Bioinformation
www.bioinformation.net

open access
Hypothesis

Clustering tendency in the protein fold space

Oliviero Carugo

Department of General Chemistry, Pavia University, viale Taramelli 12, I-27100 Pavia, Italy and Department of Biomolecular Structural Chemistry, MFPL - Vienna University, Campus Vienna Biocenter 5, A-1030 Vienna, Austria. Phone:+43 1 4277 52208, Fax:+43 1 4277 9255, Email:oliviero.carugo@univie.ac.at

Received June 20, 2009; Accepted July 23, 2009; Published February 28, 2010

Abstract:
Several non-redundant ensembles of protein three-dimensional structures were analyzed in order to estimate their natural clustering tendency by means of the Cox-Lewis coefficient. It was observed that, despite proteins tend to aggregate into different and well separated groups, some overlap between different clusters occurs. This suggests that classifications bases only on structural data cannot allow a systematic classification of proteins. Additional information are in particular needed in order to monitor completely the complex evolutionary relationships between proteins.

Keywords: Cluster analysis; clustering tendency; protein fold; protein structural domains; protein structure classification.

Background:
During the last few years, the common paradigm that protein folds tend to be mutually exclusive and to cluster into well separated groups started to be criticized. The expression "gregariousness" was used to indicate the number of close neighbors of each fold and such a concept was applied to examine weather the fold space is a continuum, where existing motifs are used to enlarge old folds and create new types of structures **[1].** It was observed that high levels of gregariousness are observed when different folds contain the same motif **[1].** The protein structure space was also analyzed by Kim and co-workers, by using several representative data sets and several computational approaches **[2, 3].** It was observed that protein structures can be discriminated essentially by three features: the prevalence of residues that adopt a α or a β secondary structure and the presence of α-β-α motifs (two flanked parallel β strands separated by an antiparallel α helix) **[2, 3].** However, significant overlaps were observed amongst different structural classes. Folds in the overlap regions contain features of both classes, and this was considered the main reason why structure-based function predictions are not very efficient **[2].** A detailed review on the nature of the protein fold space showed both advantages and disadvantages of considering it a continuous and multidimensional object rather that an ensemble of discrete categories **[4].**

In the present paper, a robust statistical approach is adapted to the problem of the estimation of the degree of clustering within the fold space. The term "clustering tendency" refers to the problem of deciding weather the subjects have an intrinsic predisposition to cluster into distinct groups or they are randomly arranged. This is also referred to as the spatial randomness problem and while intrinsically aggregated subjects are characterized by mutual attraction, randomly arranged subjects show mutual repulsion **[5].** The clustering tendency was estimated with the Cox-Lewis coefficient **[5, 6]** on different datasets and it was observed that protein folds are partially overlapping.

Methodology:
Data Selection:
Four types of data sets were selected. Care was taken to avoid redundancies between the data. In fact, a purely random selection of a set of protein structures might produce results completely biased. The clustering tendency measured in a data set that contains several proteins nearly identical to each other would be considerably overestimated since many experimental points would be extremely close to each other.

Scop_fold data set. A representative example of each protein fold was taken from the Scop database of protein domain structures **[7].** Only the four most populated classes (α, β, α/β, and α+β) were considered. Entries containing "unobserved" residues were disregarded. 624 files were retained.

Scop_fold/X data sets. 10 subsets, each containing 62 structures, of the *Scop_fold* data set were randomly built (X = 1, 10). They do not overlap with each other.

Pisces data set. Protein chains were taken from the Pisces database **[8].** Their crystal structures were determined at resolution not worse than 2 Å and the maximal sequence identity between two of tem is 25%. Entries containing residues "unobserved" in the electron density maps were removed. 2237 structures were retained.

Pisces/X data sets. 10 non-overlapping subsets, each containing 223 structures, of the *Pisces* data set were randomly built (X = 1, 10).

Structural similarity between proteins
A very wide variety of techniques were used to compare pairs of protein structures **[9-11]**. In the present manuscript, we used a technique that allows one to represent a structure with a geometrical point in a n-dimensional space and to select a random (geometrical) point in the fold space (this is necessary to evaluate to clustering tendency; see below). This task cannot be accomplished, in general, by using protein structure comparison techniques, since the similarity scores are nearly never metrics, in the mathematical sense. An exception is the program GI, in which the protein topology is described always by 30 numbers - its number of residues and its Gauss integrals - independently of the protein dimension **[12]**. Therefore, each protein, either large or small, is associated with a point in a space defined by 30 variables. The distance between two protein structures can be measured with the Euclidean distance between two points in this space. A further advantage of this method of protein structure comparison is its extreme velocity (thousands of comparisons can be made in few minutes).

Cox-Lewis coefficient of clustering
The Cox-Lewis coefficient is defined in the following way **[5, 6].** Given m proteins, each characterized by 30 variables, $k \ll m$ geometrical points are randomly selected in the 30-dimensional space. The smallest distance u_i between the i^{th} point and one of the m proteins is then recorded together with the minimal distance w_i between such a protein and another protein (see **Figure 1**). The ratio R is given using equation 1 & 2 **(supplementary material)** is the Cox-Lewis coefficient. R values close to 1 are expected for uniformly distributed data, because $w_i \approx u_i$. Values much larger than 1 are on the contrary expected to arise if the subjects tend to cluster into well defined groups. Given that the R values clearly depend on the procedure of selection of random points and on the treatment of the border conditions, this procedure is described in detail in the next paragraph.

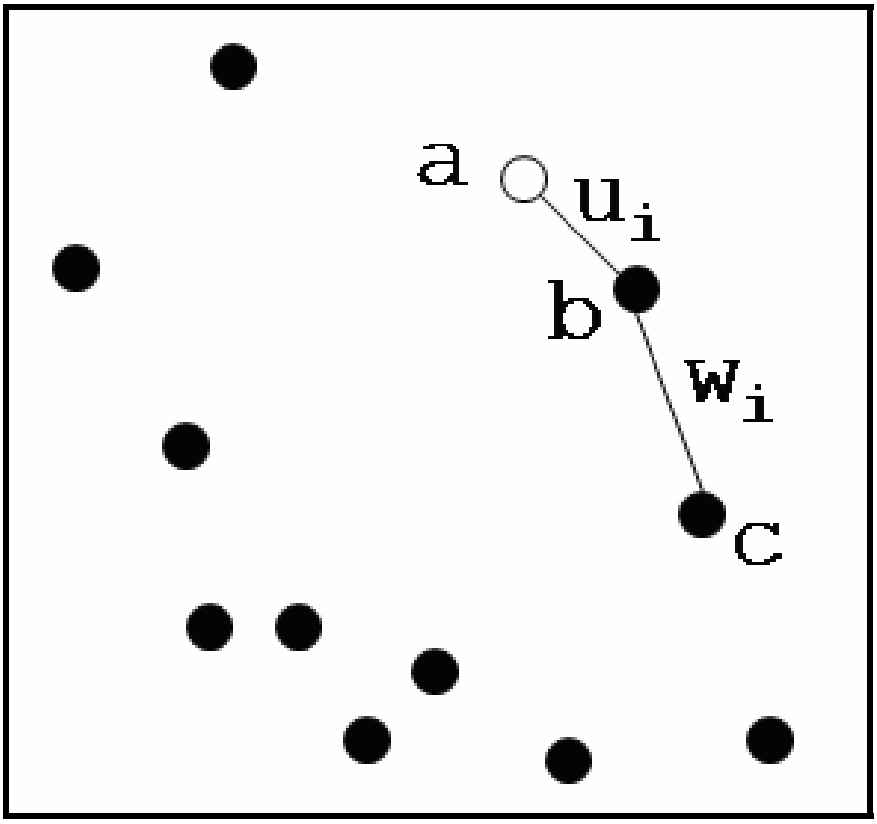

Figure 1: Example of Cox-Lewis statistics. Filled circles represent subjects in a bi-dimensional space. Open circles represent geometrical points. u_i is the minimal distance between a geometrical point (a) and a subject (b). w_i is the minimal distance between the same subject (b) and another subject (c). The geometrical point (a) is randomly selected. The geometrical point (a) is randomly selected. The Cox-Lewis coefficient can be computed with equation (2) on the basis of a set of randomly selected points (a).

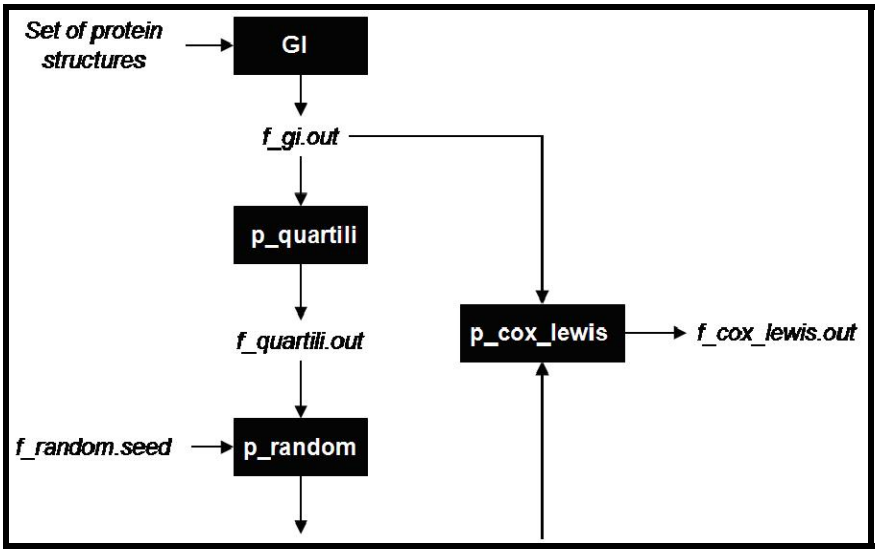

Figure 2: Scheme of the computations performed to obtain the Cox-Lewis coefficient of clustering tendency.

Details about the computations

Computations were performed as schematized in **Figure 2**. The output file of program **GI**, *f_gi.out*, contains the 30 variables necessary to describe each protein. If there are *np* proteins, *f_gi.out* is a table of *np* lines and 30 columns. The program **p_quartili** determines, for each of the 30 variables, the first and the last quartile, outputted in the file *f_quartili.out*. Together with the file *f_random.seed*, which contains a randomly selected integer number, the file *f_quartili.out* is read by the program **p_random** that generates *k* random numbers ranging from the first to the last quartile of each of the 30 variables that represent a protein structure. The value of *k* was defined as *np*/10, where *np* is the number of proteins described in the file *f_gi.out*. It is essential to select

random numbers within the two quartiles in order to avoid insidious problems at the boundaries of the protein fold space. Eventually, the files *f_gi.out* and *f_random.out* are read by the program **p_cox_lewis**, which computes the Cox-Lewis coefficient of clustering tendency. 500 different values of the seed (contained in the file *f_random.seed*) were randomly generated and 500 values of the Cox-Lewis coefficient were computed for each data set and averaged.

Discussion:

Table 1 (see supplementary material**)** shows the minimal, maximal, and average values (and the standard deviation of the mean) of the Cox-Lewis coefficient computed on 22 different data sets. The average values oscillate amongst different data sets. However, they are always significantly larger than one, the value that would indicate that protein fold structures are uniformly distributed. This first conclusion is therefore that protein structures tend to cluster into separate groups. Noteworthy, similar results were obtained also by using the Hopkins coefficient, another measure of clustering tendency.

A second observation is interesting. The Cox-Lewis coefficients were computed on two types of data sets. Eleven of them (Scop and Scop/X) were based on the Scop database of protein domain structures, where the redundancy was reduced essentially on the basis of structural features **[7].** In the other eleven of them (Pisces and Pisces/X), the redundancy was reduced only on he basis of the amino acid sequences **[8].** Despite these approaches are different (structure- and sequences-based redundancy reduction), the Cox-Lewis coefficients are nearly the same. The fact that smaller values tend to be observed for the Pisces data sets is likely dependent on the fact that these data sets contain entire protein chains, sometime made by more than a single domain and sometime participating to permanent oligomeric assemblies. Consequently, and this is the second conclusion, coefficients close to 1.3-1.4 are likely to be rather reliable estimations of the clustering tendency of protein structures.

Are these values really high? To answer this question, one would need to "see" the 30-dimensional fold space defined by the GI approach **[12].** This is impossible. Human perception is limited to two or three dimensions. In principle, reductions of dimensionality are possible, for example by using principal component analysis [13]. However, this is impossible in our case, since the first three principal components, out of the original 30 variables, accounts for an insufficient fraction of the original, overall variance (less than 70%). The visualization in a reduced 3-dimensional space would be useless, since about one third of the overall variability amongst structures would be ignored.

The only possibility to get a visual assessment is thus based on simulations, the simplest of which are in two dimensions. Four sets of bi-dimensional data, each containing 1000 points, were generated by using a pseudo random number generator. One contained 1000 points within distance = *R* from the point (-1, 0) (arbitrary units, a.u.); the second was centred on the point (1, 0); the third on the point (0, -1); and the fourth on the point (0, 1). By increasing *R*, it is possible to reduce arbitrarily the clustering tendency of the entire data set of 4,000 points. It is thus possible to see, with a simple plot, how the clustering tendency changes and to relate is with the Cox-Lewis coefficient. This is shown in **Figure 3**. For small *R* values, the data are clearly segregated into four clusters and the Cox-Lewis coefficient is high (6.93). For large values of *R*, the four clusters are completely superposed to each other and the Cox-Lewis coefficient decreases to 1.04 and approaches the value of one, expected for data uniformly distributed. If the Cox-Lewis coefficient is close to 1.3-1.5 (values close to those that are shown in **Table 1 – supplementary material**), the four clusters are partially superposed. About 6-7% of the members of a cluster tend to invade a neighbour group.

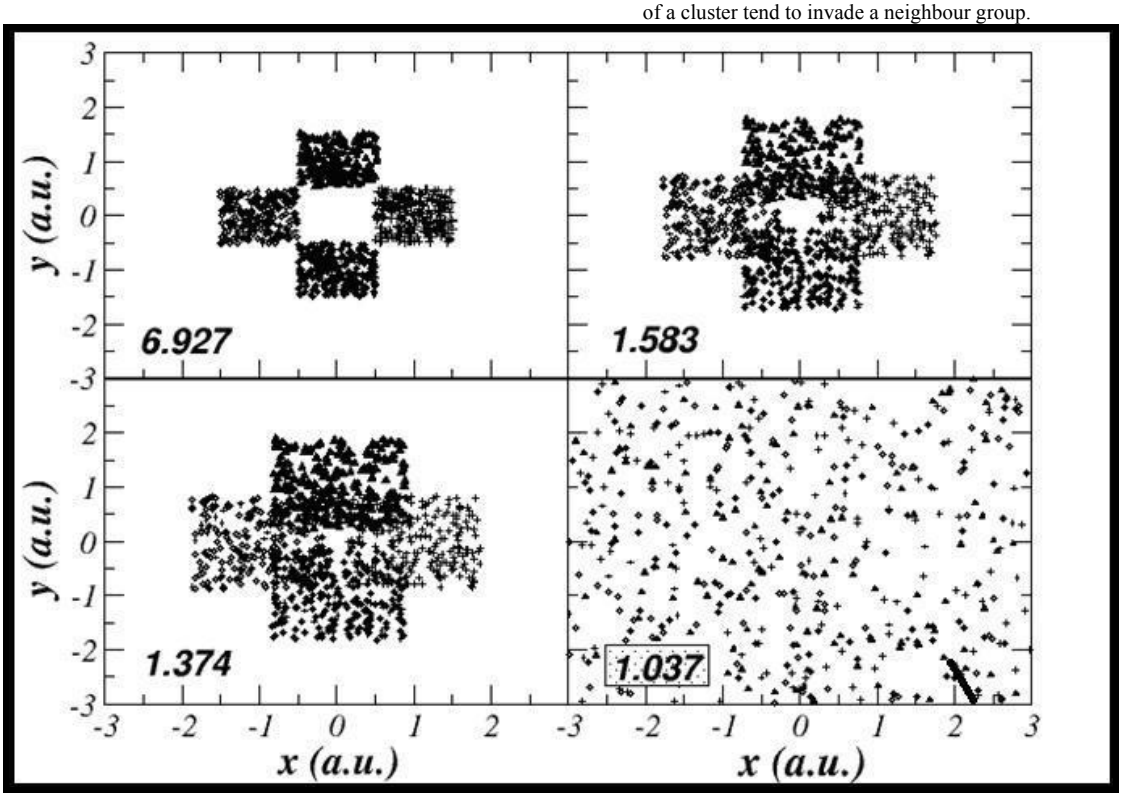

Figure 3: Examples of various levels of clustering tendency in a bi-dimensional space. Four types of objects are plotted in each graph (white diamonds, black diamonds, plus symbols, and black triangles). The Cox-Lewis coefficient decreases from 6.927 to 1.037 if the separation between the four clusters decreases. Both axes are in arbitrary units (a.u.).

Conclusion:
This implies that protein fold structures have a natural tendency to aggregate into different groups and that it is, as a consequence, infrequent that a structure of a certain type is observed into a cluster that groups other structure types. However, some overlap between different clusters is possible and seldom observed, with the consequence that false positives or negatives cannot be completely avoided in the various structure-based prediction methods that were designed. As expected and as already observed **[2, 3],** the superposition between different types of protein 3D structure clusters tends to occur for the cases that are known to be relatively similar. Typically, this mix up is observed for structures that are essentially α or β, on the one hand, and α+β, on the other, according to the SCOP classification. Obviously, this does not mean that the classification adopted in the Scop database is useless or inappropriate. This only means that a description of the fold space based only on structural features cannot produce well isolated islands.

Acknowledgment:
This work was partially funded by the BIN-II and BIN-III programs of the Austrian GEN-AU.

References:
[1] A Harrison *et al., J Mol Biol* , 323: 909 (2002) [PMID: 12417203]
[2] J Hou *et al.*, *Proc Natl Acad Sci U S A,* 102: 3651 (2005) [PMID: 15705717]
[3] J Hou *et al., Proc Natl Acad Sci U S A,* 100: 2386 (2003) [PMID: 12606708]
[4] P Kolodny *et al., Curr. Opin. Struct. Biol.,* 16: 393 (2006) [PMID: 16678402]
[5] S Theodoridis & K. Koutroumbas, Pattern Recognition, Second edn. San Diego, U.S.A.: Academic Press; (2003)
[6] AK Jain & RC Bubes, Algorithms for Clustering Data. Englewoods Cliffs, New Jersey, U.S.A.: Prentice Hall; (1988)
[7] AG Murzin *et al., J. Mol. Biol.,* 247: 536 (1995) [PMID: 7723011]
[8] G Wang & R. L. J. Dunbrack, *Bioinformatics,* 19: 1589 (2003) [PMID: 12912846]
[9] O Carugo, *Curr. Protein Pept. Sci.,* 8: 219 (2007) [PMID: 17584118]
[10] O Carugo, *Curr. Bioinformatics,* 1: 75 (2006)
[11] O Carugo & S. Pongor, *Curr. Protein Pept. Sci.,* 3: 441 (2002) [PMID: 12370006]
[12] P Rogen & B Fain, *Proc Natl Acad Sci USA,* 100: 119 (2003) [PMID: 12506205]
[13] O Carugo, *Acta Crystallogr*., B51: 314 (1995)

Edited by P. Kangueane
Citation: Carugo, Bioinformation 4(8): 347-351 (2010)

Supplementary material:

Equation 1: The ratio *R* is

$$R_i = \frac{u_i}{w_i} \quad \rightarrow \quad (1)$$

is then computed and the average value *R* of the *k* R_i values

$$R = \frac{\sum_{i=1}^{k} R_i}{k} \quad \rightarrow \quad (2)$$

Table 1: Cox-Lewis coefficients computed on various data sets. 500 values were computed on each data set.

Dataset	min	max	ave	std
Scop	**1.356**	**1.544**	**1.445**	**0.065**
Scop/1	1.104	1.773	1.369	0.062
Scop/2	1.195	1.561	1.400	0.063
Scop/3	1.103	1.609	1.334	0.060
Scop/4	1.312	1.818	1.559	0.070
Scop/5	0.983	1.281	1.128	0.051
Scop/6	0.934	1.465	1.163	0.052
Scop/7	0.932	1.659	1.294	0.058
Scop/8	0.964	1.285	1.120	0.050
Scop/9	1.013	1.355	1.175	0.053
Scop/10	0.935	1.555	1.258	0.056
Pisces	**1.320**	**1.422**	**1.373**	**0.061**
Pisces/1	1.073	1.691	1.364	0.061
Pisces/2	1.086	1.356	1.215	0.054
Pisces/3	1.117	1.435	1.262	0.057
Pisces/4	1.290	1.978	1.537	0.069
Pisces/5	1.051	1.344	1.193	0.053
Pisces/6	1.152	1.434	1.282	0.057
Pisces/7	1.147	1.497	1.332	0.060
Pisces/8	1.077	1.417	1.193	0.053
Pisces/9	1.133	1.418	1.280	0.057
Pisces/10	1.035	2.120	1.288	0.058