Journal of Biological Physics **20**: 331–334, 1994.

REPETITIVE ELEMENTS OF PROTEIN SEQUENCES AS FOSSILS OF EARLY LIFE

ISTVÁN SIMON
Institute of Enzymology, BRC, Hungarian Academy of Sciences, Budapest, Hungary and
International Centre for Theoretical Physics, Trieste, Italy

AND

SÁNDOR PONGOR
International Centre for Genetic Engineering and Biotechnology, Trieste, Italy

Conformational energy calculation shows that most of the small overlapping segments of polypeptide appear in one of the low energy conformations of the respective oligopeptide in proteins. It was shown that from a properly selected set of overlapping segments of polypeptide the protein structure can be calculated as an ansamble of low energy conformations of the overlapping segments [1]. The number of low energy structures are much smaller than the total number of possible conformations of an oligopeptide, therefore the amino acid sequences of proteins have to be well edited to ensure the possibility of simultaneous low energy structure for most of the overlapping segments [1–4] . To appreciate the level of editing of the amino acid sequence with overlapping low energy segments, one should try to create a row of one hundred letters in which most of the overlapping short segments are real English words. One soon realizes that this is almost impossible unless there is some kind of translational symmetry in the sequence of the letters. In a similar manner, a polypeptide with random amino acid sequence can not adopt a unique, stable structure. For a polypeptide with a sequence of ABCDE... etc., it is not easy to ensure that tripeptides ABC and BCD have low energy conformations in which the dipeptide BC appears in the same conformation. At the same time there has to be a low energy conformation of the tripeptide CDE, which has the same dipeptide CD conformation as one of those BCD's, which have a common dipeptide comformation with the low energy ABC and so on.

Translational symmetry makes the situation much simpler. For example, in a poly-dipeptide ABABAB... etc. if there is a low energy structure for the tripeptide ABA, which shares a conformation of the dipeptide BA with one of the low energy structures of tripeptide BAB, and if this also shares a conformation of the dipeptide AB with a low energy structure of tripeptide ABA, the chain can continue endlessly with all the tripeptides in a low energy conformation.

It is worth mentioning that if in such a simple sequence one residue, A or B, is replaced by residue X at one point of the sequence it influences only 3 tripeptides directly. It might influence the local structure, for example a new bend can be formed only if the three new tripeptides in which the mutation is included have low energy conformation. Applied step by step, this kind of amino acid replacement can result in heteropolymers, similar to the polypeptides of globular proteins in which most of the overlapping short segments are in fact in low energy conformations.

Since it is less likely that evolution would produce a simple symmetrical sequence starting from a more complicated asymmetric one, we can suppose that at least a significant part of the homopolymeric portions appearing in globular proteins are fossils of the early stage of evolution rather than the product of this evolution.

The PIR 34 data-base was analyzed to locate repetitive sequence motifs by using a special tailor-made computer programme (Tusnady, E. G. unpublished). Assuming random sequences, the size of which are comparable to the PIR 34, no long homopolymers are expected. However, in the protein data base we found homopolymers from A, D, E, G, H, L, M, N, P, Q, R, S, and T. At the homo-dipeptide level K and V were added to the list. In homo- tripeptides we also find I and Y. At the homo-tetrapeptide level F and C were added to the the list and W appears in the repeated hexapeptides.

These findings would suggest the that aromatic residues F, W and Y as well as C, are added to the list of amino acids in a later stage of protein evolution. It is also possible that I was added to the list later than L and V, the two amino acids of similar chemical character.

In the foregoing we focused on the appearance of various residues in the repetitive sequence motifs. It is very probable that the location of these motifs in the primary and 3D structure and their biochemical functions can also provide information on the early evolution of proteins. Naturally, homopolymeric or repeating protein sequences can appear as a result of various genetic mechanisms. Some non-globular proteins are especially rich in repeats and one could hypothesize that a certain physicochemical property of the amino acids which constitute the repeat is the important factor, not the structure of the repeat itself. For example, salivary and glue-forming

proteins are especially rich in repeats, and the repeats of the same protein may be of different length in different species. It appears that duplication of short repeats is greatly facilitated in some cases. Such fast duplications result in exact copies of the DNA sequence. In other cases, we see mutations in the DNA sequence but there is conservation at the amino acid level. These are the cases when the protein structure may be the decisive factor. Thirdly, there are proteins in which repeats also mutate at the amino acid level.

We should thus expect that DNA structure and the *in vivo* environment also have be taken into account when studying the evolution of protein repeats. One interesting example is that of proline rich repeats whose abundance is a peculiar phenomenon in the non-globular segments of many protein families [5]. From the structural point of view, one can hypothesize that the turn-forming ability, or the propensity to form proline-helices, are the properties which are required in these segments. One has to note, however, that the repeating segments very rarely have a stable structure. We have to notice that proline is a hydrogen-bond donor. This property provides a more plausible explanation for the occurrence of prolin-rich repeats since many or most of these segments which are present in ligand-binding domains. In addition, one has to note that many repetitive DNA sequences that do not code any protein (such as repeats of viral genomes), can be translated into hypothetical proline-coding protein sequences. The abundance of these DNA-repeats is not well understood; one has to note, however that they posess some peculiar structural features [Pongor, S. unpublished]. Notably, proline is encoded by GC-rich codons, and GC stretches are both rigid [6] and, also exhibit a peculiar conformational behaviour [7]. It thus appears that, at least in the case of this class, there may be a structural preferrence at the DNA-level which coincides with a preferrence at the protein structure level.

References

1. Simon, I., Glasser, L. and Scheraga, H.A. (1991) Calculation of protein conformation as an assembly of stable overlapping segments: Application to bovine pancreatic trypsin inhibitor, *Proc. Natl. Acad. Sci. USA.*, **Vol. no. 88**, pp. 3661–3665
2. Simon, I. (1985) Investigation of Protein Refolding: a Special Feature of Native Structure Responsible for Refolding Ability, *J. Theor. Biol.*, **Vol. no. 113**, pp. 703–710
3. Vondervisztt,F., Matrai, Gy. and Simon, I. (1986) Characteristic sequence residue environment of amino acids in proteins, *Int. J. Pept. Prot. Res.*, **Vol. no. 27**, pp. 483-492
4. Cserzo, M. and Simon, I. (1989) Regularities in the primary structure of proteins, *Int. J. Pept. Prot. Res.*, **Vol. no. 34**, pp. 184-195
5. Williamson, M.P. (1994) The structure and function of proline-rich regions in proteins, *Biochem. J.*, **Vol. no. 297**, pp. 249–260
6. Brukner, I., Sanchez, R., Suck, D. and Pongor, S. (1994) Sequence-dependent bend-

ing propensity of DNA as revealed by DNase I: parameters for trinucleotides, *EMBO Journal* (in Press)
7. Brukner, I., Susic, S., Dlakic, M., Savic, A., and Pongor, S. (1994) Physiological concentration of Mg^{2+} ions induces a strong macroscopic curvature in GGGCCC-containing DNA *J. Mol. Biol.*, **Vol. no. 236**, pp. 26-32

ACKNOWLEDGEMENT

Thanks are due to Ms Suzanne Kerbavcic and Ms Valeria Bevilacqua for their help with the manuscript. This work was performed in the framework of a joint project of the authors supported by the ICGEB/UNIDO (93/232; CRP/HUN93-03). I.S. also acknowledges support from the ICTP and from OTKA Foundation, Hungary (1361 and T12890).