Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Isoelectric points of multi-domain proteins

Oliviero Carugo[1, *]

[1]Department of General Chemistry, University of Pavia, Viale Taramelli 12, I-27100 Pavia, Italy and Department of Biomolecular Structural Chemistry, Max F. Perutz Laboratories, Vienna University, Campus Vienna Biocenter 5, A-1030 Vienna, Austria;
Oliviero Carugo* - E-mail: oliviero.carugo@univie.ac.at; * Corresponding author

received August 22, 2007; revised October 23, 2007; accepted November 09, 2007; published online December 05, 2007

Abstract:
Although the distribution of protein isoelectric points is multi-modal, large proteins show isoelectric points less variable than small proteins and their isoelectric points tend to converge to a unique value, close to the pH of the milieu in which the proteins are functional, as far as the protein dimension increases. This study demonstrates that large proteins, which contain more than a single domain, do have isoelectric points less variable than small proteins, which contains a single domain. However, the distribution of the isoelectric points of the single domains, contained in large proteins, resembles that of small proteins, which contain a single domain. Thus, large proteins can be soluble even if their pI is very close to the pH of the milieu, in which they perform their function, since they can contain several domains, the electrostatic properties of each of which mirror those of small proteins.

Keywords: isoelectric point; domain; pH; protein

Background:
Long ago it was observed that protein isoelectric points are not normally distributed. **[1]** Later, computational analyses showed that *Mycobacterium bovis* protein pIs have a bi-modal distribution **[2]**, with two main groups, one at pI < 7 and the other at pI > 7, and bi-modal distributions were observed in several other bacteria, with peaks centered around pI = 5.5 and pI = 9. **[3]** Further similar analyses confirmed these results in several bacterial and archeal genomes and revealed a tri-modality in eukaryotic genomes, and it was proposed that the third peak is related to the emergence of nuclear proteins in eukaryotes. **[4]** Subsequently, it was shown that most organisms, either archea, bacteria, plants or animals have similar tri-modal pI distributions, with prominent peaks at 6.0, 7.6, and 9.0, with an additional peak at 11.3 in some organisms. **[5]**

Such a multimodal distribution is supposed to be a consequence of the type of amino acids selected by nature to build proteins **[5, 6]** and it might be related to the fact that proteins with pI values different from the pH of the milieu, in which they are active, are charged and thus more soluble in water **[7, 8]** and have also an increased folding stability. **[9]**

Interestingly, it was observed that these multimodal distributions tend to coalesce towards neutral pH for very long proteins, both in real proteomes and in the human proteome simulated by considering the restraints due to either the protein length and/or the amino acidic composition. **[5]** This result was recently confirmed **[10]** and it was also shown that the pI distributions tend to be markedly different in different proteomes, depending on the ecological niche of the organisms, taxonomy, proteome size and subcellular location.

The observation that long proteins can maintain the nearly neutral pI seems to contradict the fact that a better solubility is ensured by non-neutral values. It was hypothesized that such a feature depends essentially on statistical reasons, since long proteins contain a larger number of ionizable residues than small proteins and thus buffer better the fluctuations in their amino acidic composition, though a possible role of biological restraints was not excluded. **[10]**

Given that large proteins tend to be formed by several domains **[11]**, it is also important to compare the pI values of large, multi-domain proteins with the pIs of their individual domains. Here it is shown that single domains have pI distributions similar to small proteins. It can thus be hypothesized that the nearly neutral pI values of long, multi-domain proteins is due to a combination of acidic and basic domains and do not depend only on statistical reasons.

Methodology:
Here the pI distributions of 61104 proteins with 100 to 115 residues, of 128806 proteins with at least 1000 residues, and of 22342 individual domains of these long proteins were determined. Proteins containing about 100 residues are typically constituted by a single domain and proteins longer than 1000 residues are hardly expected to contain a single domain. **[12]**. Amino acidic sequences were found in (and downloaded from) the UniProt database **[13]** by using the SRS browsing tool (www.expasy.ch/srs5/). Sequence redundancy was reduced to 40% identity by using the cd-hit software. **[14]** Domain boundaries were extracted from the SwissProt formatted protein sequence files. The pI values were determined by solving

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

numerically the equations 1 and 2 given in supplementary material.

The pI is by definition the pH value at which *nPLUS = nMINUS*. The values of pKa for the various ionizable groups were taken from Hamaguchi (1992). The following pKa values were considered: for the N-terminus 7.4 and 7.5; for lysine 10.0, 10.1, 10.2, 10.3, and 10.4; for arginine 12.5; for histidine 6.3, 6.4, 6.5, and 6.6; for aspartic acid 3.9 and 4.0; for glutamic acid 4.4 and 4.5; for cysteine 7.5, 8.0, and 8.5; for the C-terminus 3.4, 3.5, 3.6, 3.7, and 3.8. While the pI values clearly depend on the set of pKa values, the distributions of the pI values are independent of the pKa and the same multi-modality was observed for each type of pKa values (data not shown).

Discussion:

The pI distribution of the proteins containing 100-115 residues is shown in Figure 1a. It presents four maxima, one around pI = 5, the second around pI = 6.5, the third around pI = 9.5, and the fourth around pI = 11.5. Although these values differ a bit from those observed in similar studies **[5]**, because of different choices of pKa values, it is verified that pIs tend to be distributed in a tetra-modal manner. It is also verified that large proteins tend to have pI values different from small proteins. In fact, the pI distribution for the proteins containing at least 1000 residues, shown in figure 1b, presents a broad maximum around pI = 6.5 and a prominent shoulder around pI = 9. Notably, pI values larger than 10 are very uncommon for these large proteins. On the contrary, the pI distribution for individual domains contained into the large proteins (Figure 1c) is quite similar to the distribution for small proteins (Figure 1a). In fact, four maxima are observed also in Figure 1c, one around pI = 4.5, one around pI = 6.5, one around pI = 9.5, and the last one around pI = 12. Very similar results were also obtained by examining individual proteomes (*Homo sapiens*, *Arabidopsis taliana*, *Thermotoga maritima*, and *Escherichia coli*) and by determining the pI distributions for distinct subsets of proteins of increasing sizes.

Although the exact values of the pIs can hardly be estimated on the basis of the amino acid sequence alone, since the pKa values of ionizable groups might be significantly affected by the three-dimensional structure **[15]** and since the post translational modifications are neglected (some of them, like phosphorylations or arginine methylations have an obvious and substantial effect on the real pI) **[16]**, it can be concluded that individual domains, included in large proteins, tend to have pI values distributed like those of small, generally single-domain proteins. The pI of large, generally multi-domain proteins is thus the result of a series of contributions from each of their domains. It can be postulated that the overall pI of large proteins are, at least in part, determined by their domain composition. Besides general statistical considerations, which may be used to predict that very large proteins tend to assume nearly neutral pI values **[10]**, it can be hypothesized that large proteins can be soluble even if their pI is very close to the pH of the milieu, in which they perform their function, since they can contain several domains, the electrostatic properties of each of which mirror those of small proteins.

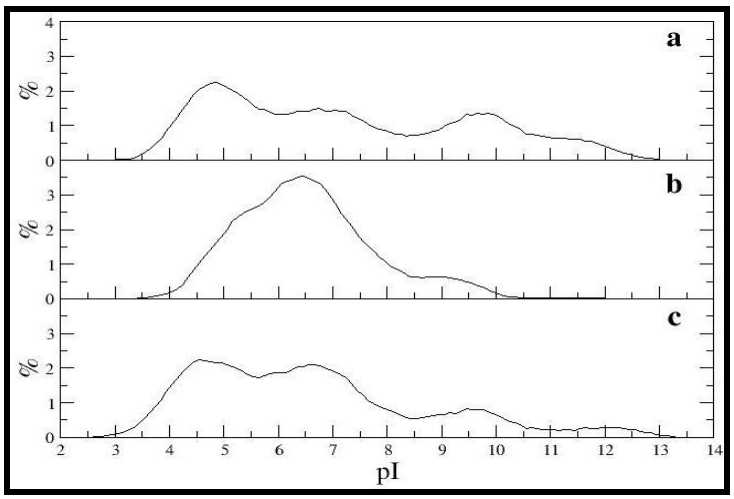

Figure 1: Distribution of the pI values in a set of small proteins (a), long proteins (b), and in a set of domains contained in long proteins (c). see text for details

Conclusion:

The analysis of the pI values of small, single-domain and large, multi-domain proteins shows that the latter ones tend to have pI values close to the pH of the milieu in which they are functional. However, the pI values of the individual domains contained in large proteins closely resemble those of small, single-domain proteins. This clearly demonstrates that the pI value of large, multi-domain proteins results from the sum of the pIs of their individual domains: high pI values of some of the domains are compensated by low pI values of other individual domains

within the same protein chain. As a consequence, large, multi-domain proteins can be soluble and functional despite their overall pI values are close to the neutral pH.

Acknowledgement:
This work was supported by the Austrian GEN-AU project BIN-II. Björn Sjöblom is gratefully acknowledged for helpful discussions.

References:

[01] P. G. Righetti & E. Gianazza, *J. Chromatogr.*, 184: 415 (1980) [PMID: 7438457]
[02] B. L. Urquhart, *et al.*, *Electrophoresis*, 18: 1384 (1997) [PMID: 9298652]
[03] R. A. VanBogelen, *Electrophoresis*, 20: 2149 (1999) [PMID: 10493120]
[04] R. Schwartz, *et al.*, *Genome Res.*, 11: 703 (2001) [PMID: 11337469]
[05] S. Wu, *et al.*, *Proteomics*, 6: 449 (2006) [PMID: 16317776]
[06] G. F. Weiller, *et al.*, *Proteomics*, 4: 943 (2004) [PMID: 15048976]
[07] P. R. Majhi, *et al.*, *Langmuir*, 22: 9150 (2006) [PMID: 17042523]
[08] D. A. Jackson, *et al.*, *J. Cell Sci.*, 90: 365 (1998) [PMID: 3075613]
[09] P. Chan, *et al.*, *Proteomics*, 6: 3494 (2006) [PMID: 16705750]
[10] J. Kiraga, *et al.*, *BMC Genomics*, 8: 163 (2007) [PMID: 17565672]
[11] J. P. Zbilut, *et al.*, *Proteins*, 66: 621 (2007) [PMID: 17154417]
[12] D. Xu & R. Nussinov, *Fold. Des.*, 3: 11 (1998) [PMID: 9502316]
[13] C. H. Wu, *et al.*, *Nucleic Acids Res.*, 34: D187 (2006) [PMID: 16381842]
[14] W. Li & A. Godzik, *Bioinformatics*, 22: 1658 (2006) [PMID: 16731699]
[15] M. R. Salaman & A. R. Williamson, *Biochem. J.*, 122: 93 (1971) [PMID: 5124820]
[16] D. Locke, *et al.*, *FASEB J.*, 20: 1221 (2006) [PMID: 16645047]

Edited by P. Kangueane
Citation: Carugo, Bioinformation 2(3): 101-104 (2007)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Supplementary material

Equation used in this article:

$$\text{nPLUS} = \frac{N_{Lys}[H^+]}{[H^+]+K_{Lys}} + \frac{N_{Arg}[H^+]}{[H^+]+K_{Arg}} + \frac{N_{His}[H^+]}{[H^+]+K_{His}} + \frac{[H^+]}{[H^+]+K_{NT}} \quad \rightarrow \quad (1)$$

$$\text{nMINUS} = \frac{N_{Asp}K_{Asp}}{[H^+]+K_{Asp}} + \frac{N_{Glu}K_{Glu}}{[H^+]+K_{Glu}} + \frac{N_{Cys}K_{Cys}}{[H^+]+K_{Cys}} + \frac{K_{CT}}{[H^+]+K_{CT}} \quad \rightarrow \quad (2)$$

where nPLUS and nMINUS are the positive and negative charges of the protein containing N_x groups X associated with the ionization constant K_x at concentration $[H^+]$.