *B* for β-sheet, and *R* for random (Table II). Visual inspection of Figure 1 verifies that a given training spectrum is locally mapped to a region of the network.

Two clusters are especially prominent, being located at the lower right and lower left corners of the network. They are separated from the rest of the map by streaks of dark hexagons. In the lower left corner, the weights have as closest training spectra both α-helix reference spectra (Ayan and Acur). Consequently, the neurons of this cluster have modified their weight vectors during training, making them similar to spectra from proteins with high α-helical content. Moreover, although the distances to the weight vectors of the neurons just outside the cluster seem to be large, judging by the color of the separating hexagons, it is interesting to note that also these neurons are labeled with spectra from all-α proteins (classification done according to SCOP.[30]) Among these are myoglobin (myo), hemoglobin (hem), and hemerythrin (hmr). Apparently, spectra from all-β proteins have been mapped to the lower left region of the network.

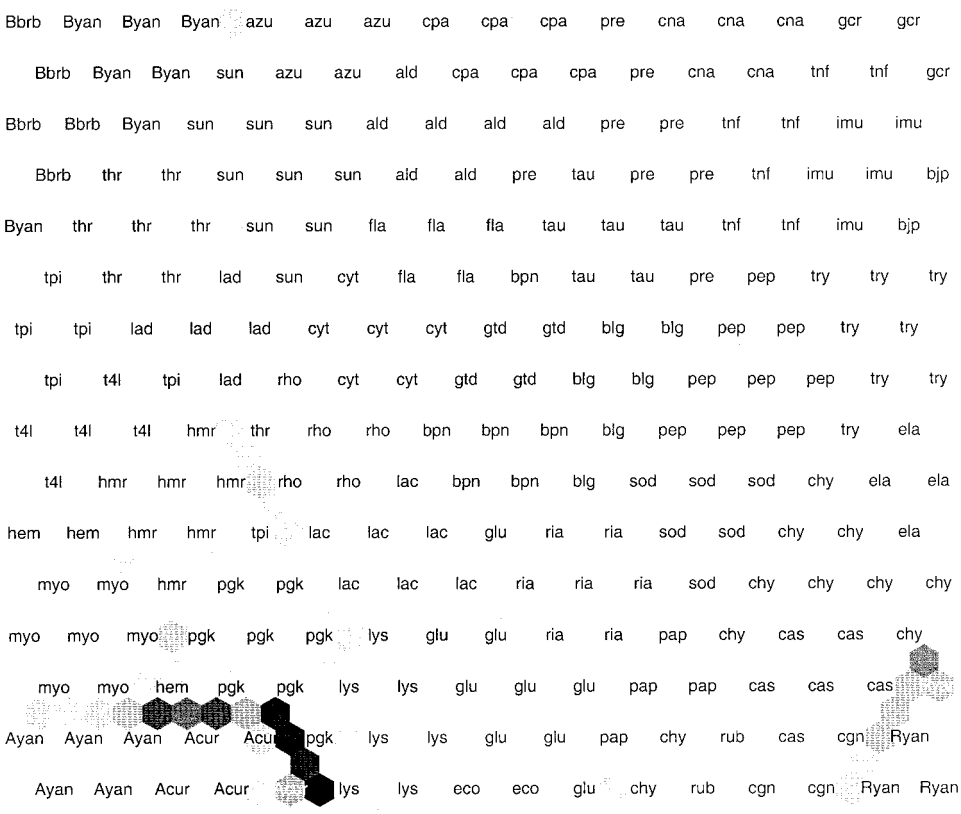

Fig. 1. Output of program umat, showing the nodes of the network surrounded by nonlabeled hexagons, which represent the distances between them. Each neuron is labeled with a three- or four-letter sequence, where the letter code represents the closest training spectrum to the neuron. The gray scales indicate the euclidean distance between the weight vectors of adjacent neurons. Darker hexagons indicate a larger distance.

In the lower right corner, a cluster of four neurons is visible. All neurons are labeled with one of the random reference spectra, indicating that the weights are similar to spectra from proteins in the other class. Although the second random reference spectrum does not appear as a label, it is the second closest training spectrum to all neurons in this cluster (data not presented).

In the upper left corner, a group of neurons are labeled with the β-sheet reference spectra. The fact that reference spectra are mapped to the corners of the network was noted already by Andrade et al.[18] and was to be expected. However, all-β proteins do not map to the upper left region of the map. Instead, all-β proteins with the highest β-sheet content of the proteins in the training set, such as γ-crystallin (gcr), immunoglobulin-λ (imu), thaumatin (tau), tumor necrosis factor-α (TNF-α) (tnf), and prealbumin (pre), are found in the upper right corner. As pointed out earlier, only the α-helix reference spectra show resemblance to spectra from proteins with high α-helical content.[9] Since the spectra of all-β proteins in the training set do not resemble the corresponding reference spectra, they are not able to form stable clusters with the reference spectra. All-β proteins with lower β-sheet content, such as pepsinogen (pep), elastase (ela), α-chymotrypsin (chy), carbonic anhydrase (cas), chymotrypsinogen A (cgn), and superoxide dismutase (sod), are mapped to the lower right corner of the map, thus surrounding the "other" cluster. Consequently, the rightmost region of the map corresponds to all-β proteins, excluding the "other" cluster.

In the remaining portions of the map, the distances between weights of neighboring neurons are relatively small. Here, mostly α/β and α + β proteins are found. Still, some anomalies can be found. The clusters of cytochrome c (cyt) and thermolysin(thr) are found in this region, with the thr cluster close to the cluster of the β-sheet reference spectra.

Despite the anomalies, the facts just presented indicate that the weight vector arrangement is related to structure. Therefore, structure information from the network remains to be extracted.

Structure Maps

After training, the algorithm has organized the spectra topologically (the terminology in Andrade et al.[18] was *proteinotopic mapping*) on a two-dimensional grid, in which each node corresponds to a weight vector, or a spectrum. The algorithm has constructed this spectrum by nonlinear interpolation of the spectra in the training set. Neighboring nodes display similar weight vectors if good self-organization has taken place.

Following the procedure developed by Andrade et al.,[18] structure maps can be constructed. The basic idea is that every node is assigned a complete set of structure fraction values. For the sake of clarity, structure values

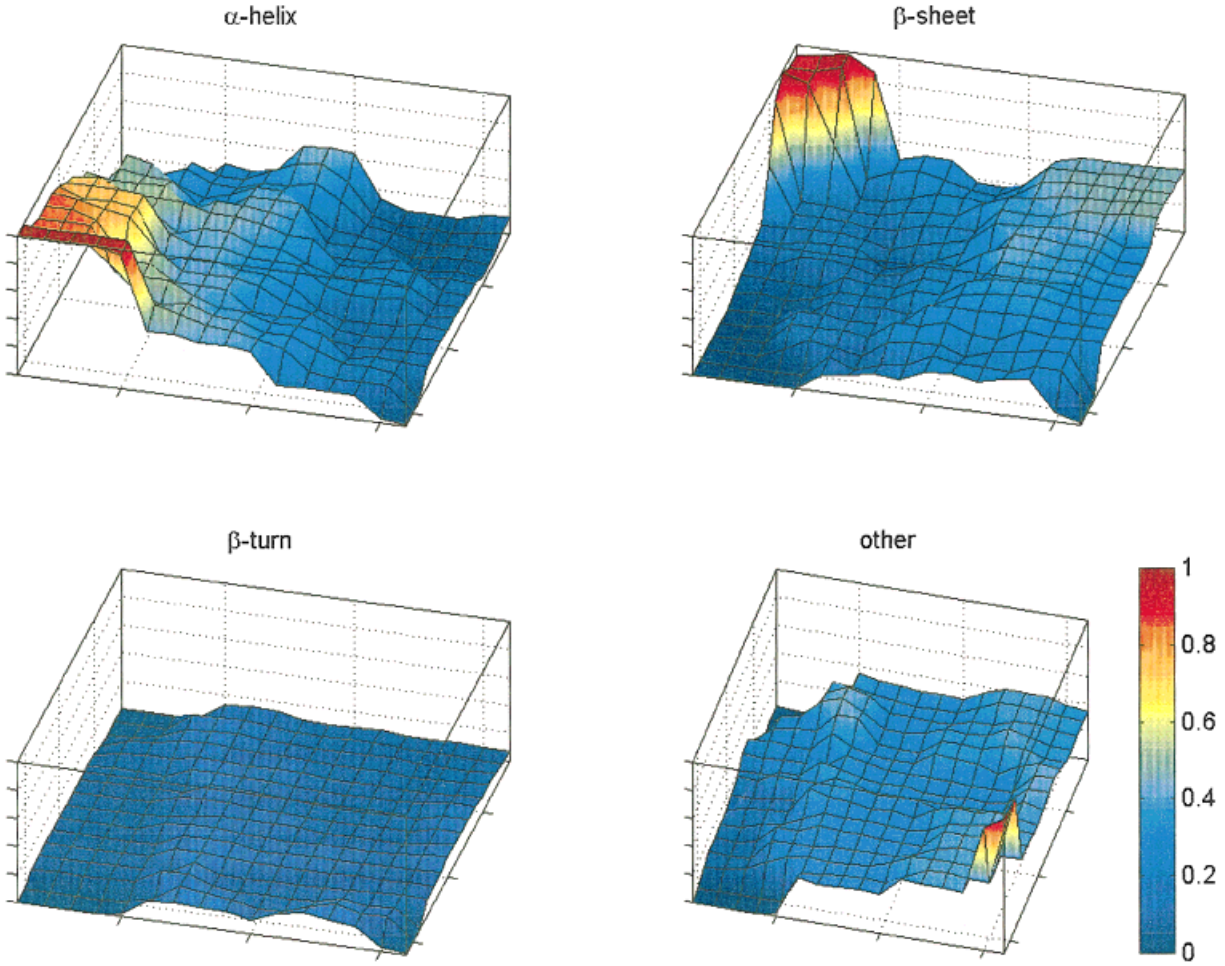

Fig. 2. Constructed structure maps using $L = 3$. Red implicates high neuron structure values, blue low values.

TABLE I. Structure Values Estimated for Thermolysin[†]

Secondary structure component[a]	Actual value	Estimate
H	0.48	0.45 ± 0.17
E	0.17	0.16 ± 0.03
T	0.09	0.08 ± 0.06
O	0.27	0.31 ± 0.14

[†]Standard deviation of secondary structure values is computed over the three structures used for averaging (see the section, Structure Maps).
[a]*H*, α-helix; *E*, β-strand; *T*, β-turn; *O*, other amides not classified as belonging to any of the previous categories.

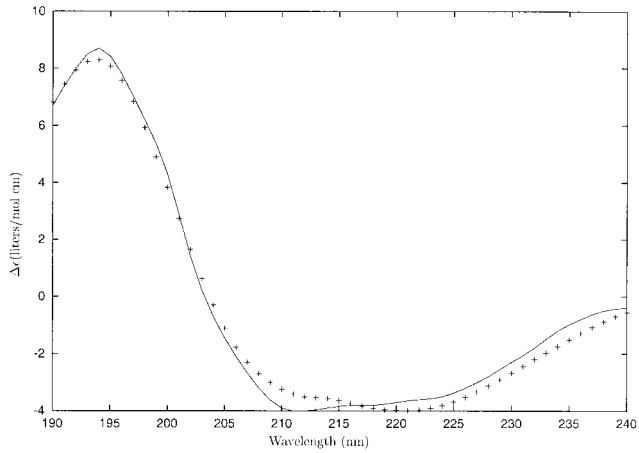

Fig. 3. Circular dichroism (CD) spectrum of thermolysin (−) and the weight vector of the winning neuron (+).

corresponding to neurons will be called neuron structure values. In our case, a complete set means the structures α-helix, β-sheet, β-turn, and other. For every weight vector, the spectra in the training set are organized in the order of increasing euclidean distance. The node is then assigned a combination of the structure fraction values of the L closest spectra in the training set. For example, in Figure 1, each label corresponds to the closest spectrum. The parameter L, which determines how much structural information is extracted from the training set, is called the *scope* of the method. More specifically, for structure s, the neuron structure value f_s for neuron j,k can be calculated as the mean of the structure values of the L closest spectra:

$$\bar{f}_{s,jk} = \frac{\sum_{i=1}^{L} \nu_i f_{s,i}}{N}, \qquad N = \sum_{i=1}^{L} \nu_i \tag{5}$$

where $f_{s,i}$ is the structure fraction value for the ith closest training spectrum, and ν_i is the inverse of the distance between the weight vector and the ith farthest training protein. Thus, the factor ν_i determines the contribution of the ith structure fraction value to the calculation of the neuron structure value. The more similar a training spectrum is to a weight vector, the more importance should be laid on the secondary structure composition of this protein.

The higher the scope, the more structural information is extracted from the training set. However, an excessively large scope will also introduce more noise. Calculation of the standard deviation of the expression in Eq. 5 gives the error of the corresponding structure value:

$$s_{jk} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{L} \nu_i (f_{s,i} - \bar{f}_{s,jk})^2}, \qquad N = \sum_{i=1}^{L} \nu_i \tag{6}$$

When a CD spectrum of a protein with unknown secondary structure is presented to the trained network, the neuron structure values of the winning neuron, derived from Eq. 5 are the estimated structure values of the protein. The error of the estimate is given in Eq. 6.

It is important to keep in mind which aspects of the method are based on linear approaches and which are based on nonlinear approaches. The neuron structure values of a neuron are a linear combination of the structure values of the proteins whose spectra are most similar to the weight vector of the neuron in question. However, the weight vectors are not linear combinations of the spectra in the training set. In other words, although the neuron structure values are assigned linearly, the weight vectors are obtained by nonlinear combinations of the spectra in the training set.

For the training set used in this work, a scope of three gave the best results. In Figure 2, the structure maps for secondary structures with $L = 3$ are shown. To facilitate visualization, the neuron structure values have been plotted separately for each secondary structure element.

These maps correspond to the network in Figure 1. Examination of the structure maps shows that for the β-sheet structure, high βvalues are found in two corners, one where the reference spectra are mapped, and one where all-β proteins are mapped. Nevertheless, it is important to include reference spectra, since they provide not only high extreme values, but zero values as well. This is important for the construction of complete structure maps.

Example

Suppose we want to estimate the structure values of thermolysin ($H = 0.48, E = 0.17, T = 0.09, O = 0.27$). The network is first trained with the training set, excluding thermolysin. Then, the weight vector that is closest to the thermolysin spectrum is chosen as the winning vector. The closest training spectrum to this weight vector is that of lactate dehydrogenase (LDH). If only one spectrum were to be used for the neuron structure values, the structure values of LDH ($H = 0.52$, $E = 0.18$, $T = 0.05$, $O = 0.24$) would be assigned the winning neuron, and consequently, also to thermolysin.

Since a scope of 3 has been used in this work (see above), the structure values of the following two closest spectra are used. The second closest spectrum is subtilisin novo ($H = 0.27$, $E = 0.14$, $T = 0.14$, $O = 0.46$), and the third closest spectrum is triose phosphate isomerase ($H = 0.57$, $E = 0.14$, $T = 0.05$, $O = 0.24$). The distances for the three closest spectra are 4.73, 6.45, and 8.34, respectively. Using the inverse values of the distances as the factor ν in Eq. 4, the final estimation of the secondary structure composition of thermolysin is as shown in Table I.

The standard deviation from Eq. (6) is given as the estimation error. Note that percentages do not necessarily have to add up to 1, since it is an approximation. In Figure 3, the winning vector is plotted with the CD spectrum of thermolysin.

Method Evaluation

The performance of the SOMCD method was evaluated by determining the secondary structure composition of all the proteins in the training set. The network was trained with all spectra except one, for which the secondary structure composition was determined. Success of structure estimation was obtained by comparing the estimated values with the X-ray crystallography data and calculating the RMS deviations and Pearson correlation coefficients.

The estimated structure values of the 39 proteins are shown in Table II. The worst results are obtained for cytochrome C (cyt). This is not surprising, since in Figure 1, the cluster of cyt is found in a region with α/β- and α + β-proteins, separated from the all-α region. When training the network without cyt, it is therefore likely that the neuron with the most similar weight vector will be found outside the all-α region, thus giving too high a β-sheet value in the estimation process.

The results of the evaluation are shown in Table III. Comparisons are made with the methods described in earlier in the section, Circular Dichroism.

Overall, SOMCD obtains a good prediction for all secondary structures except in the case of β-turn (when considering the correlation coefficient). Even in this case, results could be improved if the spectrum of β-turn possible values increases, as can be seen in Figure 4, where a histogram plot of the β-turn values in the training set is displayed. Evidently, most proteins have β-turn content close to 10%, which limits the range of possible β-turn values in the structure map.

DISCUSSION

This work presents a method that evaluates the secondary structure of a protein from UV circular dichroism spectra. It represents an improvement of the K2D algorithm[18] and also obtains results similar to those of other linear methods. More protein spectra have been added to the training set, the evaluation of the β-turn structure has

TABLE II. Estimated Structure Values Compared With X-Ray Structures[†]

Protein	PDB	CD[a]	Source[b]	Method	*H*	SD	*E*	SD	*T*	SD	*O*	SD
Alcohol dehydrogenase	2ohx	ald	P	SOMCD	0.40	0.05	0.22	0.03	0.12	0.01	0.26	0.03
				X-ray	0.33		0.24		0.11		0.32	
Azurin	2aza	azu	C	SOMCD	0.35	0.03	0.23	0.02	0.11	0.01	0.31	0.02
				X-ray	0.20		0.33		0.14		0.33	
Bence-Jones protein	1rei	bjp	C	SOMCD	0.10	0.02	0.47	0.01	0.09	0.00	0.34	0.02
				X-ray	0.05		0.49		0.12		0.34	
Bovine β-lactoglobulin	1beb	blg	C	SOMCD	0.22	0.03	0.38	0.03	0.11	0.00	0.29	0.01
				X-ray	0.24		0.43		0.09		0.25	
Carbonic anhydrase	1ca2	cas	P	SOMCD	0.15	0.01	0.29	0.03	0.17	0.02	0.40	0.01
				X-ray	0.21		0.29		0.11		0.40	
Chymotrypsinogen A	2cga	cgn	P	SOMCD	0.19	0.00	0.28	0.01	0.14	0.01	0.39	0.00
				X-ray	0.16		0.32		0.14		0.38	
α-Chymotrypsin	5cha	chy	C	SOMCD	0.15	0.02	0.33	0.02	0.13	0.01	0.39	0.01
				X-ray	0.14		0.33		0.12		0.42	
Concanavalin A (Con A)	1nls	cna	C	SOMCD	0.11	0.07	0.43	0.05	0.10	0.01	0.37	0.03
				X-ray	0.08		0.45		0.08		0.40	
Carboxypeptidase A	1arl	cpa	Y	SOMCD	0.39	0.04	0.23	0.02	0.11	0.00	0.27	0.02
				X-ray	0.46		0.16		0.09		0.30	
Cytochrome C	5cyt	cyt	C	SOMCD	0.39	0.01	0.24	0.03	0.11	0.01	0.27	0.02
				X-ray	0.50		0.00		0.15		0.36	
*Eco*RI	1eri	eco	C	SOMCD	0.46	0.02	0.13	0.03	0.16	0.01	0.26	0.01
				X-ray	0.38		0.19		0.13		0.30	
Elastase	1lvy	ela	C	SOMCD	0.17	0.02	0.31	0.01	0.12	0.01	0.41	0.01
				X-ray	0.14		0.34		0.15		0.37	
Flavodoxin	2fx2	fla	C	SOMCD	0.37	0.03	0.19	0.03	0.12	0.00	0.32	0.01
				X-ray	0.44		0.25		0.11		0.20	
γ-Crystallin	1amm	gcr	C	SOMCD	0.06	0.02	0.45	0.00	0.08	0.00	0.41	0.02
				X-ray	0.13		0.46		0.07		0.35	
Glutathione reductase	3grs	glu	P	SOMCD	0.34	0.03	0.26	0.04	0.12	0.01	0.27	0.01
				X-ray	0.41		0.24		0.09		0.27	
Glyceraldehyde 3-P dehydrogenase	1gd1	gtd	C	SOMCD	0.41	0.02	0.16	0.02	0.13	0.00	0.30	0.01
				X-ray	0.36		0.28		0.10		0.26	
Hemoglobin	1a3n	hem	C	SOMCD	0.80	0.02	0.03	0.01	0.06	0.00	0.12	0.01
				X-ray	0.86		0.00		0.06		0.09	
Hemerythrin	2hmq	hmr	C	SOMCD	0.75	0.00	0.08	0.00	0.06	0.00	0.12	0.00
				X-ray	0.77		0.00		0.06		0.17	
Immunoglobulin λ	8fab	imu	P	SOMCD	0.03	0.00	0.46	0.01	0.09	0.00	0.42	0.02
				X-ray	0.12		0.48		0.09		0.31	
Lactoferrin	1lcf	lac	P	SOMCD	0.41	0.01	0.12	0.01	0.10	0.00	0.37	0.02
				X-ray	0.40		0.18		0.16		0.26	
Lactate dehydrogenase (LDH)	9ldb	lad	C	SOMCD	0.51	0.04	0.15	0.00	0.07	0.01	0.27	0.03
				X-ray	0.52		0.18		0.05		0.24	
Lysozyme	3lzt	lys	C	SOMCD	0.38	0.01	0.20	0.02	0.12	0.01	0.30	0.01
				X-ray	0.50		0.06		0.19		0.25	
Myoglobin	1mbd	myo	C	SOMCD	0.86	0.06	0.01	0.01	0.04	0.02	0.09	0.04
				X-ray	0.86		0.00		0.07		0.08	
Papain	1ppn	pap	C	SOMCD	0.19	0.08	0.35	0.02	0.11	0.01	0.34	0.05
				X-ray	0.34		0.18		0.09		0.39	
Pepsinogen	3psg	pep	C	SOMCD	0.16	0.04	0.42	0.02	0.10	0.00	0.31	0.03
				X-ray	0.27		0.38		0.09		0.26	
3-Phosphoglyceric phosphokinase	1php	pgk	C	SOMCD	0.59	0.08	0.13	0.03	0.06	0.02	0.22	0.04
				X-ray	0.52		0.16		0.07		0.24	
Prealbumin	1tyr	pre	C	SOMCD	0.21	0.03	0.39	0.01	0.09	0.00	0.32	0.03
				X-ray	0.07		0.47		0.11		0.35	
Rhodanese	1rhs	rho	P	SOMCD	0.40	0.02	0.16	0.03	0.15	0.01	0.29	0.02
				X-ray	0.40		0.13		0.09		0.39	
Ribonuclease A	7rsa	ria	C	SOMCD	0.27	0.06	0.37	0.06	0.09	0.00	0.27	0.01

TABLE II. (Continued)

Protein	PDB	CD[a]	Source[b]	Method	*H*	SD	*E*	SD	*T*	SD	*O*	SD
				X-ray	0.26		0.33		0.13		0.28	
Rubredoxin	1iro	rub	C	SOMCD	0.19	0.01	0.30	0.00	0.12	0.00	0.39	0.00
				X-ray	0.17		0.15		0.26		0.42	
Subtilisin BPN′	1sbt	bpn	C	SOMCD	0.38	0.01	0.26	0.02	0.10	0.00	0.26	0.01
				X-ray	0.36		0.18		0.13		0.33	
Subtilisin novo	2sbt	sun	C	SOMCD	0.39	0.04	0.19	0.04	0.11	0.01	0.31	0.02
				X-ray	0.27		0.14		0.14		0.46	
Superoxide dismutase (SOD)	1sxn	sod	C	SOMCD	0.21	0.03	0.31	0.04	0.11	0.01	0.37	0.04
				X-ray	0.09		0.39		0.12		0.40	
T4-lysozyme	4lzm	t4l	C	SOMCD	0.69	0.05	0.06	0.03	0.05	0.00	0.19	0.02
				X-ray	0.77		0.09		0.06		0.09	
Thaumatin	1thv	tau	P	SOMCD	0.11	0.03	0.45	0.01	0.11	0.00	0.33	0.02
				X-ray	0.16		0.37		0.09		0.38	
Thermolysin	1hyt	thr	C	SOMCD	0.47	0.03	0.16	0.01	0.07	0.01	0.30	0.03
				X-ray	0.48		0.17		0.09		0.27	
Tumor necrosis factor (TNF-α)	1a8m	tnf	C	SOMCD	0.10	0.01	0.48	0.00	0.10	0.00	0.32	0.01
				X-ray	0.01		0.44		0.08		0.48	
Triose phosphate isomerase	1amk	tpi	C	SOMCD	0.62	0.06	0.13	0.03	0.06	0.01	0.19	0.02
				X-ray	0.57		0.14		0.05		0.24	
Trypsin	5ptp	try	C	SOMCD	0.14	0.02	0.34	0.03	0.13	0.00	0.38	0.02
				X-ray	0.11		0.32		0.14		0.42	
α-Helix reference spectrum 1[c]		Ayan	Y		1.00		0.00		0.00		0.00	
α-Helix reference spectrum 2[c]		Acur	B		1.00		0.00		0.00		0.00	
β-Sheet reference spectrum 1[c]		Byan	Y		0.00		1.00		0.00		0.00	
β-Sheet reference spectrum 2[c]		Bbrb	B		0.00		1.00		0.00		0.00	
Random reference spectrum 1[c]		Ryan	Y		0.00		0.00		0.00		1.00	
Random reference spectrum 2[c]		Rbrb	B		0.00		0.00		0.00		1.00	

[†]*H*, α-helix; *E*, β-strand; *T*, β-turn; *O*, other amides not classified as belonging to any of the previous categories (see text for details regarding secondary structure assignments). The estimated values, corresponding to lines where method = SOMCD, are the mean over 20 algorithm runs. The standard deviations (SD) are included to show the invariability in the results. The reference spectra are included for completion of the reference protein set (see text).

[a]The circular dichroism (CD) column lists the three- or four-letter abbreviations used for each CD spectrum.

[b]List of the sources for the protein CD spectra, where *C* = Manavalan and Johnson,[13] *P* = Pancoska et al.,[24] Y = Yang et al.,[23] and B = Brahms and Brahms.[26]

[c]Three reference spectra are taken from Brahms and Brahms.[26] The α-helix reference spectrum is the spectrum of sperm whale myoglobin, which has been normalized to 100% helical content. The spectrum of poly(Lys$^+$-Leu-Lys$^+$-Leu) in 0.5 M NaF pH 7 has been used as a model for the β-sheet reference spectrum, whereas the random reference spectrum was obtained from poly(Pro-Lys$^+$-Leu-Lys$^+$-Leu) in salt-free solution. The remaining three (source Y) are standard spectra for α-helix, β-sheet and random coil conformation extracted from 15 proteins by multilinear regression by Yang et al.[23] We chose to use the same reference spectra as in Andrade et al.[18]

been included, and the wavelength range has been expanded. CD spectra are used to train an unsupervised learning neural network, by which the spectra are mapped topologically onto a grid of nodes. Neighboring nodes have similar weight vectors, and correspond to similar structures, as is shown by the umat graph, which means that similar spectra are mapped to the same local region of the map. Continuity of the weights is essential for creating good structure maps.

Examination of the clusters has made it easier to verify the continuity between the weight vectors of neighboring neurons. If the cluster representation in Figure 1 does not display dark-colored regions, the distances between weights is small in all local regions of the map. This feature of the som_pak program package provides an excellent guidance in determining whether the network has achieved good self-organization or not. It is, in fact, as important to take into account as network distortion. Clustering also provides a first approach to secondary structure prediction: an estimation of the secondary structure percentages can be done on the basis of the zone of the map the spectrum falls, or the label of the neuron that is closest to it.

TABLE III. Comparison of Different Methods of Prediction of Secondary Structure Values Using Pearson Correlation Coefficients (r) and RMSD (δ)[†]

			H		*E*		*T*		*O*	
Method	Range	Proteins	δ	*r*	δ	*r*	δ	*r*	δ	*r*
Lin. reg.[a,b]	205–240	18	0.10	0.96	0.17	0.94	0.12	0.31	0.15	0.49
SVD[b]	190–260	16	0.04	0.98	0.20	−0.27	0.09	0.18	0.17	0.24
CONTIN[b]	190–240	18	0.05	0.96	0.06	0.94	0.10	0.31	0.11	0.49
VARSLC[b]	190–260	16	0.07	0.95	0.13	0.45	0.05	0.54	0.08	0.69
SELCON[c]	178–260	16	0.08	0.96	0.07	0.89	0.05	0.78	0.06	0.70
K2D[b]	200–240	24	0.11	0.91	0.14	0.73	—	—	0.13	0.76
CDSSTR[d]	178–234	22	0.06; 0.03; 0.03	0.99; 0.62; 0.76[e]	0.04	0.94	0.04	0.38	0.05	0.87
SOMCD	190–240	39	0.07	0.95	0.08	0.92	0.04	0.75	0.06	0.94

[†]*H*, α-helix; *E*, β-sheet; *T*, β-turn; *O*, other amides not classified as belonging to any of the previous categories.
[a]Linear regression.
[b]Data taken from Andrade et al.[18]
[c]Data taken from Sreerama and Woody.[31]
[d]Data taken from Johnson.[16]
[e]For this method, rms deviations and correlation coefficients are shown for α-helix, 3_{10}-helix, and poly(L-proline) II type 3_1-helix, respectively.

The root-mean-square deviation (RMSD) (δ) and Pearson correlation coefficient r were calculated using equations:

$$\delta = \sqrt{\frac{1}{n}\sum (x_i - y_i)^2} \tag{7}$$

and

$$r = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\left(\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right] \times \left[\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right]\right)^{1/2}} \tag{8}$$

where x_i and y_i are the experimental and calculated values, respectively, and n is the number of samples studied. r varies between −1 and 1, where an r of 1 indicates perfect correlation, −1 indicates anti-correlation, and 0 indicates no correlation at all.

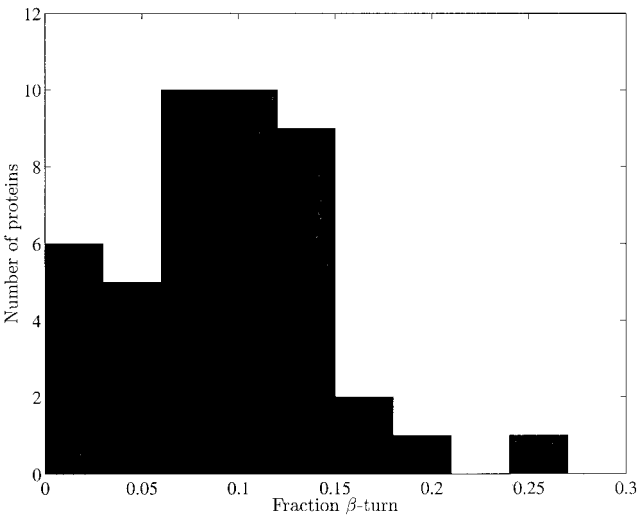

Fig. 4. Histogram plot of β-turn values.

Because of the self-organizing and cooperative feature of the map, anomalous spectra (i.e., spectra much affected by aromatic side chains etc.) are excluded or mapped to the borders of the map. One example of this is concanavalin A (Con A), labeled cna in Figure 1. It was pointed out by Andrade et al.[18] that this protein has many aromatic side chains, and that structure estimation was bad. Even if estimation in this work has improved, the spectrum of Con A is still mapped to the border of the map, as there are no similar spectra in the training set. Provided that the number of anomalous spectra is not excessive, the algorithm itself is able to filter out these bad examples and even take advantage of them, using them as anchors to the corners of the maps.

After optimal parameters have been found, it is only necessary to train the network once. Thus, the same structure maps can be used for structure estimation of a protein, while in the previous version, the average results over several maps were used. Mapping of a test spectrum is instantaneous; consequently, structure information is rapidly provided.

The overall performance of the method has improved as compared with K2D. This could be attributed to the increase in training set size. More training spectra means the algorithm has more information to interpolate, improving the continuity in the network, and giving more accurate structure maps. The wavelength range of the spectra used is also bigger, which allows for a better discrimination among the secondary structures. Moreover, an additional secondary structure, β-turn, is evaluated, which constrains the freedom of the remaining structure values.

Another thing that has improved performance is the fact that network size and parameters have been chosen based on examination of network distortion and the clusters, which was not done with K2D.

A web-based version of SOMCD has been prepared, where users can submit their CD spectra and receive instantaneous secondary structure information about the protein. (The server runs on a perl script, and can be found at URL:http://somcd.geneura.org.)

Future work on the algorithm will procede along the following lines. Systematic parameter setting, that is, a methodology designed to find the best values for the map size and learning parameters, will be in effect by utilizing global optimization algorithms such as simulated annealing or genetic algorithms. New results can be obtained immediately when new proteins can be included into the training set. In order to do that, a systematic "harvesting" effort will be done from the method web page.

REFERENCES

1. Kohonen T. The self-organizing map. Proc IEEE 1990;78:1464–1480.
2. Ripley BD. Pattern recognition and neural networks. Cambridge, UK: Cambridge University Press; 1996.
3. DeBoeck G, Kohonen T. Explorations in finance with self-organizing maps. London: Springer Finance; 1998.
4. Chang CT, Wu CS, Yang JT. Circular dichroic analysis of protein conformation: inclusion of the beta-turns. Anal Biochem 1978;91: 13–31.
5. Greenfield N, Fasman GD. Computed circular dichroism spectra for the evaluation of protein conformation. Biochemistry 1969;8: 4108–4116.
6. Chen YH, Yang JT. A new approach to the calculation of secondary structures of globular proteins by optical rotatory dispersion and circular dichroism. Biochem Biophys Res Commun 1971;44:1285–1291.
7. Saxena VP, Wetlaufer DB. A new basis for interpreting the circular dichroic spectra of proteins. Proc Natl Acad Sci USA 1971;68:969–972.
8. Provencher SW, Glockner J. Estimation of globular protein secondary structure from circular dichroism. Biochemistry 1981;20:33–37.
9. van Stokkum IH, Spoelder HJ, Bloemendal M, van Grondelle R, Groen FC. Estimation of protein secondary structure and error analysis from circular dichroism spectra. Anal Biochem 1990;191: 110–118.
10. Pancoska P, Janota V, Keiderling TA. Novel matrix descriptor for secondary structure segments in proteins: demonstration of predictability from circular dichroism spectra. Anal Biochem 1999;267: 72–83.
11. Sreerama N, Venyaminov SY, Woody RW. Estimation of the number of alpha-helical and beta-strand segments in proteins using circular dichroism spectroscopy. Protein Sci 1999;8:370–380.
12. Greenfield NJ. Methods to estimate the conformation of proteins and polypeptides from circular dichroism data. Anal Biochem 1996;235:1–10.
13. Manavalan P, Johnson WC Jr. Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra. Anal Biochem 1987;167:76–85.
14. Sreerama N, Woody RW. A self-consistent method for the analysis of protein secondary structure from circular dichroism. Anal Biochem 1993;209:32–44.
15. Perczel A, Hollosi M, Tusnady G, Fasman GD. Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. Protein Eng 1991;4:669–679.
16. Johnson WC. Analyzing protein circular dichroism spectra for accurate secondary structures. Proteins 1999;35:307–312.
17. Bohm G, Muhr R, Jaenicke R. Quantitative analysis of protein far UV circular dichroism spectra by neural networks. Protein Eng 1992;5:191–195.
18. Andrade MA, Chacon P, Merelo JJ, Moran F. Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. Protein Eng 1993;6:383–390.
19. Woody R. Ultraviolet/visible spectroscopy. Methods Enzymol 1995; 246:34–71.
20. Merelo JJ, Andrade MA, Prieto A, Morán F. Protein classification through a feature map. In: Fourth International Conference on Neural Networks and Their Applications (Neuro-Nimes, France, 1991). p 765–768.
21. Merelo JJ, Andrade MA, Prieto A, Morán F. Proteinotopic feature maps. Neurocomputing 1994;6:443–454.
22. Kohonen T, Hynninen J, Kangas J, Laaksonen J. SOM_PAK: The self-organizing map program package. Report A31. Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland. 1996.
23. Yang JT, Wu CS, Martinez HM. Calculation of protein conformation from circular dichroism. Methods Enzymol 1986;130:208–269.
24. Pancoska P, Bitto E, Janota V, Urbanova M, Gupta VP, Keiderling TA. Comparison of and limits of accuracy for statistical analyses of vibrational and electronic circular dichroism spectra in terms of correlations to and predictions of protein secondary structure. Protein Sci 1995;4:1384–1401.
25. Toumadje A, Alcorn SW, Johnson WC Jr. Extending CD spectra of proteins to 168 nm improves the analysis for secondary structures. Anal Biochem 1992;200:321–331.
26. Brahms S, Brahms J. Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. J Mol Biol 1980;138:149–178.
27. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
28. Hutchinson EG, Thornton JM. PROMOTIF—a program to identify and analyze structural motifs in proteins. Protein Sci 1996;5: 212-20.
29. Ultsch A, Siemon HP. Kohonen's self organizing feature maps for exploratory data analysis. In: Proceedings of the INNC 1990 International Neural Network Conference; 1990. Kluwer: Dordrecht, Netherland. p 305–308.
30. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
31. Sreerama N, Woody RW. Protein secondary structure from circular dichroism spectroscopy. Combining variable selection principle and cluster analysis with neural network, ridge regression and self-consistent methods. J Mol Biol 1994;242:497–507.