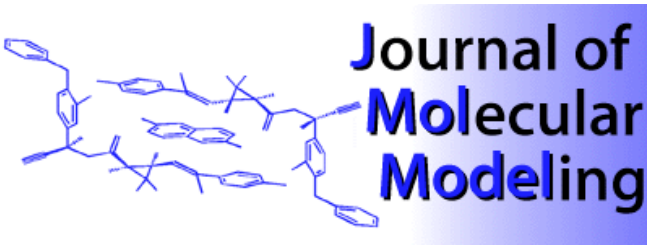
J. Mol. Model. **1998**, *4*, 347 – 354



FULL PAPER

Statistical Analysis of the Loop-Geometry on a Non-Redundant Database of Proteins

Marc A. Martí-Renom, José M. Mas, Patrick Aloy, Enrique Querol, Francesc X. Avilés, and Baldomero Oliva

Institut de Biologia Fonamental and Departament de Bioquímica, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Barcelona, Spain. Tel: +34-3-5812807; Fax: +34-3-5812011. E-mail: baldo@pug.uab.es

Received: 27 July 1998 / Accepted: 18 September 1998 / Published: 24 November 1998

Abstract The conformations of protein loops from a non-redundant set of 347 proteins with less than 25% sequence homology have been studied in order to clarify the topological variation of protein loops. Loops have been classified in five types (α-α, α-β, β-α, β-links and β-hairpins) depending on the secondary structures that they embrace. Four variables have been used to describe the loop geometry (3 angles and the end-to-end distance between the secondary structures embracing the loop). Loops with well defined geometry are identified by means of the internal dependency between the geometrical variables by application of information-entropy theory. From this it has been deduced that loops formed by less than 10 residues show an intrinsic dependency on the geometric variables that defines the motif shape. In this interval the most stable loops are found for short connections owing to the entropic energy analysed.

Keywords Statistics, Loop conformation, Protein modelling

Introduction

Loops constitute an important category of non-regular secondary structures in globular proteins. Because of their variety of forms, loops have evaded a descriptive taxonomy of folds. They are also one of the most difficult structures to delineate for X-ray crystallographers and NMR spectroscopists. Nevertheless, the conformation of loops in proteins has been intensively investigated because of their fundamental and applied interest. Evaluation of the role of the loop conformation on the fold of a protein has become one of the main objectives pursued in this field [1-4].

Correspondence to: B. Oliva

Loops were mainly treated as "random coils", and on this basis energetic studies on the involvement of loops in protein folding were performed. Thomas calculated the internal tension entropic energy involved in the fold of a loop using the simplest available model of randomly joined chains on the basis of a rubber-like elasticity [5]. Meirovitch and Hendrickson did a similar calculation by assuming a Gaussian distribution for the end-to-end extension of a loop [5,6]. The validity of these studies became limited once it was discovered that loops follow certain fixed patterns and cannot be considered as random structures [7-10]. In addition, these studies used a single variable (the end-to-end distance) to define a loop which is a simple but insufficient parameter to characterise it.

In order to define the minimum number and properties of variables to use when classifying loops, we have performed

an statistical analysis on a non-redundant database of 347 protein structures. Information-entropy shows that loops on short connections (even less than 10 residues) do not present random geometry.

Methods

Protein loops have been defined as polypeptide regions between two regular secondary structures (α-helices and β-strands). The α-helices and β-strands of a protein were defined with the program DSSP [11]. There are four types of loops: α-α) loops between two α-helices; α-β) loops between an α-helix and a β-strand; β-α) loops between a β-strand and an α-helix; and β-β) loops between two β-strands, this being β-hairpins or β-links. The total of loops extracted from the Protein Data Bank [12] statistically defines an universe of known protein loops. Here a non-redundant set of loops have been analysed. That is, a set of loops extracted from those proteins of the Protein Data Bank with homology lower than the 25% [13,14].

Geometrical definition

The geometry of the motif comprising the loop is defined by means of the main moments of inertia of the secondary structures flanking the loop [8]. This has been defined by the geometrical co-ordinates of the loop (see Figure 1):

1) D, distance between secondary structures.
2) δ, "hoist angle", angle between *M1* and *D*.
3) θ, "packing angle", angle between *M1* and *M2*.
4) ρ, "meridian angle", angle between *M2* and the perpendicular plane.

For D, the interval ranged from 0 Å to 40 Å partitioned by intervals of 2 Å. For angles δ and θ, they ranged from 0° to 180° partitioned by intervals of 45°. For angle ρ, the interval ranged from 0 to 360° degrees, being this also partitioned by intervals of 45°. This partition represents each loop in a four dimensional space with geometrically independent variables: D, δ, θ and ρ.

Statistical analysis of the protein-loops geometry

For the statistical analysis of the data we have incorporated a function that calculate the correlation between pairs of variables by two different methods: (1) contingency tables using measures of association based on χ^2 [15] and on information entropy [16]; and (2) non-parametric rank-order correlation coefficients, with the Spearman's rank [17] or Kendall's τ [18]. The result of all these methods is commonly represented by a value lying between 0 and 1: Cramer's V, contingency coefficient for measures of χ^2, and measures of the symmetrical uncertainty coefficient based on the information-entropy, equal zero when there is not association and 1 when there is a perfect association. Also a value of significance is given to determine the accuracy of is the correlation. This value is smaller than 0.05 if the correlation is accurate enough, and larger otherwise. We have also studied the results from the

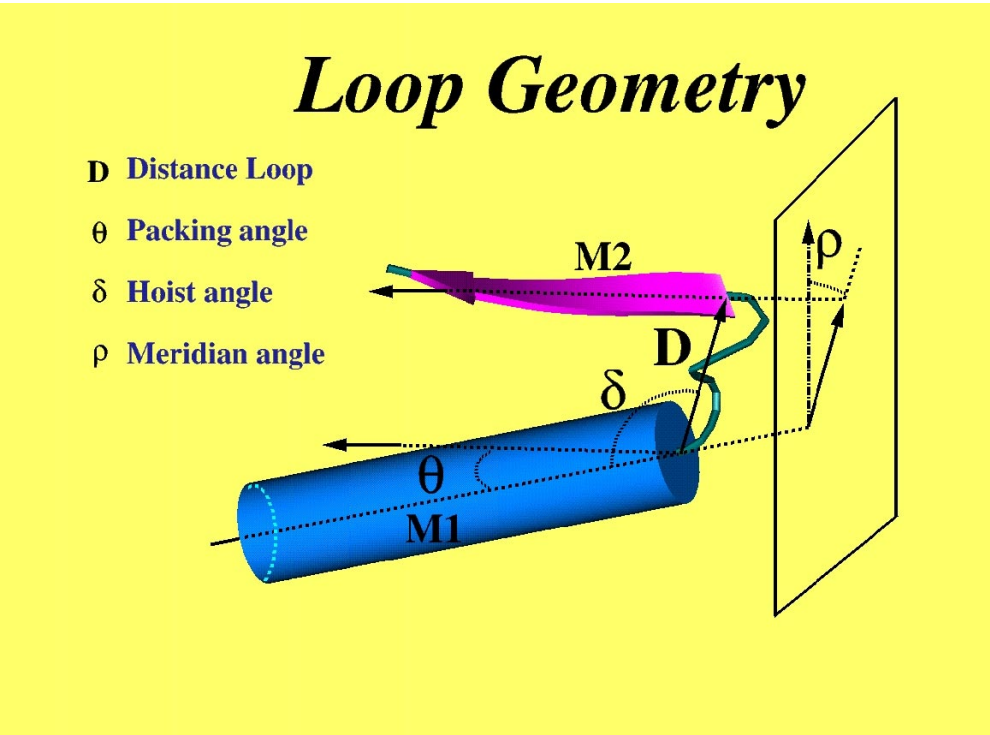


Figure 1 *Definition of the internal co-ordinates used for the geometrical description of a loop motif (i.e. for the α-β motif)*

Spearman and Kendall's analysis and we have represented their significant correlation with similar ranges (smaller than 0.05 for remarkable correlation). The values of rank correlation and Kendall's τ are calculated in order to gain insight into its strength, being positive correlation when Spearman's rank and/or Kendall's τ are positive, and being an anticorrelation when they are negative. The analysis of the non-parametric rank correlation is allowed because the statistical variables are continuous (real and ordered numbers) and also because their interval of existence is large enough. However, for a small number of data it is not possible to rely on the results. From the analysis of the contingency tables, the strength of the correlation may be evaluated. Nevertheless, values of Cramer's V or contingency coefficient lying between both extremes are meaningless and only the symmetrical uncertainty coefficient shows up the intensity of the correlation. The linear correlation of the ranks (Spearman) or Kendall's τ may also show the correlation type and strength. Moreover, for the comparison of the values obtained from the contingency tables it is necessary to define the tables with the same sizes.

In order to be able to recognise the most stable size of a loop, we have used the measure of association based on information-entropy [16]. The entropy of the set of variables in the 4D-space that belongs to an specific loop size is defined as:

$$\text{Entropy} = S = S(D,\delta,\theta,\rho) = \sum_i p_i \ln(p_i) \tag{1}$$

being $p_i = \dfrac{n(D,\delta,\theta,\rho)_i}{N}$;

and $n_{(D,\delta,\theta,\rho)_i}$ the number of loops with the vector (D,δ,θ,ρ) of the 4D-space in the interval "i", identified as a 4-cube of size ΔI=[2Å,45°,45°,45°] on a partition of the vectorial space. For independent co-ordinates, the total entropy is calculated by the addition of each individual entropic term, analogously defined as

$$p_i = \frac{n(\text{variable})}{N}$$

and being the variable either D, δ, θ, or ρ, respectively. We may represent the relation between the real and the independent entropy as:

$$r \equiv \frac{S}{(SD + S\delta + S\theta + S\rho)} \tag{2}$$

r being 1 when the variables are perfectly independent and different otherwise. Hence, r can be a good estimator of the

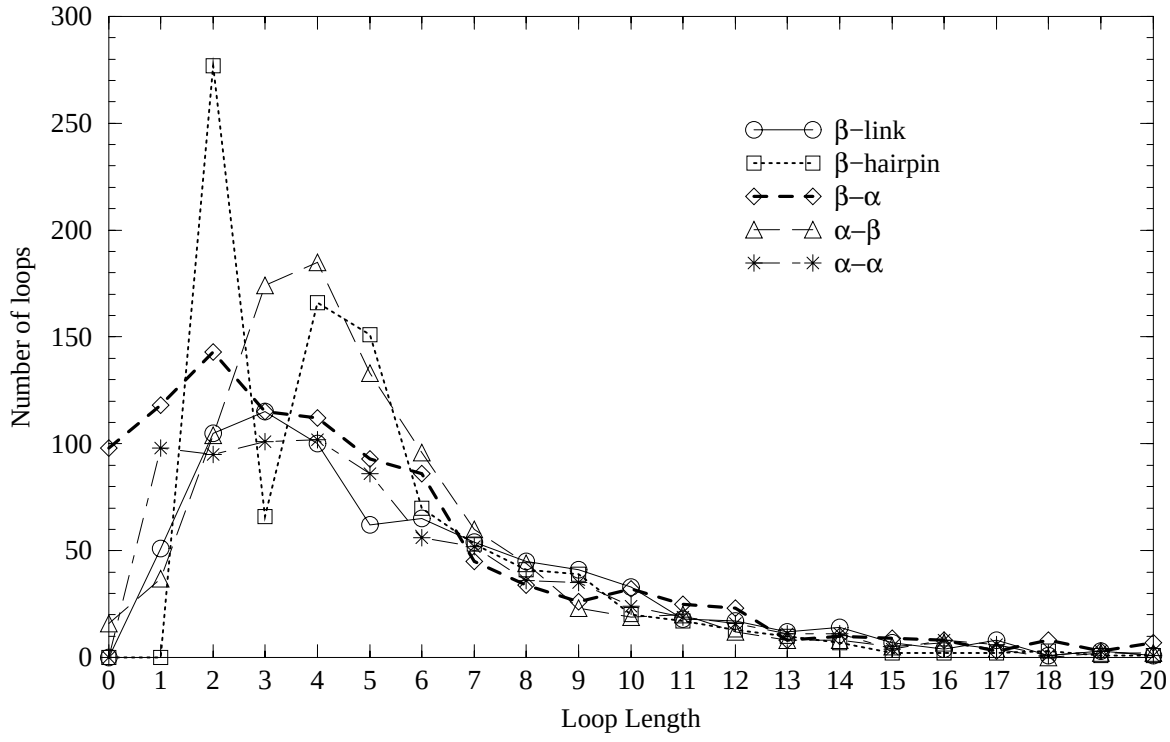


Figure 2 *Representation of the number of loop-motifs versus the number of residues forming the loop: β-hairpins, β-links, α-α motifs, α-β motifs and β-α motifs*

independence of the variables. It is also meaningful the difference calculated as energy, that defines the entropic energy necessary to produce the real correlated set of variables as:

$$E = -T\Delta S = -T\left[S - \left(SD + S\delta + S\theta + S\rho\right)\right] \quad (3)$$

with T=300K. This energy is positive when the set is non random.

Statistical significance

The geometric co-ordinates (D,δ,θ,ρ) have been obtained considering the regular secondary structures surrounding the loop. Any relation between them shows the restrictions arising from the geometrical scaffold of the motif. The problem is reduced to calculate the correlation between the set of values on this 4D-space.

To consider the statistically meaningful results it is necessary to determine first if the number of loops is large enough. This is shown by plotting of the number of loops versus its size for each type of loop (Figure 2). We assume that loops larger than 20 residues are unimportant. We represent the results obtained from the statistical analysis by the measure of the significant value of Spearman's non parametric rank correlation, the significant value of Kendall's τ and the significant value of the contingency tables based on χ^2. We also aim at representing the strength of the correlation. The results obtained for each statistical parameter did show that Cramer's V (CV), symmetrical dependency of entropy (u(x,y)), the absolute value of the Spearman's rank correlation (SR), and the absolute value of Kendall's τ (Kτ) had similar curves. Therefore, we have defined a new parameter formed by the combination of the statistical parameters. We call this parameter the "correlation strength". We require some specific properties for the construction of this new parameter: 1) it must show the existence of correlation, 2) it must distinguish the positive correlation from the anticorrelation, 3) it must be related with the intensity of the correlation, and 4) it should enable us to compare the different types of loops. The strength of correlation is defined as:

$$\text{Strength} \equiv \frac{K_\tau}{\left|K_\tau\right|}\left(\frac{CV + u(x,y) + \left|SR\right| + \left|K_\tau\right|}{4}\right) \cdot \Omega\left(\text{sig}, \text{sig}SR, \text{sig}K_\tau\right) \quad (4)$$

where "sig" is the significance of the correlation by contingency tables, "sigSR" the significance of the correlation by Spearman, "sigKτ" the significance of the correlation by Kendall, and "Ω" is a function defined as:

$$\Omega\left(sig, sigSR, sigK_\tau\right) = \begin{cases} 0 & \text{if sig} \geq 0.05; \text{sig}SR \geq 0.05; \text{sig}K_\tau \geq 0.05 \\ 1 & \text{if sig} < 0.05 \text{ or sig}SR < 0.05 \text{ or sig}K_\tau < 0.05 \end{cases}$$

The desired properties 1, 2 and 3 follow immediately from their definitions. Also property 4 is obtained if the parameters used on the linear combination have this property. This comes out from its definition because the contingency tables defined to calculate the measures of association have the same size for each pair of geometrical variables, independently of the type of loop.

Results and discussion

The results obtained for the correlation strength (Figure 3) show that for any type of loop shorter than 10 residues there are always some correlated geometrical variables. For sizes between 10 and 16 there are always some pair of correlated variables (except for the α-α motif). β-link loops present high correlation in all pair of variables that involve the distance between secondary structures (D). This correlation is observed in loops shorter than 10 residues. However, the correlation between D and the meridian angle (ρ) is not observed in loops larger than 2 residues. A similar behaviour was observed for β-hairpins. In this type of loops the correlations involving the D variable are lower than in β-links. β-α loops (Figure 3c) showed correlation between all pairs of variables for loops shorter than 7 residues indicating that an intrinsic relation between the chosen conformational variables exists. This behaviour is also observed in α-β loops. Finally, Figure 3d shows the correlation strength for loops between α-helices. In this case, the intrinsic relation between variables is lower than for loops involving an α-helix and a β-strand. However, there is significant correlation in loops shorter than 6 residues. The meridian angle (ρ) shows for all cases smaller correlation with the rest of geometrical co-ordinates. The strength of the correlation for the six possible combinations of pairs of geometrical co-ordinates (sizes than 5 residues) is smaller in both α-α motifs and β-hairpins than for the rest of motifs. It is remarkable that all the correlation strengths involving the end-to-end distance of the loop co-ordinates are not null until sizes of about 8 residues. This shows the main relation between the length of the loop and its shape. We conclude that most of the loops with size smaller than 10 residues belong to a determined set of geometry, which involves recurrences of conformation. We describe those motifs where the size ranges in the interval of minimum correlation-strength as those with preference to be constructed with minimal restraints.

We have calculated the entropic energy necessary to produce the correlated set of geometrical co-ordinates to gain insight into the problem of the energetically more stable loop size. We obtain the same conclusion as for the correlation of pairs of co-ordinates: short connections show the smallest entropic energy (Figure 4a). Furthermore, the result of the entropy estimator (Figure 4b) is also in agreement with the result obtained from the comparison of the strength of the correlation. The number of loops shorter than 10 residues with similar geometry is higher than at obtained randomly. It is also shown that the estimator is larger for lengths shorter

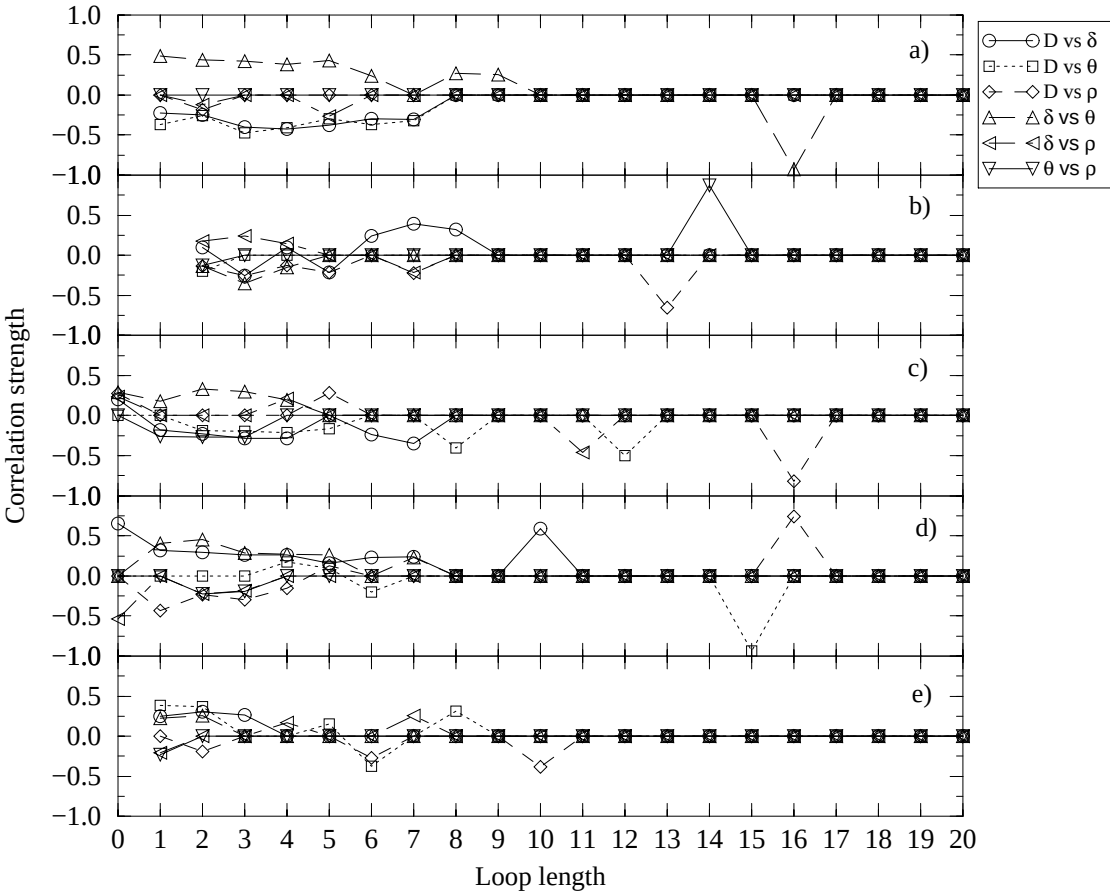


Figure 3 *Correlation strength between geometric co-ordinates: a) β-links ; b) β-hairpins; c) β-α motifs; d) α-β motifs; and e) α-α motifs*

than 6 residues than for lengths between 6 and 20 residues. This is in agreement with the small entropic energy found for short connections (around 3 kJ/mol). Therefore, loops shorter than six residues are produced with the minimal energetic effort without being random. This energy is around 8 kJ/mol for lengths between 6 and 10 residues. In particular, loops of type β-α show a minimum around 4 residues size of about 3 kJ/mol.

Conclusions

The main goal of this work has been to reduce the geometrical parameters of the loop to the minimum number of independent co-ordinates and to analyse the independence of their variables. Its aim has not been to cluster a set of non-homologous loops but to demonstrate that this is possible. Although this has been shown by other authors [19,20], this demonstration was not shown in a previous clustering work using the same motif-geometry definition [8], this being a requisite before tackling the classification of protein loops. The geometry of the loop involves several restrictions on the 4D-space and leads to an additional projection into a subspace. These restrictions have been demonstrated by statistical analysis with the independence of the 4 geometrical co-ordinates. The entropic estimator of independence has demonstrated the intrinsic relation between the geometrical co-ordinates of the 4D space for loops shorter than 10 residues (r<50%). It has also been obtained the energetic expense due to the construction of the restrictions on the 4D space. This energy is involved in the increase of information-order with respect to the chaos of a randomised system. The results have shown a minimum energy (around 3 kJ/mol calculated at 300 K) for short connections (shorter than 6-residues). This energy should not be interpreted in terms of physico-chemical optima but as the optimum size according with the statistics taken from the data. In this sense, this is analogous to the pseudo-energies used on statistic potentials derived from a similar set of non-homologous proteins [21]. The result obtained gives an answer to the results obtained by Rooman *et al*. [22], Fidelis *et al*. [23] and Bystroff and Baker [24], it explains why it is easy to reconstruct and/or predict the conformation of short segments when building a region embedded in a protein structure. Also, we have shown that the use of a database method

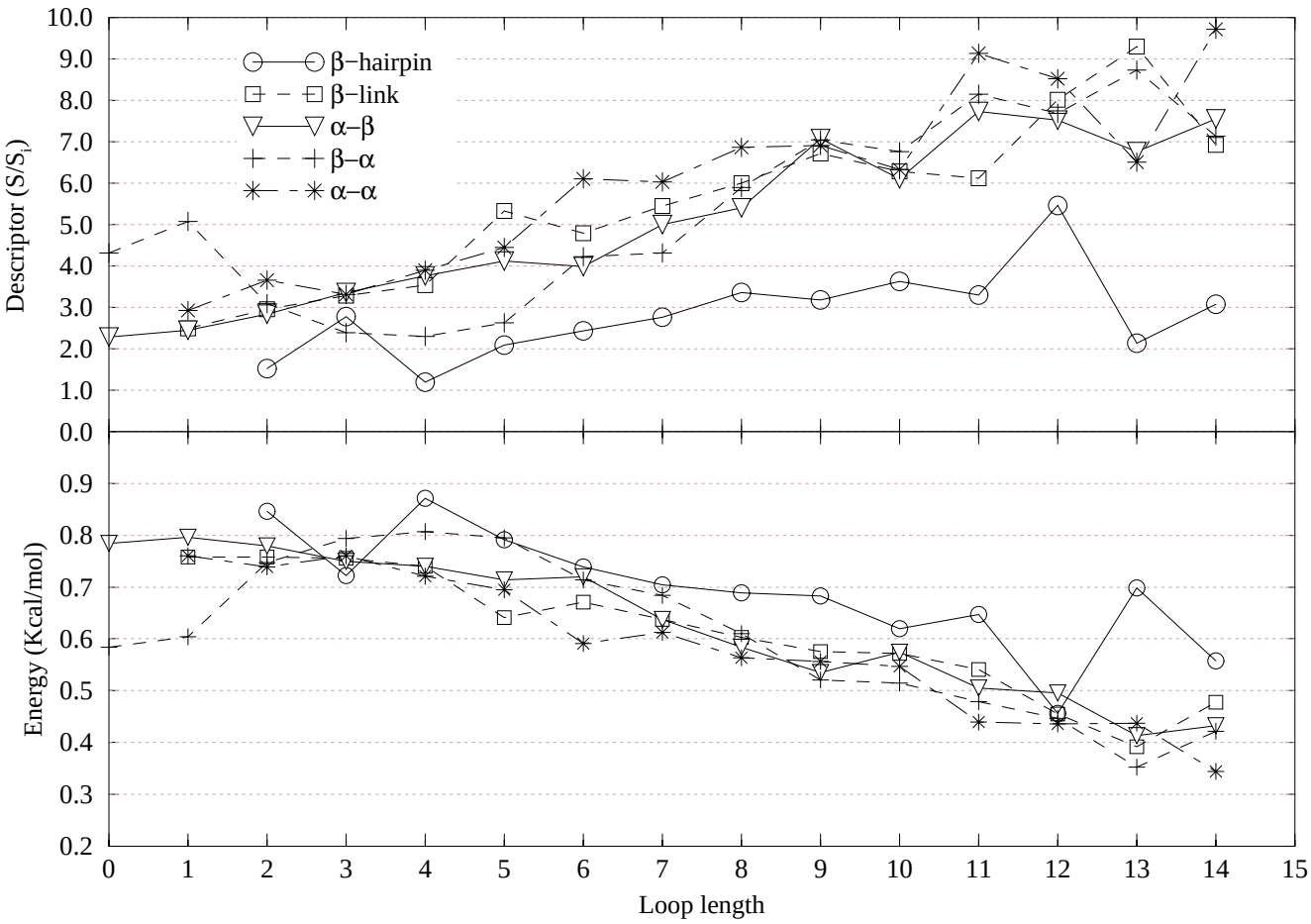


Figure 4 *Information-entropy energy and total correlation represented by the entropic association estimator for the five loop-motif types*

was not effective for comparative modelling in Fidelis *et al*.'s work because the same conformation may cause different geometries, whilst for a given geometry, the number of available conformations for short segments is restricted and non-randomly determined (also shown in ref. 20).

The statistics presented in this work demonstrate that the four variables chosen to describe the loop geometry (D, δ, θ, and ρ) are enough to explain the geometrical clustering of loops described by Oliva and co-workers [8] and used in the classification of antibody CDR3 loops by the same authors [25]. Therefore, the results show that it is possible to obtain clusters of short connections because their geometry is not built randomly. The use of this 4D space to describe a loop geometry could be used in other protein loop classification with a limit on the number of residues involved in the loop. Statistic shows that these four variables can be used in protein loop clustering when the loop is not longer than 10 residues. The conformation of loops longer than 10 residues do not follow a pattern and the clustering leads to non distinguishable classes on a non-homologous database. The statistical approach presented in this work differs from that presented by Oliva *et al*. [8]. We are not clustering loops but analysing the possibilities of obtaining statistically meaningful clusters, and therefore the objectives of the two studies are different. The present work is intended to justify the existence of an inner relation for a given geometry of a loop (not necessarily involved with the conformation of the loop). On the other hand, the aim of the previous work was to cluster the loops of a non-homologous set of proteins according with their geometry and conformation as well as to extract the main inner interactions. Also, most preceding works [7,9,23,26] needed the use of an RMS cut-off to get clusters with similar conformation restricted to two end points involving the specific geometry of the motif. We have shown that for a given motif-geometry and a short number of residues an inner relation can be found that determines the geometry and, therefore, there is less probability to be obtained randomly. This was also shown by van Vlijmen and Karplus [20] providing a direct correlation between the stem residues (those adjacent to the loop) and the loop conformation. In our work we have shown the same relation for the internal co-ordinates that define the motif-geometry instead of using

the stem residues, which improves the results already presented for the clustering method that uses these co-ordinates [8].

The work is mainly a retroactive confirmation of the methodology and finds a new specific insight. First, we raise the need of demonstrations to be done on the cluster analysis of protein loops [7-10,19,26]. All the previous authors found the need of defining restrictions of loops, both related to the conformation and to the motif geometry. A recent method based on the iterative refinement of clusters has also improved the clustering of short segments and cross-validated its predictive capabilities [24], without need of restrictions on the geometry by means of a jack-knife test. However, the work of van Vlijmen and Karplus [20] gave an important insight into the use of this geometry to predict the most likely conformation for a target loop, ranking them according to the van der Waals energies. The demonstration of how statistically meaningful were the clusters was mainly tackled in the works of Ring *et al*. [19], where the existence of recurrences of loop conformations was proved, and van Vlijmen and Karplus [20], where the statistical correlation between loop conformation and the stem residues was shown. Second, we have obtained an energetic point of view to suggest the optimum size of a motif according to the current methods based on statistical potentials. This may be helpful for the *ab initio* protein engineering of a loop segment embedded in the structure of a protein. However, it is important to note that we can not carelessly relate this statistic pseudo-energy with the atomic interactions within loops. In this sense, the work of Tramontano *et al*. [27] and van Vlijmen and Karplus [20] did succeed in tackling the problem of the energetic interactions at the atomic level and illustrated the most feasible reasons to explain these results. Tramontano *et al*. [27] showed that the structural determinants involved interactions with other parts of the protein. They also identified medium-sized loops with characteristic packing interactions and/or main-chain hydrogen bonds involved in the loop conformation, although some of their conclusions were based on the complementary determining regions (CDR) of immunoglobulins (a set of homologous proteins). Therefore, the conclusions of our analysis cannot be straightforward compared with their results because it only takes into account the motif-geometry and uses a set of non-homologous proteins. On the other hand, short and some medium sized loops present main inner interactions within the motif or primarily depend on their sequence, which could explain the statistical optimum size found in our work and relate both, physico-chemical energy and pseudo-energy, and also the use of clusters for the prediction of conformation of short segments [24]

Our result is very important in order to save computational time and efforts trying to classify long loops because it has been proved that this is not possible. Actually, most of the papers presented in the literature about classification of loops have been focused on classification of short connections, whilst for long loops it has not been achieved. In this work we show that this is an impossible task with the present database of proteins. Although Bystroff and Baker [24] have improved the procedure for the prediction of the conformation of short segments, this is not yet available for long segments. Our work presents a similar conclusion: loops of around 4 residues may show specific patterns because they are constructed non-randomly. The energetic terms calculated by means of the statistical potentials further support this conclusion. Finally, this conclusion can be used as a benchmark on the construction of loops by protein engineering according to the classification of short connections and a given loop geometry.

Acknowledgements Authors wish to thank Dr. Núria Romero from the Inorganic Chemistry Department (Autonomous University of Barcelona) for helpful discussions. This work has been supported by grants BIO94-0912002 and BIO95-0848 from the CICYT (Ministerio de Educación, Spain) and by the Centre de Referència de R+D de Biotecnologia de la Generalitat de Catalunya. The support by CESCA and Fundaciò Roviralta is also acknowledged.

References

1. Thomas, D. J. *J. Mol. Biol*. **1990**, *216*, 459-465.
2. Thomas, D. J. *J. Mol. Biol.* **1991**, *222*, 805-817.
3. Carter, C. W. Jr. *Structure* **1995**, *3*, 147-150.
4. Brandy, J. P.; Sharp, K. A. *Curr. Opin. Struct. Biol.* **1997**, *7*, 215-221.
5. Treolar, L. R. G. *The physics of Rubber Elasticity*, Clarendon Press, Oxford, 1975.
6. Meirovitch, H.; Hendrickson, T. F. *Proteins* **1997,** *29*, 127-140.
7. Wintjens, R. T.; Rooman, M. J.; Wodak, S. J. *J. Mol. Biol.* **1996**, *255*, 235-253.
8. Oliva, B.; Bates, P. A.; Querol, E.; Avilés, F. X.; Sternberg, M. J. E. *J. Mol. Biol*. **1997**, *266*, 814-830.
9. Donate, L. E.; Rufino, S. D., Canard, L. H. J.; Blundell, T. L. *J. Mol. Biol*. **1996**, *5*, 2600-2616.
10. Rufino, S. D.; Donate, L. E.; Canard, L. H. J.; Blundell, T. L. *J. Mol. Biol.* **1997,** *267*, 352-367.
11. Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577-2637.
12. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535-542.
13. Hobohm, U.; Scharf, M.; Schneider, R.; Sander, C. *Protein Sci.* **1992**, *1* ,409-417.
14. Hobohm, U.; Sander, C. *Protein Sci.* **1994**, *3*, 522-524.
15. Downie, N. M.; Heath, R. N. *Basic Statistical Methods*, Harper and Row, New York, 1965.
16. Fano, Robert M. *Transmission of information.* New York Wiley and MIT Press, New York, 1961.
17. Norman, H. N.; Hull, C. H. *SPSS Statistical Package for the Social Science.* Mc Graw-Hill. New York, 1975.
18. Lehmann, E. L. *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco, 1975.
19. Ring,C.S.; Kneller,D.G.; Langridge, R.; Cohen, F. *J. Mol. Biol*. **1992**, *224*, 685-699.

20. van Vlijmen, H.W.T; Karplus, M. *J.Mol. Biol.* **1997**, *267*, 975-1001.
21. (a) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859-883. (b) Skolnick, J.; Jaroszewski, L.; Kolinski, A.; Godzik, A. *Protein Sci*. **1997**, *6*, 676-688. (c) Aloy, P.; Moont. G.; Gabb, H.A.; Querol, E.; Avilés, F.X.; Sternberg, M.J.E. *Proteins: Struct. Func. Genet*. **1998**, *in press*.
22. Rooman, M.J.; Rodriguez , J.; Wodak, S.J. *J. Mol. Biol.* **1990**, *213*, 327-336.
23. Fidelis, K.; Stern, P.S.; Bacon, D.; Moult, J. *Protein Eng.* **1994**, *7*, 953-960.
24. Bystrof, C.; Baker, D. *J. Mol. Biol*. **1998**, *281*, 565-577.
25. Oliva, B.; Bates, P.A.; Querol, E.; Avilés, F. X.; Sternberg, M. J. E. *J. Mol. Biol*. **1998**, *279*, 1193-1210.
26. Kwasigroch, J.M.; Chomilier, J.; Mornon, J.P. *J. Mol. Biol.* **1996**, *259*, 855-872.
27. Tramontano, A.; Chothia, C.; Lesk, A.M. *Proteins: Struct. Funct. Genet*. **1990**, *6*, 382-394.