INDEXING AND MAPPING OF PROTEINS USING A MODIFIED NONLINEAR SAMMON PROJECTION

Nonlinear Sammon Projection of Compositional Space of Proteins can Predict Protein Folding Classes.

Izydor Apostol[1*] and Wojciech Szpankowski[2]

[1]Somatogen Inc., Boulder, CO 80301,

[2]Purdue University, Department of Computer Science, West Lafayette, IN 47907

* To whom correspondence should be addressed: Somatogen, Inc., 2545 Central Ave. Boulder, CO 80301

Abstract

A modified Sammon's algorithm was applied to display a relationship between proteins based on their amino acid composition. In the first stage of the method the 19-dimensional compositional space of representative proteins was mapped into 2-dimensinal space using the original Sammon projection to create a contour map. In the second stage, the contour map was used as a reference for newly projected proteins. Data analysis showed that proteins belonging to the same structural class form characteristic and distinct clusters which can be utilized in prediction of structural classes. However, significant overlapping of the clusters has been observed which may explain the limited success of previous protein folding predictions based solely on amino acid composition. Additionally, the modified Sammon's projections can generate a unique index for each individually projected protein related to its amino acid composition which can be a useful parameter in classification of proteins.

Introduction

Classification of proteins and prediction of their structural classes is an important task in the characterization of newly discovered proteins. Unfortunately, classification of proteins is limited only to proteins with enzymatic activities, and no general indexing has been widely accepted. Also, prediction of protein folding class based on primary sequence information remains a difficult task (Fasman, 1989; Mayoraz et al. 1995; Muskal and Kim, 1992; Lesk and Boswell, 1992; Tuckwell et al. 1995; Garnier et al. 1996; , 1989).

The comparison of more than two protein sequences is generally not a straightforward process. Aligning them, measuring similarity/dissimilarity distances requires complicated computing in multidimensional spaces equal to the length of the protein sequences. Therefore, the comparison of proteins of different length creates an additional difficulty of brining them to the same dimensional space and necessitates the introduction of complicated gaps. A simpler way to compare proteins is to contrast their amino acid composition (AAC). In this case all proteins can be compared in the same 19-D compositional space, based on the 20 amino acids used to create functional proteins. But, even in this reduced space, comparison of vectors of proteins is not easy primarily due to the limitation of humans to adequately visualize objects in spaces with greater than 3 dimensions. Several attempts have been made to use AAC information to predict protein folding classes (Hatch, 1965; Harding, 1984; Chou, 1989; Zhang and Chou, 1992; Dubchak et al. 1993; Nakai et al. 1988; Nakashima et al. 1986; Chun-Ting et al. 1992). In these cases the compositional information has been brought into a linear format. In this linear fitting process, different kinds of weighting factors and averaging have been introduced. These procedures resulted in a significant reduction of information and have shown limited predictive utility (Nakashima et al. 1986; Chun-Ting et al. 1992; Landes and Risler, 1994).

The Sammon nonlinear algorithm offers the possibility to project multidimensional spaces into 2 or 3 dimensional spaces while approximately preserving the original information (Sammon, 1969; Agrafiotis, 1997). The algorithm works by projecting protein vectors from compositional space onto a plane display in such a way that the Euclidean distances between the projected images (points) approximates, as closely as possible, the corresponding Euclidean distance in the original compositional space. No introduction of averaging and/or correction/weighting factors is necessary. The ability of this algorithm to capture the essential features of protein sequence similarities was recently demonstrated by Agrafiotis for a set of protein kinases (Agrafiotis, 1997). In this work we are attempting to use Sammon mapping to project compositional space of proteins into 2-D space to observe the relationship between them in that newly created space.

RESULTS

Helix example of Sammon’s nonlinear projection

We start here with a non biological example to show how Sammon’s algorithm preserves certain dependencies/shapes (e.g. helixes) when projected from 3-D to 2D space. Figure 1 displays the results obtained using the nonlinear Sammon algorithm to project 50 points distributed evenly along a 3-dimensional helix. The parametric equations for this helix are: $X = cos\ Z$, $Y = sin\ Z$, $Z = t / 2^{1/2}$. The points were distributed at one-unit intervals along the curve. To initiate the algorithm we selected corresponding 50 random points in 2-D space (Fig. 1A). Each point was assigned to represent one of the 50 points on the helix. Application of the Sammon algorithm caused 2-D points to organize in such a fashion that the Euclidean distances between points in 2-D space closely reflect their Euclidean distances on the helix in 3-D space. After 250 iterations for each point using the steepest descent algorithm (MF =0.3) described in the methods section, the projected points formed a highly organized wave shape (Fig. 1B) as described by Sammon (Sammon, 1969). Figure 1C displays the results of an experiment in which one random point was excluded from the helix. The remaining, 49 points were projected creating a gap in the projected wave shape. In the third experiment, the missing point was added back and all 50 point were projected but this time X, Y coordinates of 49 points were preserved as on figure 1C. After only 50 iterations, the missing point fell back in the gap and completed the wave shape. The orientation of the wave in panel C and D of Fig. 1 are different from panel B because the optimization started from different randomly distributed points in 2-D space and resulted in a different approximated projection.

These experiments demonstrate that a set of vectors in 3-dimensional space can be projected into 2 dimensional space and recreated the dependencies from a higher dimension even if one element of the set is missing. The missing element can be added back to the pre-computed set and complete the structure without the need to re-optimize the entire set from the beginning. This approach represents a significant advantage over traditional approaches because a full re-optimization "costs" n^2 versus n computation for one point optimization into the constant contour map.

The Sammon projection of proteins belonging to definite structure classes.

The AAC of 64 proteins classified by Chou (Chou, 1989) to specific structural classes were used to calculate the mole percent for all 20 amino acids. After that, each protein was described as a unique vector in composition space with coordinates in the range of 0 to 1 corresponding to mole percent for each particular amino acid. We numbered the 64 proteins as 1,2,3…,64 according to the order listed in table 5-8 of Chou (Chou, 1989) or tables 2-5 of Zhang and Chou (Zhang and Chou, 1992). All 64 protein vectors in 19-D composition space were then projected into 2-D space using the Sammon nonlinear algorithm. Typically results were analyzed after 250 iterations using steepest descent optimization with a learning factor MF = 0.3. Several other projections have been made starting from different random distributions of points. In all cases the results were similar differing primarily in the orientation of the projection and magnitude of the resulting error. The most variability was observe for proteins belonging to α+β class. The best projection, based on the smallest Sammon's error (0.0747), accomplished in these experiments is presented in Fig. 2. The protein names are listed in tables 1-4. The four different colors correspond to four different structural classes: red (1-19) - 19 α proteins; purple (20-34) - 15 β proteins; blue (35-48) - 14 α+β proteins; green (49-64) - 16 α/β proteins. It is apparent that structural classes formed recognizable groupings. However, a few points are clear outliers and are located within other sets. In general, points which were "misplaced" correspond to the proteins which were also mis-assigned by Zhang and Chou’s prediction algorithm (Zhang and Chou, 1992). For example: point 32 corresponds to Rubredoxin which was initaily clasified as β protein. Its unusual position can be explain by the lowest level of β structure (25%) found in that set. The Zhang and Chou algorithm predicts α/β structural class for Rubredoxin. Similar arguments can be made for point 38, which represents a High-potential Iron Protein, from α+β class. In this case only 27 % of the total protein has defined structure, and the α structural class was predicted by Zhang and Chou. In our projection this protein also falls into the cluster of α proteins. Similar arguments could be made for other outlying points: 23, 33, 46, 62. This suggests, that the irregular portion of these proteins can be the source of their structural missassigment. The presence of irregular protein structure can significantly influence the AAC and may result in inaccurate class prediction.

The relationship within each set of proteins was analyzed by complete-linkage cluster analysis. Cluster analysis can provide an objective automated way to group objects into the clusters in multidimensional spaces. The correlation coefficient of each cluster corresponds to the maximum distance at which two objects are still consider to have similar properties. We defined the correlation coefficient as the median 2-D Euclidean’s distance for the set after excluding outliers (points: 23, 32, 33, 38, 46, 62). The results of the cluster analysis are presented as the shading surrounding each set in figure 2. It appears that the structural classes of proteins are separated in 2-D space which corresponds to their unique AAC. This result supports the early hypothesis that the folding of proteins is determined by their AAC (Hatch, 1965; Harding, 1984; Chou, 1989; Zhang and Chou, 1992; Dubchak et al. 1993; Nakai et al. 1988; Nakashima et al. 1986;

Chun-Ting et al. 1992). However, we observed a significant degree of overlapping between clusters which may explain the inaccurate prediction of protein folding based solely on AAC. The irregular portion of proteins, which represents some fraction of all proteins, may have signification contributions to the overall ambiguity of the predicted classification.

Projection of proteins with unusual amino acid composition.

We decided to investigate if Sammon's projection can be used to predict folding class of "unknown" proteins. Several hundred random proteins from the Pir1 protein data base were projected (one at a time) into the developed map. In this modified projection, coordinates of the 64 representative proteins which form the contour map were held constant and only the coordinates of new proteins were optimized by Sammon's algorithm. Each newly projected protein was treated as the 65th element. It was noticed that several proteins fell outside the area occupied by the four clusters of the representative set of structurally distinct proteins. We have noticed that these proteins were small and/or showed unusual amino acid composition. Several of these proteins were described as unusual by Cornish-Bowden during his research on dependencies between the size and AAC of proteins (Cornish-Bowden, 1983). In addition, we observed that if these proteins were projected several times over, results of the optimization were significantly different. This is due to missing reference points outside the area occupied by set of 64 representative proteins. To avoid that ambiguity, an extra 27 proteins with unusual AAC were selected from the Pir1 protein data base and added into the representative set. These added proteins expanded the representative set from 64 to 91.

These proteins were projected using the original Sammon's algorithm. As a result, a new extended contour map was formed (Fig. 3). A list of proteins used in this projection, references/accession number, and the final coordinates (indices) are listed in table 1. A large number (>1200) of random proteins for Pir1 were projected into the new extended contour map using modified Sammon's mapping procedure. This time we observed significantly less variability during a multiple projection of the same proteins. Each protein gave a distance point in 2-D map. It appears that each newly projected protein could be characterized by unique index of X,Y coordinates on a 2-D map which reflects their unique amino acid composition. This could offer a new way to classify the proteins based solely on their AAC.

Projection of amino acid composition of hemoglobins into the extended contour map.

The amino acid composition of porcine hemoglobin alpha chain was projected into this extended contour map as an additional 92nd protein. As previously described, the X,Y coordinates of 91 proteins from the extended contour map were displayed as fixed points. The projected porcine hemoglobin fell into the cluster of the α proteins, close to the other hemoglobins used in the representative set. The observed index for the α chain of the porcine hemoglobin was 0.623;0.5713 which is very close to the other hemoglobins used in the representative set (see table 1). This operation of the projection of the 92nd protein was repeated for the beta chain of porcine hemoglobin and 322 other different alpha and beta globins with a similar result. All of the projected α and β chains formed a distinct well defined cluster (Fig. 4). Approximately 30-40 iterations (MF =0.3) were required to finish the projection based on insignificant changes in the Sammon error.

Additionally, a subset of proteins use by Nakashima *et al*. (Nakashima et al. 1986) for their predictions were projected into the extended contour map. Interestingly, almost all of them fell into the clusters to which they were previously classified except for a few α+β proteins. This is not surprising since the projection of α+β protein from 64 Chou's proteins also created a several outliers. These results indicate that the Sammon nonlinear projection can be used to predict the structural classes of unknown proteins. Although some ambiguity in the assignment may remain due to the overlapping of structural class clusters.

The Sammon projection using modified distances or reduced alphabet of amino acids.

Landes and Risler reported successful use of reduced amino acid alphabet in searching protein data bases (Landes and Risler, 1994). In their work they reduced the 20 amino acids to 10 symbols as follows: (A=T=S), C, (D=E=N), (F=Y), G, H, (I=L=M=V),(K=Q=R), P, W. We were interested if the reduction of the original compositional space into 9-dimension would affect the clustering of the projected proteins. The application of the reduced alphabet into the Sammon projection resulted in an irregular distribution of points (Fig. 5A). Clusters were still visible, however they were not as clearly separated as in the case of the nominal projection from 19-D compositional space.

Chou’s set of 64 proteins were again projected into 2-D space using Sammon’s algorithm. However in this particular case the calculation of Euclidean’s distance was modified as follows:

$$D^*_{pq} = \sqrt{\sum_{i=1}^{20} w_i (P_i - Q_i)^2}$$

The distances in compositional space were calculated as the product of w_i the weighting factor and Euclidean’s distance for each amino acid, as described by Chun-Ting *et al.*(Chun-Ting et al. 1992). These researchers performed predictions of structural class of proteins from AAC based on a linear-programming approach. We were interested to what degree this weighting factor developed for a linear fit would affect nonlinear projection. The results (Fig. 5B) indicated that these weighting factors did not improved clustering of the proteins belonging to the same structural class.

Conclusion

The nonlinear Sammon’s mapping algorithm may be a useful tool to examine complicated multidimensional systems in protein science. When applied to project 19-D compositional space of proteins, it can provide a new way of mapping and indexing. We demonstrated that proteins with similar functionality can be mapped to the same region in 2-D space. This may allow prediction folding classes and potentially functional properties of newly sequenced proteins based on compositional indices.

Our results suggest that prediction of protein folding based on amino acid composition may never overcome certain limitations such as overlapping clusters. Although, different structural classes of proteins form distinct clusters which can be visualized after projection to 2-D space, these clusters have a tendency to overlap. This may result in ambiguous assignment of structural classes of unknown proteins especially proteins sharing significant contribution of irregular folding. It is possible that extending the number of proteins in the representative (learning) set may limit overlapping and improve prediction accuracy.

We must also point out the computational advantage of the contour map: The original Sammon's algorithm, and its application to the protein classification as proposed by Agrafiotis (Agrafiotis, 1997), required n^2 computations per iteration of the steepest decent algorithm discussed in the Method section, where *n* is the number of projected proteins. In the proposed modified method, proteins have been projected onto contour map (in our experiments *n*=91), and new proteins were added by comparing them only one by one to these *n* points. Therefore, projection of compositional space for a new protein costs only 'n' steps per iteration. This translates to a significant saving in computation time.

Material and Methods

Calculation of Euclidean's distance between proteins in the composition space.

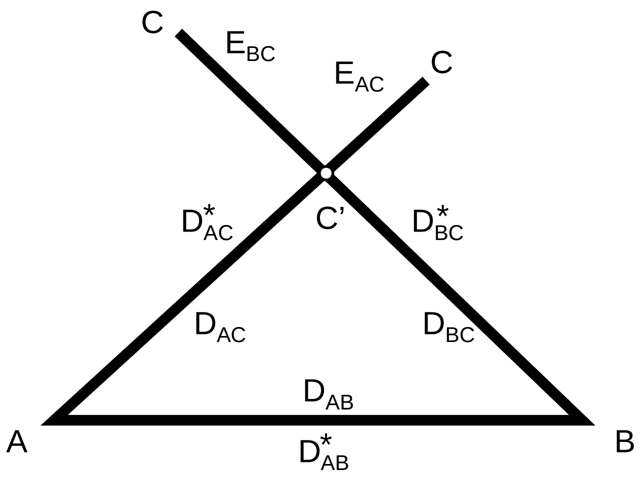
In the first experiment presented in Figure 1 we used the standard Euclidean's distance between two points. For all other experiments, we computed the distance between two proteins (level of dissimilarity) based on the amino acid composition, that is:

$$D^*_{pq} = \sqrt{\sum_{i=1}^{20} \left(P_i - Q_i\right)^2}$$

where P_i , Q_i are mole percent of the *i*th kind of amino acid in the proteins P and Q. Each protein corresponds to a point whose coordinates are given by the mole percent of the 20 constituent amino acids. Different modifications of this approach have been reported by other researchers (Chou, 1989; Zhang and Chou, 1992; Nakai et al. 1988; Nakashima et al. 1986; Chun-Ting et al. 1992; Cornish-Bowden, 1983).

The Sammon projection

The projection of 19-D compositional space onto the 2-D Euclidean space was obtained according to the original Sammon nonlinear projection algorithm (Sammon, 1969; Agrafiotis, 1997). This algorithm tries to approximate in the best possible way (i.e., in the squared error sense) a relation ship between points in a multidimensional space when projected into 3-D spaces. The algorithm is presented in details by Sammon(Sammon, 1969), so we only illustrate its meaning here on a simple example. Observed, that distance relationship in a higher dimension cannot be preserved in a lower dimensional space. For example, imagine three points A, B and C in 3D space with given distances between them, say D^*_{AB}, D^*_{AC} and D^*_{BC}. These distances cannot be preserve in the projected 2-D space (D_{AB}, D_{AC} and D_{BC}). The best you can do is to preserve the distance between points A and B and make a small error (E_{AC} and E_{BC}) as possible when locating point C. This is illustrated in figure bellow



In general, the goal of Sammon's algorithm is to minimize the discrepancy in "distance" which was defined as the error of the projection:

$$E(x,y)=\frac{1}{c}\sum_{i<j}^{n}\frac{\left(D_{ij}^{*}-D_{ij}(x,y)\right)^{b}}{D_{ij}^{*}} \qquad (1)$$

where D_{ij}^{*} is Euclidean distance in old space (19 dimensional compositional space),

$D_{ij}(x,y)=\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2}$ Euclidean distance in 2-D space which is function of *x*, *y* coordinates, *b* is a parameter, *c* is constant and *n* is a number of proteins in the learning set. It should be pointed out the parameter *b* can model a variety of situations (Szpankowski, 1993). Throughout the computation, as in Agrafiotis (Agrafiotis, 1997), we assume b=2 and $c=\sum_{i<j}^{n}D_{ij}^{*}$.

To find optimal coordinates of a point (x, y) a numerical optimization procedure called the steepest descent was applied. In this method, the optimal solution is found in several iteration starting from a (random) initial point. In each iteration we move towards the gradient of the function E(x,y). In the m^{th} iteration we compute the m^{th} estimate of *E(x,y)*, written as $E^{(m)}_{(x,y)}$. The next iteration coordinates of each point $x^{(m+1)}$, $y^{(m+1)}$ are computed according to the following formula:

```
for $m=1$ to number of iterations
    for $i=1$ to $n$
        $x_i^{(m+1)} = x_i^{(m)} - MF \frac{\left|\frac{\partial E^{(m)}_{(x,y)}}{\partial x}\right|}{\left|\frac{\partial^2 E^{(m)}_{(x,y)}}{\partial x^2}\right|}$
        $y_i^{(m+1)} = y_i^{(m)} - MF \frac{\left|\frac{\partial E^{(m)}_{(x,y)}}{\partial y}\right|}{\left|\frac{\partial^2 E^{(m)}_{(x,y)}}{\partial y^2}\right|}$
    end
end.
```

where MF (‘‘magic factor”) is an experimentally determined coefficient. The first and the second derivative of *E(x,y)* with respect to x are shown below (in similar manner one can compute the derivatives with respect to y):

$$\frac{\partial E^{(m)}_{(x,y)}}{\partial x} = \frac{-2}{c}\sum_{\substack{i=1\\ i\neq j}}^{n} \frac{\left(D^*_{ij} - D_{ij}(x,y)\right)}{D_{ij}(x,y)}\left(x_i - x_j\right) \qquad (2)$$

$$\frac{\partial^2 E_{(x,y)}^{(m)}}{\partial x^2} = \frac{-2}{c}\sum_{\substack{i=1\\ i\neq j}}^{n} \frac{1}{D_{ij}^{*}D_{ij}}\left[\left(D_{ij}^{*}-D_{ij}\right)-\frac{\left(x_i-x_j\right)^2}{D_{ij}}\left(1+\frac{D_{ij}^{*}-D_{ij}}{D_{ij}}\right)\right] \qquad (3)$$

In the implementation of the steepest descent method we stopped the iteration procedure after 250 cycles.

In the experiments when additional protein *n+1* was projected into the contour map the same algorithms were used, except that in the iteration procedure was used only to optimize the distance of the new protein without affecting the distances already optimized between elements of the learning set. The following modified formula was used:

```
**for** $m=1$ **to** *number of iterations*
    $x_{n+1}^{(m+1)} = x_{n+1}^{(m)} - MF\frac{\left|\frac{\partial E_{(x,y)}^{(m)}}{\partial x}\right|}{\left|\frac{\partial^2 E_{(x,y)}^{(m)}}{\partial x^2}\right|}$
    $y_{n+1}^{(m+1)} = y_{n+1}^{(m)} - MF\frac{\left|\frac{\partial E_{(x,y)}^{(m)}}{\partial y}\right|}{\left|\frac{\partial^2 E_{(x,y)}^{(m)}}{\partial y^2}\right|}$
**end.**
```

In addition the summation of error E_{xy} (eq. 1) is over single index *i* (this is only in terms of the sum).

Computer programs.

Source code of computer programs were written in Borland Turbo Pascal® version 7. Executable versions of programs for Windows® (Helix.exe and SammProj.exe) used in this paper and corresponding contour maps are available at http://www.cs.purdue.edu/people/spa/.

Acknowledgments

We express our special thanks to Marcin Apostol for many insightful suggestions during writing source code of computer programs. We also thank our colleagues Dr. Spencer Anthony-Cahill, Dr. Jeff Etter and Dr. Michael Weickert for reviewing the manuscript.

References

Fasman GD. Ed. 1989. *Prediction of Protein Structure and the Principles of Protein Conformation*, New York: Plenum Press,

Agrafiotis DK. 1997. A new method for analyzing protein sequence relationships based on Sammon maps. *Protein Sci*. 6. 287-293.

Chou PY. 1989. Prediction of protein structural classes from amino acid composition. In: Fasman GD. Ed. *Prediction of Protein Structure and the Principles of Protein Conformation*. New York: Plenum Press, pp. 549-586.

Chun-Ting Z, Xinhua X, Genfa Z. 1992. A weighting method for predicting protein structural class from amino acid composition. *Eur J Biochem*. 210. 747-749.

Cornish-Bowden A. 1983. The amino acid compositions of proteins are correlated with their molecular sizes. *Biochem J*. 213. 271-274.

Dubchak I, Holbrook SR, Kim SH. 1993. Prediction of protein folding class from amino acid composition. *Proteins*. 16. 79-91.

Garnier J, Gibrat JF, Robson B. 1996. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol*. 266. 540-553.

Harding JJ. 1984. The prediction of repetitive protein sequences from amino acid compositions. *Biochem J*. 217. 339-340.

Hatch FT. 1965. Correlation of amino-acid composition with certain characteristics of proteins. *Nature*. 206. 777-779.

Landes C, Risler JL. 1994. Fast databank searching with a reduced amino-acid alphabet. *Comput Appl Biosci*. 10. 453-454.

Lesk AM, Boswell DR. 1992. Does protein structure determine amino acid sequence? *Bioessays*. 14. 407-410.

Mayoraz E, Dubchak I, Muchnik I. 1995. Relation between protein structure, sequence homology and composition of amino acids. *Ismb*. 3. 240-248.

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

Nakai K, Kidera A, Kanehisa M. 1988. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng*. 2. 93-100.

Nakashima H, Nishikawa K, Ooi T. 1986. The folding type of a protein is relevant to the amino acid composition. *J Biochem (Tokyo)*. 99. 153-162.

Sammon JW. 1969. A nonlinear mapping for data structure analysis. *IEEE Trans Comp*. C-18. 401-409.

Szpankowski W. 1993. A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors

. *SIAM J Computing*. 22. 1176-1198.

Tuckwell DS, Humphries MJ, Brass A. 1995. Protein secondary structure prediction by the analysis of variation and conservation in multiple alignments. *Comput Appl Biosci*. 11. 627-632.

Zhang CT, Chou KC. 1992. An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci*. 1. 401-408.

Table 1. The X,Y indexes obtained using Sammon mapping for 19 α proteins

I.D.	Name	X	Y
1	Calcium-binding parvalbumin (carp)	0.607	0.645
2	Cytochrome b_{562} (*E. coli*)	0.546	0.660
3	Cytochrome c (tuna)	0.455	0.556
4	Cytochrome c_2 (*R. rubrum*)	0.540	0.616
5	Cytochrome c_{550} (*P. denitrificans*)	0.499	0.585
6	Cytochrome c_{555} (*C. thiosulfatophilim*)	0.699	0.501
7	Hemerythrin B (*G. Gouldii*)	0.436	0.580
8	Methemerythrin (*T. dyscritum*)	0.427	0.561
9	Methemerythrin (*T. pyroides*)	0.461	0.585
10	α-methemoglobin (horse)	0.617	0.562
11	β-methemoglobin (horse)	0.573	0.596
12	α-deoxyhemoglobin (human)	0.616	0.581
13	β-deoxyhemoglobin (human)	0.581	0.580
14	γ-deoxyhemoglobin (human fetal)	0.550	0.542
15	Hemoglobin (glycera)	0.651	0.535
16	Hemoglobin (lamprey)	0.586	0.530
17	Hemoglobin (midge larva)	0.599	0.522
18	Myoglobin (scal)	0.509	0.628
19	Myoglobin (sperm whale)	0.501	0.624

Table 2. The X, Y indexes obtained using Sammon mapping for 15 β proteins

I.D.	Name	X	Y
20	α-chymotrypsin (bovine)	0.573	0.454
21	Concanavalin A (jack bean)	0.577	0.475
22	Elastase (porcine)	0.555	0.407
23	Erabutoxin B (sea snake)	0.453	0.338
24	Immunoglobulin Fab (V_H and C_H, human)	0.593	0.405
25	Immunoglobulin Fab (V_L and C_L, human)	0.588	0.444
26	Immunoglobulin B-J MCG (human)	0.573	0.427
27	Immunoglobulin B-J REI (human)	0.551	0.359
28	Penicillopepsin (*P. janthinellum*)	0.624	0.417
29	Prealbumin (human)	0.554	0.490
30	Protease A (*S. griseus*)	0.627	0.386
31	Protease B (*S. griseus*)	0.618	0.357
32	Rubredoxin (*C. pasteurianum*)	0.315	0.504
33	Superoxide dismutase (bovine)	0.453	0.472
34	Trypsin (bovine)	0.540	0.388

Table 3. The X, Y indexes obtained using Sammon mapping for 14 α+β proteins

I.D.	Name	X	Y
35	Actidin (kiwi fruit)	0.504	0.429
36	Cytochrome b_5 (bovine)	0.423	0.541
37	Ferredoxin (*P.aerogenes*)	0.534	0.300
38	High-potential iron protein (chromatium)	0.666	0.560
39	Insulin (A and B chains, porcine)	0.422	0.399
40	Lysozyme (bacteriophage T_4)	0.474	0.531
41	Lysozyme (chicken)	0.466	0.422
42	Papain (papaya)	0.475	0.427
43	Phospholipase A_2 (bovine)	0.404	0.432
44	Pibonuclease S (bovine)	0.503	0.394
45	Staphylococcal nuclease (*Staphylococcus aureus*)	0.482	0.596
46	Subtilisin inhibitor (streptomyces)	0.636	0.485
47	Thermolysin (*B. thermoproteolyticus*)	0.534	0.436
48	Trypsin inhibitor (bovine)	0.385	0.449

Table 4. The Y, Y indexes obtained using Sammon mapping for 16 α/β proteins

I.D.	Name	X	Y
49	Adenyl kinase (porcine)	0.447	0.523
50	Alcohol dehydrogenase (horse)	0.521	0.497
51	Carbonic anhydrase B (human)	0.537	0.480
52	Carbonic anhydrase C (human)	0.496	0.540
53	Carboxypeptidase A (bovine)	0.521	0.459
54	Carboxypeptidase B (bovine)	0.497	0.468
55	Dihydrofolate reductase (*E. coli*)	0.474	0.500
56	Flavodoxin (*Clostridium* MP)	0.407	0.512
57	Glyceraldehyde 3-P dehydrogenase (lobster)	0.547	0.511
58	Glyceraldehyde 3-P dehydrogenase (*B. stearothermophilus*)	0.567	0.542
59	Lactate dehydrogenase (dogfish)	0.511	0.524
60	Phosphoglycerate kinase (horse)	0.531	0.550
61	Rhodanese (bovine)	0.500	0.500
62	Subtilisin BPN’ (*B. amyloliquefaciens*)	0.631	0.453
63	Thioredoxin (*E. coli*)	0.531	0.590
64	Triose phosphate isomerase (chicken)	0.530	0.537

Table 5. X, Y indices obtained using Sammon mapping for 27 extra proteins

I.D.	Accession #	Name	X	Y
65	HHBYD8	Heat shock protein DDR48 - Yeast	0.200	0.517
66	TNHUA	Prothymosin alpha - Human	0.333	0.747
67	TISYD	Proteinase inhibitor (Bowman-Birk) D-II - soybean	0.389	0.275
68	FECF	Ferredoxin - Chlorobium sp.	0.399	0.342
69	PIHUPF	Basic proline-rich peptide P-F - Human	0.096	0.344
70	HSBOS	Sperm histone - Bovine	0.170	0.100
71	EWBY8	H+-transporting ATP synthase (EC 3.6.1.34)	0.592	0.755
72	C32038	mu-agatoxin III - funnel-weaving spider	0.504	0.220
73	QMVHMM	mastoparan M - hornet	0.854	0.663
74	JTJG3	Tremerogen a-13 - Basidiomycete	0.667	0.168
75	XASNBA	Bradykinin-potentiating peptide B - Mamushi	0.112	0.273
76	AKLQ	Adipokinetic hormone - Migratory locust	0.423	0.790
77	SPPGNK	neuromedin K - pig	0.861	0.407
78	TPRBTS	Troponin T, skeletal muscle - Rabbit	0.410	0.656
79	SMHU1F	Metallothionein 1F - Human	0.554	0.151
80	UNBO	neurotensin - bovine	0.237	0.582
81	QMWAVV	mastoparan - yellowjacket	0.854	0.663
82	MXKN1	mu-conotoxin GIIIA - cone shell	0.345	0.174
83	QFBO	micro glutamic acid-rich protein - bovine	0.302	0.920
84	FDFI8G	antifreeze protein GS-8 - grubby sculpin	0.842	0.926
85	EEWTG	gamma-gliadin B precursor - wheat	0.742	0.274
86	SNUMP	sillucin - Rhizomucor pusillus	0.465	0.209
87	KGZQHF	histidine/alanine-rich protei	0.677	0.813
88	DNVPBF	DNA-binding protein - budgerigar fledgling diseases	0.262	0.726
89	W5WLEB	E5 protein - bovine papillomavirus type 1	0.827	0.575
90	QQBE3	BHLF1 protein - human herpesvirus 4	0.292	0.377
91	VHNVBM	nucleocapsid protein	0.097	0.435

Figure 1. Sammon projection of a helix: A-starting reandom distriubutin of 50 points, B-optimized projection (map) of 50 points after 250 iterations, C-contour map of 49 points (1 selected point is missing form the map of 50 points), D-projection of the missing 50th point into the contour map.

Figure 2. Contour map of compositional space of Chou’s 64 proteins (Chou, 1989) belonging to four different folding classes. Red points (1-19) - 19 a proteins, purple points (20-34) - 15 b proteins, blue points (35-48) - 14 a+b proteins, green points (49-64) - 16 a/b proteins. Shading represents clustering for each class.

Figure 3. Contour map of 91 proteins, The set of 91 proteins included Chou’s (Chou, 1989) 64 proteins (table 1-4) and additional 27 proteins (table 5) with unusual composition.

Figure 4. Mapping of 322 alpha and beta globins of different hemoglobins (open squares) into the 91 proteins contour map.

Figure 5. Sammon projection of Chou’s 64 proteins (Chou, 1989): A - using a reduced amino acid alphabet (Landes and Risler, 1994), B - using weighting factors (Chun-Ting et al. 1992).