Neural, Parallel & Scientific Computations 15 (2007) 75-90

COMPUTATION OF CONSENSUS HYDROPHOBICITY SCALES WITH SELF-ORGANIZING MAPS AND FUZZY CLUSTERING ALONG WITH APPLICATIONS TO PROTEIN FOLD PREDICTION

NIKHIL RANJAN PAL, SOMITRA KUMAR SANADHYA, AND ANIMESH SHARMA

{nikhil, somitra_r}@isical.ac.in,
sharma.animesh@gmail.com
Indian Statistical Institute, 203 B.T. Road, Calcutta- 700108, India.

ABSTRACT. Hydrophobicity is probably the most important property of amino acid that is being often exploited to predict protein folds. There are many amino acid hydrophobicity scales available in the literature. Here we propose two computational approaches based on self-organizing map (SOM) and fuzzy clustering to find some consensus scales. Although SOM and fuzzy clustering produce centroids, we propose new schemes to compute more effective representative scales that exploit the properties of SOM and fuzzy memberships. To demonstrate the utility of the new scales, we apply them to predict the protein folds of a benchmark data set using neural networks. Our experiments show that it is possible to generate useful scales with better utility compared to some existing scales. There can be other applications of the proposed scales.

Key Words Hydrophobicity, Fuzzy C-means Clustering, Self Organizing Maps, Protein Fold prediction.

1. INTRODUCTION

Hydrophobicity of amino acids is an important feature for the still unsolved problem of protein structure prediction. However, the hydrophobicity is a difficult property to measure. According to Charton and Charton [2] "There is no special phenomenon denoted by hydrophobicity in amino acids. It is the natural and predictable result of differences in the intermolecular forces between water and the amino acid and those between the amino acid and some other medium." There are many hydrophobicity scales available in [4]. Some of these scales are radically different from others (having negative correlation), while some others are strongly related. Cornette et al.[4] describe 46 scales, out of which 12 are experimental, 22 are based on statistical studies, 9 that combine experimental and statistical data and 3 are averages of other scales. Some of these scales do not assign values to all amino acids. The scales analyzed by Cornette et al.[4] are first normalized but a uniform normalization

Received December 20, 2006

scheme has not been used for all scales because some scales were anomalous. We believe that it is not appropriate to take different centers for different scales. If a scale is anomalous then we should discard it, and the choice of the center should affect only the magnitude, not the order relationship of the amino acids in that scale. Neumaier et al.[14] analyzed 39 complete scales from [4] and the q-values scale by Li et al. [12] using Principal Component Analysis (PCA) [7]. Each amino acid is represented by an ordered tuple of 40 real numbers corresponding to the 40 scales. The principal component analysis is used and the top 3 principal components are used to construct a minimum spanning tree (MST). It is concluded from the MST that the nearest neighbor relation between amino acids is not fully linear. The closest approximation to their average scale is the scale by Li et al.[12] and the scale MIJER by Mijazawa and Jernigan [13].

Given such a diversity of hydrophobicity scales, we explore the possibility of finding consensus scales computationally. Our objective in this paper is to design computational methods to find useful scales with neural and fuzzy computing. In this regard, we have successfully used the self-organizing map and fuzzy clustering to derive new hydrophobicity scales that capture the salient properties of existing scales. The superiority of the new derived scales are demonstrated in protein fold prediction using a well known benchmark data set.

2. METHODOLOGY

2.1. **Data Set.** Among the 46 scales available in [4], we consider experimental and statistical scales in this study and do not consider average scales and scales that combine experimental and statistical data. We consider only complete scales because our methods are not suitable for incomplete scales. This leaves us with 28 of the 46 scales. In addition to these, we use another experimental scale of the so called "q-values scale" by Li et al.[12] since it has been used in a previous study, and thus finally we have 29 scales for analysis. The original 29 scales are shown in Table 1. In this Table, as well as in the rest of this paper, we use the same abbreviated names of the scales as used in [4].

2.2. **Normalization Schemes.** We use three different normalization schemes each of which preserves the relative ordering of hydrophobicity among different scales :

1. mean is made zero and standard deviation one;
2. hydrophobicity of Glycine is made zero and then the average is made equal to 1 (by dividing each scale by it's average);
3. hydrophobicity of Glycine is made zero and then the ranges equal (by dividing each scale with the difference of its maximum and minimum values).

Our objective is to compute representative scales. This can be done by finding homogeneous gropes of scales and replacing each group by a representative scale. To motivate that such approaches we first look at the correlation structure in the data.

2.3. **Correlation Analysis.** Both correlation and rank-correlation divide the 29 scales into 2 groups of 23 scales and 6 scales, such that one group has positive correlation with all the scales in that group and negative correlation with all the scales in the other group [Table 2]. This property of the scales is independent of the normalization scheme used for the scale as only the magnitude of the hydrophobicity values change by normalization. Note that the correlation values in Table 2 are rounded to single digit to accommodate the table into one page for display purposes only.

These two groups of scales are described in Table 3. From this grouping, we observe that the analysis in [14] tried to group highly dissimilar scales together. In the averaging scheme of theirs, MIJER and LI scales are found to be the closest to the first principal component. However, these two scales have negative correlation with each other [Table 2]. Thus we should be careful in averaging the scales, if they are not well correlated [Table 3].

Our analysis suggests that the set of scales has both positively and negatively (strongly) correlated subgroups. Hence, next we try to compute one or more representative scale(s) using neural networks and fuzzy clustering. Groups in a data set can be found using any clustering algorithm. We use here the fuzzy c-means [1] algorithm and the self-organizing maps [8] because the chances of finding poor prototypical scales due to bad initialization are less with both approaches.

2.4. **Fuzzy C-Means Based Scale Computation.** The Fuzzy C-Means (FCM) clustering algorithm [1] attempts to cluster data vectors into C groups based on the distances between them.

The FCM algorithm minimizes the objective function

$$J = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^{m} \|\mathbf{x}_j - \mathbf{v}_i\|^2,$$

where C is the number of clusters, $\mathbf{x}_j \in \Re^p$ is the i^{th} data vector, N is the number of data vectors, $m > 1$ is the fuzzifier, u_{ik} denotes the membership of k^{th} data vector to i^{th} cluster and $\mathbf{v}_i \in \Re^p$ is the centroid of the i^{th} cluster.

First order necessary conditions for $\mathbf{U}$ and $\mathbf{V}$ at a local minima of J are:

$$u_{ik} = \left[\sum_{j=1}^{C} \left(\frac{\|\mathbf{x}_i - \mathbf{v}_i\|}{\|\mathbf{x}_i - \mathbf{v}_j\|} \right)^{\frac{2}{m-1}} \right]^{-1}, \ \forall i, k \tag{2.1}$$

TABLE 1. Original Hydrophobicity Scales

No	Scale Name	Amino acids																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	S	T	W	Y	V	P
1	ZIMMR	0.83	0.83	0.09	0.64	1.48	0.00	0.65	0.10	1.10	2.52	3.07	1.60	1.40	2.75	2.70	0.14	0.54	0.31	2.97	1.79
2	JONES	0.87	0.85	0.09	0.66	1.52	0.00	0.67	0.10	0.87	3.15	2.17	1.64	1.67	2.87	2.77	0.07	0.07	3.77	2.67	1.87
3	FAUPL	0.31	-1.01	-0.60	-0.77	1.54	-0.22	-0.64	0.00	0.13	1.80	1.70	-0.99	1.23	1.79	0.72	-0.04	0.26	2.25	0.96	1.22
4	KUNTZ	1.50	3.00	2.00	6.00	1.00	2.00	7.50	1.00	4.00	1.00	1.00	4.50	1.00	0.00	3.00	2.00	2.00	2.00	3.00	1.00
5	ABODR	5.10	2.00	0.60	0.70	0.00	1.40	1.80	4.10	1.60	9.30	10.00	1.30	8.70	9.60	4.90	3.10	3.50	9.20	8.00	8.50
6	MEEK	0.50	0.80	0.80	-8.20	-6.80	-4.80	-16.90	0.00	-3.50	13.90	8.80	0.10	4.80	13.20	6.10	1.20	2.70	14.90	6.10	2.70
7	BULDG	-200	-120	80	-200	-450	160	-300	0	-120	2260	2460	-350	1470	2330	-980	-390	-520	2010	2240	1560
8	CHOTH	0.38	0.01	0.12	0.15	0.50	0.07	0.18	0.36	0.17	0.60	0.45	0.03	0.40	0.50	0.18	0.22	0.23	0.27	0.15	0.54
9	WERSC	0.52	0.49	0.42	0.37	0.83	0.35	0.38	0.41	0.70	0.79	0.77	0.31	0.76	0.87	0.35	0.49	0.38	0.86	0.64	0.72
10	JANIN	1.70	0.10	0.40	0.40	4.60	0.30	0.30	1.80	0.80	3.10	2.40	0.05	1.90	2.20	0.60	0.80	0.70	1.60	1.50	2.90
11	OLSEN	1.38	0.00	0.37	0.52	1.43	0.22	0.71	1.34	0.66	2.32	1.47	0.15	1.78	1.72	0.85	0.86	0.89	0.82	0.47	1.99
12	MEIRO	0.93	0.98	0.98	1.01	0.88	1.02	1.02	1.00	0.89	0.79	0.85	1.05	0.84	0.78	1.00	1.02	0.99	0.83	0.93	0.81
13	NNEIG	50.76	48.66	45.80	43.17	58.74	46.09	43.48	50.27	49.33	57.30	53.89	42.92	52.75	53.45	45.39	47.24	49.26	53.59	51.79	56.12
14	CHDLG	-0.27	2.00	0.61	0.50	-0.23	1.00	0.33	-0.22	0.37	-0.80	-0.44	1.17	-0.31	-0.55	0.36	0.17	0.18	0.05	0.48	-0.65
15	WSDLG	0.05	0.12	0.29	0.41	-0.84	0.46	0.38	0.31	-0.41	-0.69	-0.62	0.57	-0.38	-0.45	0.46	0.12	0.38	-0.98	-0.25	-0.46
16	JADLG	0.30	-1.40	-0.50	-0.60	0.90	-0.70	-0.70	0.30	-0.10	0.70	0.50	-1.80	0.40	0.50	-0.30	-0.10	-0.20	0.30	-0.40	0.60
17	GUY	0.10	1.91	0.48	0.78	-1.42	0.95	0.83	0.33	-0.50	-1.13	-1.18	1.40	-1.59	-2.12	0.73	0.52	0.07	-0.51	-0.21	-1.27
18	NIOII	0.23	-0.26	-0.94	-1.13	1.78	-0.57	-0.75	-0.07	0.11	1.19	1.03	-1.05	0.66	0.48	-0.76	-0.67	-0.36	0.90	0.59	1.24
19	MIJER	5.33	4.18	3.71	3.59	7.93	3.87	3.65	4.48	5.10	8.83	8.47	2.95	8.95	9.03	3.87	4.09	4.49	7.66	5.89	7.63
20	ROSEF	0.74	0.64	0.63	0.62	0.91	0.62	0.62	0.72	0.78	0.88	0.85	0.52	0.85	0.88	0.64	0.66	0.70	0.85	0.76	0.86
21	SWEET	-0.40	-0.59	-0.92	-1.31	0.17	-0.91	-1.22	-0.67	-0.64	1.25	1.22	-0.67	1.02	1.92	-0.49	-0.55	-0.28	0.50	1.67	0.91
22	SWEIG	-0.41	-0.58	-0.92	-1.31	0.16	-0.91	-1.22	-0.68	-0.63	1.24	1.22	-0.67	1.02	1.94	-0.50	-0.56	-0.29	0.51	1.70	0.90
23	PRIFT	-0.96	0.75	-1.94	-5.68	4.54	-5.30	-3.86	-1.28	-0.62	5.54	6.81	-5.62	4.76	5.06	-4.47	-1.92	-3.99	0.21	3.34	5.39
24	PRILS	-0.26	0.08	-0.46	-1.30	0.83	-0.83	-0.73	-0.40	-0.18	1.10	1.52	-1.01	1.09	1.09	-0.62	-0.55	-0.71	-0.13	0.69	1.15
25	ALFT	-0.73	-1.03	-5.29	-6.13	0.64	-0.96	-2.90	-2.67	3.03	5.04	4.91	-5.99	3.34	5.20	-4.32	-3.00	-1.91	0.51	2.87	3.98
26	ALTLS	-1.35	-3.89	-10.96	-11.88	4.37	-1.34	-4.56	-5.82	6.54	10.93	9.88	-11.92	7.47	11.35	-10.86	-6.21	-4.83	1.80	7.61	8.20
27	TOTFT	-0.56	-0.26	-2.87	-4.31	1.78	-2.31	-2.35	-1.35	0.81	3.83	4.09	-4.08	3.11	3.67	-3.22	-1.85	-1.97	-0.11	2.17	3.31
28	TOTLS	-1.37	-1.33	-6.29	-8.93	4.47	-3.88	-4.04	-3.39	1.65	7.92	8.68	-7.70	7.13	7.96	-6.25	-4.08	-4.02	-0.79	4.73	6.94
29	LI	-0.12	-0.02	-0.02	0.03	-0.27	0.01	0.04	-0.05	-0.11	-0.39	-0.44	0.07	-0.33	-0.44	-0.05	-0.01	-0.06	-0.30	-0.23	-0.31

TABLE 2. correlation coefficient of amino acid scales

No	Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	ZIMMR	1	.7	.6	.6	.5	.6	.4	.5	.4	.4	.4	.3	.5	.5	.5	.7	.7	.6	.7	.6	.6	.6	.6	-.2	-.5	-.4	-.4	-.5	-.7
2	JONES	.7	1	.8	.7	.7	.7	.4	.7	.4	.4	.5	.4	.6	.7	.6	.7	.7	.5	.6	.5	.5	.5	.6	-.3	-.7	-.4	-.7	-.5	-.7
3	FAUPL	.6	.8	1	.8	.7	.7	.8	.9	.8	.7	.8	.8	.9	.9	.9	.8	.8	.8	.8	.8	.8	.8	.8	-.6	-.9	-.8	-.9	-.9	-.9
4	KUNTZ	.6	.7	.8	1	.8	.9	.6	.7	.5	.7	.6	.6	.6	.8	.7	.9	.9	.7	.7	.8	.7	.8	.7	-.5	-.8	-.6	-.6	-.7	-.8
5	ABODR	.5	.7	.7	.8	1	.7	.4	.6	.3	.4	.5	.4	.5	.6	.5	.7	.7	.5	.5	.5	.5	.5	.5	-.7	-.6	-.4	-.5	-.5	-.7
6	MEEK	.6	.7	.7	.9	.7	1	.5	.8	.5	.5	.6	.5	.7	.8	.7	.9	.9	.8	.8	.8	.8	.8	.8	-.5	-.8	-.5	-.7	-.7	-.9
7	BULDG	.4	.4	.8	.6	.4	.5	1	.7	.9	1	.8	.9	.8	.8	.9	.7	.7	.8	.8	.7	.7	.8	.8	-.6	-.8	-.9	-.7	-.9	-.8
8	CHOTH	.5	.7	.9	.7	.6	.8	.7	1	.8	.7	.9	.7	.9	.9	.9	.8	.8	.9	.9	.9	.9	.9	.9	-.6	-.9	-.6	-1	-.9	-.9
9	WERSC	.4	.4	.8	.5	.3	.5	.9	.8	1	.8	.9	.9	.9	.8	.9	.7	.6	.8	.8	.7	.7	.8	.8	-.6	-.7	-.8	-.8	-.8	-.8
10	JANIN	.4	.4	.7	.7	.4	.5	1	.7	.8	1	.8	.9	.7	.8	.8	.6	.6	.7	.7	.7	.7	.7	.7	-.6	-.8	-.9	-.6	-.8	-.8
11	OLSEN	.4	.5	.8	.6	.5	.6	.8	.9	.9	.8	1	.8	1	.9	1	.8	.8	.9	.9	.8	.8	.9	.9	-.7	-.9	-.7	-.9	-.8	-.9
12	MEIRO	.3	.4	.8	.6	.4	.5	.9	.7	.9	.9	.8	1	.8	.8	.9	.6	.6	.7	.7	.7	.7	.7	.7	-.7	-.8	-.9	-.7	-.9	-.8
13	NNEIG	.5	.6	.9	.6	.5	.7	.8	.9	.9	.7	1	.8	1	.9	.9	.8	.8	.9	.9	.8	.9	.9	.9	-.6	-.9	-.6	-.9	-.8	-.9
14	CHDLG	.5	.7	.9	.8	.6	.8	.8	.9	.8	.8	.9	.8	.9	1	1	.9	.9	.9	.9	.9	.9	.9	.9	-.7	-1	-.7	-.9	-.9	-1
15	WSDLG	.5	.6	.9	.7	.5	.7	.9	.9	.9	.8	1	.9	.9	1	1	.8	.8	.9	.9	.9	.9	.9	.9	-.7	-.9	-.8	-.9	-.9	-.9
16	JADLG	.7	.7	.8	.9	.7	.9	.7	.8	.7	.6	.8	.6	.8	.9	.8	1	1	.9	.9	.9	.8	.9	.9	-.6	-.8	-.6	-.7	-.8	-.9
17	GUY	.7	.7	.8	.9	.7	.9	.7	.8	.6	.6	.8	.6	.8	.9	.8	1	1	.9	.9	.9	.8	.9	.9	-.6	-.8	-.6	-.7	-.8	-.9
18	NIOII	.6	.5	.8	.7	.5	.8	.8	.9	.8	.7	.9	.7	.9	.9	.9	.9	.9	1	1	.9	.9	1	1	-.6	-.9	-.6	-.8	-.8	-.9
19	MIJER	.7	.6	.8	.7	.5	.8	.8	.9	.8	.7	.9	.7	.9	.9	.9	.9	.9	1	1	.9	.9	1	1	-.6	-.9	-.6	-.8	-.8	-.9
20	ROSEF	.6	.5	.8	.8	.5	.8	.7	.9	.7	.7	.8	.7	.8	.9	.9	.9	.9	.9	.9	1	1	1	1	-.6	-.9	-.6	-.8	-.8	-.9
21	SWEET	.6	.5	.8	.7	.5	.8	.7	.9	.7	.7	.8	.7	.9	.9	.9	.8	.8	.9	.9	1	1	1	1	-.5	-.9	-.6	-.8	-.8	-.9
22	SWEIG	.6	.5	.8	.8	.5	.8	.8	.9	.8	.7	.9	.7	.9	.9	.9	.9	.9	1	1	1	1	1	1	-.6	-.9	-.6	-.8	-.8	-.9
23	PRIFT	.6	.6	.8	.7	.5	.8	.8	.9	.8	.7	.9	.7	.9	.9	.9	.9	.9	1	1	1	1	1	1	-.6	-.9	-.6	-.8	-.9	-.9
24	PRILS	-.2	-.3	-.6	-.5	-.7	-.5	-.6	-.6	-.6	-.6	-.7	-.7	-.6	-.7	-.7	-.6	-.6	-.6	-.6	-.6	-.5	-.6	-.6	1	.6	.5	.5	.6	.7
25	ALFT	-.5	-.7	-.9	-.8	-.6	-.8	-.8	-.9	-.7	-.8	-.9	-.8	-.9	-1	-.9	-.8	-.8	-.9	-.9	-.9	-.9	-.9	-.9	.6	1	.7	.9	.9	1
26	ALTLS	-.4	-.4	-.8	-.6	-.4	-.5	-.9	-.6	-.8	-.9	-.7	-.9	-.6	-.7	-.8	-.6	-.6	-.6	-.6	-.6	-.6	-.6	-.6	.5	.7	1	.6	.8	.7
27	TOTFT	-.4	-.7	-.9	-.6	-.5	-.7	-.7	-1	-.8	-.6	-.9	-.7	-.9	-.9	-.9	-.7	-.7	-.8	-.8	-.8	-.8	-.8	-.8	.5	.9	.6	1	.8	.9
28	TOTLS	-.5	-.5	-.9	-.7	-.5	-.7	-.9	-.9	-.8	-.8	-.8	-.9	-.8	-.9	-.9	-.8	-.8	-.8	-.8	-.8	-.8	-.8	-.9	.6	.9	.8	.8	1	.9
29	LI	-.6	-.7	-.9	-.8	-.7	-.9	-.8	-.9	-.8	-.8	-.9	-.8	-.9	-1	-.9	-.9	-.9	-.9	-.9	-.9	-.9	-.9	-.9	.7	1	.7	.9	.9	1

and

$$\mathbf{v}_i = \frac{\sum_{k=1}^{N} u_{ik}^m \mathbf{x}_k}{\sum_{k=1}^{N} u_{ik}^m}, \ \forall i. \tag{2.2}$$

The algorithm iterates between equations 2.2 and 2.1 in that order, as described below.

Algorithm:Fuzzy C-Means

1. Initialize a valid fuzzy c-partition U.
2. Compute a new set of prototypes using eq. (2.2).
3. Compute a new partition matrix using eq. (2.1) with these new prototypes.
4. Repeat this process (Steps 2 and 3 alternately) till the entries of the partition matrix stabilize.
5. Defuzzification : Assign the data vector $\mathbf{x}_k$ to the cluster j for which it's membership value u_{jk} is largest.

The same procedure can be carried out by initializing the prototypes instead of the partition matrix in which case the algorithm iterates between equations 2.1 and 2.2 in that order. As the value of m increases the algorithm produces more fuzzy partitions [1].

We clustered the hydrophobicity scales with various parameters in the above algorithm and obtained consistent results for C =3 clusters. Experiments with more than 3 clusters resulted in clusters, 3 of which contained most of the scales and the rest were almost empty or completely empty. Therefore we settled for C =3 clusters. The fuzzification parameter m was chosen to be 2 as suggested in [1]. We call a cluster strong if the membership values of all data points in that cluster are high. Among the 3 clusters, the strongest cluster is the one which contains the 6 scales which are grouped together in Table 3. The membership values of the scales in this cluster are : 0.41 (KUNTZ), 0.93 (MEIRO),0.49 (CHDLG), 0.90 (WSDLG), 0.93 (GUY), and (0.96) LI.

We call the FCM computed centroid as 'the FCM Cluster Center scale'. However, the scales which are not members of a particular cluster also affect to some extent the cluster center computation in FCM. Therefore, we create a new scale by taking a weighted average of the 6 scales grouped together. This new scale is computed by Equation (2.3). This averaging is unaffected by scales which are not part of this group. We call this scale 'the scale obtained by FCM averaging'.

$$NewScale = \frac{\sum_{k=1}^{6}(\ membership\ of\ scale\ k\ in\ cluster\ ) * (\ scale\ k)}{\sum_{k=1}^{6}(\ membership\ of\ scale\ k\ in\ cluster\ )} \tag{2.3}$$

Both these scales are shown in Table 5.

TABLE 3. Groups obtained in scales by Correlation Coefficient Analysis

Group 1	Group 2
ZIMMR, JONES, FAUPL, ABODR, MEEK, BULDG, CHOTH, WERSC, JANIN, OLSEN, NNEIG, JADLG, NIOII, MIJER, ROSEF, SWEET, SWEIG, PRIFT, PRILS, ALFT, ALTLS, TOTFT, TOTLS	KUNTZ, MEIRO, CHDLG, WSDLG, GUY, LI

2.5. **Self Organizing Map Based Scale Computation.** Kohonen's self-organizing map (SOM) [8] has the interesting property of achieving a distribution of the weight vectors that approximates the distribution of the input data. This property of the SOM can be exploited for clustering. The visual display produced by a SOM helps to form hypotheses about topological structure present in the data [11]. It is a two layered network with complete connection between the layers. The number of nodes in the input layer is equal to the dimension of the data. The output layer has a set of neurons typically having a two dimensional lattice structure. Let the input dimension be p, then each node in the output layer has an associated weight vector in $\Re^p$. For example, the i^{th} node in the output layer has the weight vector $\mathbf{w}_i \in \Re^p$ representing the connection weights of this node with the p nodes in the layer.

Algorithm:SOM

Input:

- a set of vectors $\mathbf{x}_i \ : \ i = 1, 2, \ldots n$ =number of data vectors,
- learning rate coefficient $\alpha(t)$ which decreases with iteration t,
- a lateral feedback function $g_t(r, i)$ which decreases with iteration t and distance of the node i from winner node r,
- neighbourhood function $N_c(t)$ which denotes the set of nodes which is the neighbourhood of node c at iteration t, and
- the grid size $m \times n$.

1. Randomly initialize each weight vector $\mathbf{w}_i(0)$ for $i \in \{1, 2, ..., m \times n\}$.
2. For each $\mathbf{x}_i$, in iteration no. t:
   - (a) Identify the winner node (weight vector) c , that is closest to $\mathbf{x}_i$ according to euclidean distance (i.e. $\|\mathbf{w}_c(t) - \mathbf{x}_i\|^2$ is smallest).
   - (b) update the weights of all nodes in the neighbourhood of the winner node, as
     $\mathbf{w}_k(t+1) = \mathbf{w}_k(t) + \alpha(t).g_t(c,k)(\mathbf{x}_i - \mathbf{w}_k(t))$ for each $\mathbf{w}_k(t) \in N_c(t)$, the neighborhood of winner node.
     Usually $g_t(c,k) = \exp(\frac{\|c-k\|^2}{\sigma(t)^2})$ is taken where $\sigma(t)$ decreases with iteration no. t.

3. Repeat 2 with reducing learning rate α, reducing lateral function g and reducing neighbourhood size, until all the weight vectors $\mathbf{w}_i$ stabilize.

In this paper, use a Map of size 5 by 5. In 15000 iterations the learning rate is reduced from 0.50 to 0.05 and the square neighborhood size is reduced from 3 to 0. Then the training is continued for 5000 more iterations with learning rate 0.05. This reveals the structure in data. The trained SOM of a typical run is shown in Table 4.

TABLE 4. Self Organizing Map

3	.	3	1	3
1	.	.	.	1
2	.	2	.	3
.	.	1	.	2
1	1	2	1	2

The numbers in this map represent the number of different scales that have converged at a particular node. The top left corner has scales 12, 15 and 29; and just below it are the scales 17; 4 and 14. It can be seen from the figure that other scales are quite distributed and no strong grouping tendency is observed. This is consistent with the correlation analysis of the data. The above structure is obtained for the third normalization of the original data. Similar structures are obtained for the other two normalizations as well. The same experiment was repeated 4 times with different learning rate and neighborhood update parameters, and each time a similar structure was observed.

In the above map we have 3 weight vectors, one at each node in the top left corner of the map, which represent the 6 scales. Each of these 3 vectors can be assigned importance equal to the number of scales that are represented by it. For example, the scale which represents the top left corner node in SOM in Table 4 is assigned importance 3. A weighted average of the 3 weight vectors obtained by SOM yields a new hydrophobicity scale. The average scale is calculated according to Equation (2.4). This scale is also included in Table 5.

(2.4)

$$NewScale = \frac{\sum_k (\ no.\ of\ scales\ converging\ at\ node\ k) * (\ weight\ vector\ of\ node\ k\ )}{\sum_k (\ no.\ of\ scales\ converging\ at\ node\ k\ )}$$

It is interesting to note that although the three scales in table 5 were computed by two widely different methods, the correlation coefficient and the rank correlation coefficient between the SOM scale and the FCM averaging scale is 0.9963 and 0.9868 respectively.

TABLE 5. Proposed Hydrophobicity Scales

Amino acid	By SOM Analysis	By FCM Averaging	By FCM Cluster Center
A	-0.11	-0.12	-0.14
R	0.20	0.16	0.11
N	0.06	0.04	0.02
D	0.19	0.16	0.12
C	-0.37	-0.41	-0.44
Q	0.17	0.15	0.13
E	0.22	0.18	0.14
G	0.00	0.00	0.00
H	-0.15	-0.18	-0.23
I	-0.45	-0.52	-0.54
L	-0.37	-0.47	-0.51
K	0.29	0.26	0.23
M	-0.34	-0.42	-0.45
F	-0.46	-0.56	-0.60
S	0.12	0.08	0.07
T	0.04	0.04	0.03
W	0.04	0.01	0.00
Y	-0.37	-0.41	-0.45
V	-0.09	-0.17	-0.21
P	-0.39	-0.44	-0.47

3. EVALUATION OF THE NEW SCALES

First we assess the quality of the new scales in terms of the grouping of amino acids that they impose. Then we evaluate the scales in prediction of protein folds.

3.1. **Assessment of the Hydrophobicity Groups.** We consider five scales here. Three of these are proposed in this paper, the fourth one is the Kyte-Dolittle scale [10] (it is widely used by the bioinformatics community) and the last one is used by Dubchuk et al.[5]. The last scale was proposed by Chothia and Finkelstein [3] and we denote it by CHOFIN in this paper.

Table 5 clearly suggests that the amino acids can be divided into three natural groups. For example, the SOM scale gives the following groups: Polar (hydrophobicity values ranging from 0.12 to 0.29), Neutral (values from -0.15 to 0.06) and Hydrophobic (values from -0.46 to -0.34). The scales relating to the FCM Cluster Center, FCM Averaging and the Kyte-Dolittle scale can similarly be grouped. The grouping for CHOFIN scale is as used in [5]. The groups for all the five scales are shown in Table

6. It is evident from the table that the two new scales obtained by FCM analysis divide the amino acids into identical groups. These groups are almost identical to the grouping obtained by the SOM scale, except the position of amino acid Serine(S). Serine is on the boundary of the groups "Polar" and "Neutral". In the SOM scale, Serine falls in polar category whereas in the FCM scale it is categorized as neutral.

TABLE 6. Hydrophobicity based grouping of amino acids

	Hydrophobic	Neutral	Polar
SOM	C,F,I,L,M,P,Y	A,G,H,N,T,V,W	D,E,K,Q,R,S
FCM Average	C,F,I,L,M,P,Y	A,G,H,N,S,T,V,W	D,E,K,Q,R
FCM Cluster Center	C,F,I,L,M,P,Y	A,G,H,N,S,T,V,W	D,E,K,Q,R
Kyte-Dolittle	A,C,F,I,L,M,V	G,P,S,T,W,Y	D,E,H,K,N,Q,R
CHOFIN	C,F,I,L,M,V,W	A,G,H,P,S,T,Y	D,E,K,N,Q,R

3.2. **Dubchuk et al.'s feature set for protein fold prediction.** One of the most successful set of features used for prediction of protein folds consists of global description of the chain of amino acids representing proteins [5]. This set of features contains 125 features which have been used extensively in the context of SCOP classification [6]. The data sets with these features are available at `http://www.nersc.gov/~cding/protein`.

In the present study we also use this set of 125 features. Of the 125 features, 21 features are computed based on assigning hydrophobicity values in 3 groups: Polar, Neutral and Hydrophobic. The grouping is shown in Tabel 6.

We use the following four datasets :

1. 125 features from [5].
2. 104 features from [5] and 21 features corresponding to hydrophobicity generated by using the proposed SOM hydrophobicity scale.
3. 104 features from [5] and 21 features corresponding to hydrophobicity generated by using the proposed FCM hydrophobicity scales. (We need to generate only one dataset for the two FCM scales since both these scales group the amino acids in identical groups.)
4. 104 features from [5] and 21 features corresponding to hydrophobicity generated by the widely used Kyte-Dolittle scale [10].

3.3. **Extended feature set for protein fold prediction.** The dataset used in the previous section depends only on the hydrophobicity groups generated by the scale used, and not on the actual magnitudes of the hydrophobicity assigned. A new set of features consisting of 447 features was proposed in [15] which has been shown to be superior to the SCOP dataset for protein fold prediction [15]. We generated

447 features, as described in [15], for the CHOFIN scale [3], Kyte-Dolittle Scale and the three scales proposed by us. The three scales proposed in this paper are linear combinations of 6 normalized scales as already described in previous sections. The most dominant scale in our grouping is that scale which contributes the highest weight in the linear combination. Similarly, the least dominant scale is the one which has the least weight in the linear combiner. To evaluate whether the combination of scales used by us has been effective in improving the protein fold prediction accuracy, we also generated 447 features for the most and least dominant scales in our grouping of scales, namely the scale due to Li et al. [12] and the scale due to Kuntz [9]. Multi layered perceptron network is used to evaluate the protein fold prediction accuracy.

TABLE 7. Grouping of amino acids based on attributes (an extended version of Table 1 in [5])

Property	Group 1	Group 2	Group 3	No of features
Hydrophobiicty (H)	Polar R,K,E,D,Q,N	Neutral G,A,S,T,P,H,Y	Hydrophobic C,V,L,I,M,F,W	21
Volume (V)	0 - 2.78	2.95 - 4.0	4.43 - 8.08	21
Polarity (P)	4.9 - 6.2	8.0 - 9.2	10.4 - 13.0	21
Polarizability (Z)	0.0 - 0.108	0.128 - 0.186	0.219 - 0.409	21
Predicted Secondary Structure (S)	Helix	Strand	Coil	21
Composition (C)	-	-	-	20

3.4. **Results.**

3.4.1. *Dubchuk et al.'s features.* We use the same data set used by Dubchuk et al. [5]. The dataset consists of 313 training and 385 test protein sequences. Each sequences is converted to features as described in [5]. Since the only group that we changed is hydrophobicity, we only need to compute the 21 features related to hydrophobicity; rest of the features are the same as used in [5]. From the features obtained, we have experimented on various combinations of features. Results are reported on hydrophobicity alone (H), composition and hydrophobocity (CH) and all six types of features namely composition, predicted secondary structure, hydrophobiicty, polarity, volume and polarizability (CSHPVZ). We have used a multi layered perceptron (MLP) network consisting of 3 layers: (a:b:c), where a,b and c are the number of nodes in the input, hidden and output layer respectively. We use the architectures (21:70:27); (41:75,27) and (125,9,27) for H, CH and CSHPVZ respectively. The number of hidden layer neurons is increased with the increase in the dimension of input vector.

The network is trained by backpropagation with no momentum and learning rate 0.3. Each experiment was repeated 10 times with different initializations of weights and the best results are reported in Tables 8 and 9.

It is evident from the tables that the proposed new scales have good discriminatory power with respect to fold prediction. Table 8 shows that we are able to achieve a 4.2% increase in classification accuracy over the CHOFIN scale by using the new FCM scales with features genearted from hydrophobicity alone. All the three new scales consistently give better or equal accuracy than CHOFIN and Kyte-Dolittle scales even with more features. When all the 6 types of features (namely CSHPVZ) are used, the new scales are better than CHOFIN scale and equal to the Kyte-Dolittle scale in accuracy. Among themselves, the FCM scales are better than the SOM scale. Note that, at this stage, we cannot distinguish between the two FCM scales as both produce the same grouping of amino acids.

From Table 9, we can see that the FCM scales are better than or equal to the Kyte-Dolittle and CHOFIN scales in 18 and 21 out of 27 classes when using the features derived only from hydrophobicity (H). The FCM scales are better than or equal to Kyte-Dolittle scale in 18 and 16 out of 27 classes for CH and CSHPVZ; and are better than or equal to CHOFIN scale in 21 and 23 out of 27 classes in the case of CH and CSHPVZ.

The SOM scale is better than or equal to Kyte-Dolittle in 17,16 and 27 classes out of 27; and better than or equal to CHOFIN in 18,18 and 23 classes out of 27 when using feature sets H, CH and CSHPVZ respectively.

TABLE 8. Protein fold classification accuracy for 27 classes

Feature Sets	Network Size	FCM	SOM	CHOFIN	Kyte-Dolittle
H	21:70:27	29.1	28.3	24.9	28.3
CH	41:75:27	44.2	43.9	43.9	43.6
CSHPVZ	125:95:27	50.9	50.9	47.0	50.9

3.4.2. *Extended features.* Using the same training - test partition as used in the previous section, we have generated an extended feature set containing 447 features for the CHOFIN scale [3], the Kyte-Dolittle scale, the three proposed scales, the scale by Kuntz [9] and the q-values scale of Li et al. [12]. Here also MLP network with one hidden layer is used. The number of neurons in the hidden layer is kept fixed at 135 i.e., we used (447:135:27) architecture. Fixed learning rate of 0.3 is used without any momentum term. The results for the extended feature set are summarised in Table 10. In this case, the SOM scale and the two FCM scales exhibit a marginally improved performance over the other scales. The scale obtained by FCM averaging is found to produce the most accurate results among all the scales.

TABLE 9. Class-wise protein fold classification accuracy

Folding Class	No of Samples	H (%)				CH (%)				CSHPVZ (%)			
		FCM	SOM	CHO	KD	FCM	SOM	CHO	KD	FCM	SOM	CHO	KD
1	6	83.3	83.3	50.0	66.7	83.3	83.3	66.7	83.3	83.3	83.3	83.3	83.3
3	9	55.6	66.7	44.4	22.2	66.7	66.7	44.4	77.8	88.9	88.9	77.8	88.9
4	20	30.0	30.0	20.0	25.0	40.0	35.0	35.0	45.0	55.0	40.0	50.0	40.0
7	8	12.5	37.5	0.0	0.0	62.5	50.0	37.5	25.0	62.5	50.0	62.5	50.0
9	9	55.6	55.6	55.6	55.6	88.9	88.9	77.8	77.8	100.0	100.0	88.9	100.0
11	9	11.1	22.2	11.1	22.2	33.3	44.4	44.4	33.3	22.2	44.4	33.3	44.4
20	44	20.5	36.4	27.3	36.4	47.7	45.5	50.0	50.0	50.0	56.8	47.7	56.8
23	12	25.0	16.7	25.0	8.3	50.0	41.7	50.0	33.3	58.3	41.7	41.7	41.7
26	14	64.3	21.4	35.7	50.0	57.1	57.1	57.1	42.9	57.1	71.4	64.3	71.4
30	6	0.0	0.0	0.0	16.7	33.3	16.7	16.7	33.3	50.0	33.3	33.3	33.3
31	8	37.5	0.0	50.0	37.5	62.5	62.5	50.0	37.5	50.0	50.0	37.5	50.0
32	19	31.6	10.5	15.8	15.8	21.1	21.1	21.1	5.3	15.8	21.1	15.8	21.1
33	4	0.0	0.0	25.0	50.0	50.0	25.0	50.0	50.0	50.0	50.0	50.0	50.0
35	4	0.0	0.0	0.0	0.0	25.0	25.0	25.0	0.0	25.0	25.0	25.0	25.0
39	7	0.0	14.3	57.1	14.3	42.9	42.9	42.9	57.1	28.6	42.9	42.9	42.9
46	48	25.0	29.2	35.4	37.5	31.3	45.8	54.2	56.3	58.3	43.7	52.1	43.7
47	12	25.0	16.7	8.3	25.0	66.7	66.7	25.0	33.3	41.7	50.0	41.7	50.0
48	13	23.1	7.7	7.7	23.1	23.1	15.4	23.1	15.4	46.2	53.8	38.5	53.8
51	27	18.5	25.9	18.5	33.3	33.3	33.3	40.7	37.0	40.7	40.7	22.2	40.7
54	12	25.0	33.3	25.0	16.7	58.3	58.3	41.7	33.3	41.7	33.3	41.7	33.3
57	8	25.0	37.5	0.0	12.5	25.0	37.5	12.5	25.0	37.5	50.0	37.5	50.0
59	14	14.3	14.3	21.4	28.6	28.6	21.4	28.6	35.7	42.9	50.0	50.0	50.0
62	7	42.9	28.6	28.6	28.6	57.1	57.1	42.9	57.1	57.1	57.1	42.9	57.1
69	4	25.0	25.0	25.0	0.0	0.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
72	8	12.5	25.0	0.0	0.0	0.0	25.0	0.0	0.0	12.5	25.0	12.5	25.0
87	27	22.2	11.1	18.5	7.4	25.9	22.2	33.3	33.3	29.6	33.3	29.6	33.3
110	26	69.2	65.4	34.6	50.0	92.3	76.9	88.5	88.5	100.0	100.0	96.2	100.0
Total	385	29.1	28.3	24.9	28.3	44.2	43.9	43.9	43.6	50.9	50.9	47.0	50.9

TABLE 10. Results for extended feature set

Scale	Classification Accuracy %
CHOFIN	50.1
Kyte-Dolittle	48.8
LI	50.6
KUNTZ	49.6
SOM	50.9
FCM Cluster Center	50.6
FCM Averaging	51.2

4. CONCLUSIONS

A large number of scales for amino acid hydrophobicity are available in literature. We have chosen 28 scales from [4] and the scale due to Li et al. [12] in this study. We first use correlation analysis to reveal that in the set of these 29 hydrophobicity

scales, there are two distinct groups of scales, such that the members in each group have positive correlation with all scales in that group and negative correlation with all scales in the other group. Then we have used self organizing feature map and fuzzy clustering algorithm to generate three new hydrophobicity scales. We also proposed new schemes to compute more useful representative scale exploiting properties of SOM and fuzzy memberships.

To demonstrate the superiority and effectiveness of the proposed scales, we have used them in protein fold prediction in the context of SCOP database. The three scales proposed in this study are found to be at least as good as the widely used Kyte-Dolittle scale and better than the CHOFIN scale [3] for protein fold prediction. The new scales can also be used in other bioinformatics problems where hydrophobicity of amino acids has a role to play.

REFERENCES

[1] James C. Bezdek, James M. Keller, Raghu Krishnapuram, and Nikhil R. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.* Boston, 1999.

[2] M Charton and BI Charton. The structural dependence of amino acid hydrophobicity parameters. *Journal of Theoritical Biology*, 99:629–644, 1982.

[3] C Chothia and AV Finkelstein. The classification and origin of folding patterns. *Anual Rev. Biochem.*, 59:1007–1039, 1990.

[4] JL Cornette, KB Cease, H Margalit, JL Spouge, JA Berzofsky, and C DeLisi. Hydrophobiicty scales and computational techniques for detecting amphiphatic structures in proteins. *Journal of Molecular Biology*, 195:659–685, 1987.

[5] I Dubchuk, I Muchnik, SR Holbrook, and SH Kim. Prediction of protein folding class using global description of amino acid sequence. *Proceedings of National Academy of Sciences, USA*, 92:8700–8704, 1995.

[6] I Dubchuk, I Muchnik, C Mayor, I Dralyuk, and SH Kim. Recognition of a protein fold in the context of the scop classification. *PROTEINS: Structure, Function and Genetics*, 35:401–407, 1999.

[7] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis.* Prentice Hall, 5th edition, 2002.

[8] Teuvo Kohonen. *Self-Organizing Maps*, volume 30. Springer Series in Information Sciences, 2nd edition, 1997.

[9] ID Kuntz. Hydration of macromolecules. *J. Amer. Chem. Soc.*, 93(2):514–518, 1971.

[10] J Kyte and RF Dolittle. A simple method for displaying the hydropathic character of proteins. *Journal of Molecular Biology*, 157:105–132, 1982.

[11] Arijit Laha and Nikhil Ranjan Pal. Dynamic generation of prototypes with self-organizing feature maps for classifier design. *Pattern Recognition*, 34:315–321, 2001.

[12] Hao Li, Chao Tang, and Ned Wingreen. Nature of driving force for protein folding-a result from analyzing the statistical potential. *Phys. Rev. Lett.*, 79:765, 1997.

[13] S Miyazawa and RL Jernigan. Residue-residue potentials with a favo-rable contact pair term and an unfavorable high packing den-sity term, for simulation and threading. *Journal of Molecular Biology*, 256:623–644, 1996.

[14] A Neumaier, W Huyer, and E Bornberg-Bauer. Hydrophobicity analysis of amino acids. world wide web, `http://www.mat.univie.ac.at/~neum/software/protein/aminoacids.html`.

[15] Nikhil Ranjan Pal and Debrup Chakraborty. Some new features for protein fold prediction. In *ICANN*, pages 1176–1183. Springer-Verlag, Heidelberg, 2003.