FOR THE RECORD

Amino acid composition and protein dimension

OLIVIERO CARUGO

Department of General Chemistry, Pavia University, I-27100 Pavia, Italy
Department of Biomolecular Structural Chemistry, Structural and Computational Biology Programme of the Max F. Perutz Laboratories, Vienna University, Campus Vienna Biocenter 5, A-1030 Vienna, Austria

(Received July 22, 2008; Final Revision August 26, 2008; Accepted September 2, 2008)

Abstract

There is indirect evidence that the amino acid composition of proteins depends on their dimension. The amino acid composition of a nonredundant set of about 550,000 proteins was determined and it was observed that, in the range of 50–200 residues, the percentage of occurrence of most of the residue types significantly depends on protein dimension. This result should prove useful in analyzing protein sequences and genomics.

Keywords: amino acids; amino acid composition; protein dimension; protein sequence; UniProt database

Recently, we designed a computational technique for predicting, on the basis of the amino acid composition, if a protein is monomeric or if it forms permanent homo- or hetero-oligomers, together with other polypeptide chains (Carugo 2007). We observed that better predictions were possible by considering the protein dimension, measured by the number of residues (Carugo 2007). In practice, a query of 50 residues was processed by using learning sets of proteins containing less than 100 residues and a query of 150 residues was handled with learning sets of protein containing 100–200 residues, and so forth.

The dependence of the prediction reliability on the protein dimension was not unexpected. At least for relatively small proteins, containing only one globular structural domain, the volume increases more than the solvent-accessible surface if the radius of the globule increases (Rose and Wetlaufer 1977). Enlarging the protein, by adding a residue, implies that the protein core increases more than the protein surface. Since the core is essentially apolar, while the surface is essentially polar, it must be expected that the amino acid composition is not independent of the number of residues.

Despite that, amino acid composition was often used to describe protein sequences and to design predictive algorithms, like, for example, the tendency of proteins to crystallize (Chen et al. 2007), for the protein structural class (Chen et al. 2006), for membrane proteins (Shen and Chou 2005), or for protein contact numbers (Yuan 2005).

However, despite early observations (Fisher 1964; Cornish-Bowden 1983), the dependence of the amino acid composition on the number of residues of proteins was not examined in detail. This is done here, where unexpected trends are described.

Results and Discussion

The percentage of occurrence $pc_{aa,i}$ of the amino acid *aa* in the *i*th protein was computed for each of the 20 types of amino acids in each protein as

$$pc_{aa,i} = 100\frac{n_{aa,i}}{nres_i},$$

where $n_{aa,i}$ and $nres_i$ are the number of residues of type *aa* observed in protein *i* and the total number of residues in protein *i*, respectively. Then, the $pc_{aa,i}$ values were

Reprint requests to: Oliviero Carugo, Department of Biomolecular Structural Chemistry, Structural and Computational Biology Programme, Max F. Perutz Laboratories, Vienna University, Campus Vienna Biocenter 5, A-1030 Vienna, Austria; e-mail: oliviero.carugo@univie.ac.at; fax: 43-1-4277-9522.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.037762.108.

averaged for all protein that contains the same number of residues R, by computing

$$pc_{aa,R} = \frac{\sum_{i=1}^{n} pc_{aa,i}\delta_i}{\sum_{i=1}^{n} \delta_i},$$

where $\delta_i = 1$ if $nres_i = R$, $\delta_i = 0$ if $nres_i \neq R$, and R is the number of residues. The resulting quantity $pc_{aa,R}$ is thus the percentage with which residue aa (aa = A, C, D, . . ., V, W, Y) is observed in proteins containing R residues. All integer values of R from 50 to 200 were examined. On average, there are 3640 protein sequences for each value of R (standard deviation = 28; minimum = 2805 for R = 52; maximum = 4271 for R = 121). Larger values were disregarded since $R = 200$ is close to the natural domain size upper limit (Krishnan et al. 2007). It is the optimal value that allows one to discriminate single-domain proteins from multidomain proteins (as can be verified by considering the domain databases CATH [Orengo et al. 1997] and SCOP [Murzin et al. 1995]; data not shown). We did not want to consider multidomain proteins, the amino acid composition of which reflects the fact that they are constituted by a series of smaller structural domains interconnected by linkers, the amino acid composition of which are expected to be different (Coeytaux and Poupon 2005). R-values smaller than 50 were also disregarded, since extremely small proteins are rather infrequent. A total of 549,616 proteins were analyzed.

Figure 1A shows, for example, how the percentage of alanines ($pc_{A,R}$) varies by increasing R from 50 to 200. Clearly, it is not constant. It increases from ~7% to nearly 9%. The analogous results for cysteine are shown in Figure 1B. The values of $pc_{C,R}$ are larger for small proteins and tend to decrease, nearly linearly, from ~2.3% to ~1.5% in going from 50-residue to 200-residue proteins.

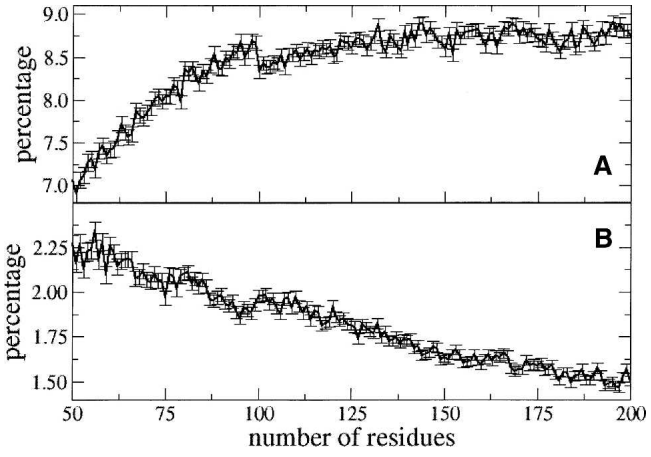

Figure 1. Dependence of $pc_{aa,R}$ on R for alanines (A) and cysteines (B). Standard deviations of the mean values are shown by the vertical bars.

Clearly, the percentage of observation of alanines and cysteines depends on the dimension of the protein, and, also, the type of dependence is different for these two types of residues.

The dependence of $pc_{aa,R}$ on R for all the 20 types of residues is summarized in Figure 2. It clearly appears that different types of amino acids show different trends. Moreover, it is rather surprising that these trends are not very serrated, with large oscillations from one R-value to the next. On the contrary, they delineate quite well continuous curves.

In order to get a quantitative estimation of the statistical significance of the data shown in Figure 2, we compared each $pc_{aa,N}$ value for $50 \leq N \leq 199$ with the $pc_{aa,R}$ value for $R = 200$, through a t-test like

$$t_N = \frac{\left|pc_{aa,N} - pc_{aa,200}\right|}{\sqrt{\sigma^2_{aa,N} + \sigma^2_{aa,200}}},$$

where $50 \leq N \leq 199$ is the number of residues, $\sigma_{aa,N}$ is the standard deviation of the mean value of $pc_{aa,N}$, and $\sigma_{aa,200}$ is the standard deviation of the mean value of $pc_{aa,200}$. For example, $pc_{aa,200}$ for alanine is equal to 8.74 (standard deviation = 0.08) and $pc_{aa,50}$ is equal to 7.06 (standard deviation = 0.10); therefore, t_{50} is equal to 13.12, indicating that the values of $pc_{aa,50}$ and of $pc_{aa,200}$ are significantly different. The t_N values of all the 20 types of amino acids are shown in Figure 3. It appears that in the large majority of the cases, they are very large and prove that the amino acid composition really depends on the protein dimension. Exceptions are tryptophan, glutamine, and threonine, where the t-values are very close to zero for the entire range of protein dimensions. For all the other residues, there is at least one region, along the x-axis, where the t-values are extremely large. For example, they are much larger than zero for protein shorter than 120 residues in the case of alanine, or they are much higher than zero in the case of tyrosine for protein containing 75–150 residues.

Obviously, if the percentage of some residues decreases, the percentage of other residues must increase. However, different trends are observed for different residues. For some amino acids (A, D, E, G, P, and V) the percentage of occurrence tends to increase with the protein dimension until a plateau is reached, where the percentage does not increase any more. For other residues (C, F, H, I, K, M, N, and S), on the contrary, the percentage tends to decrease if the protein dimension

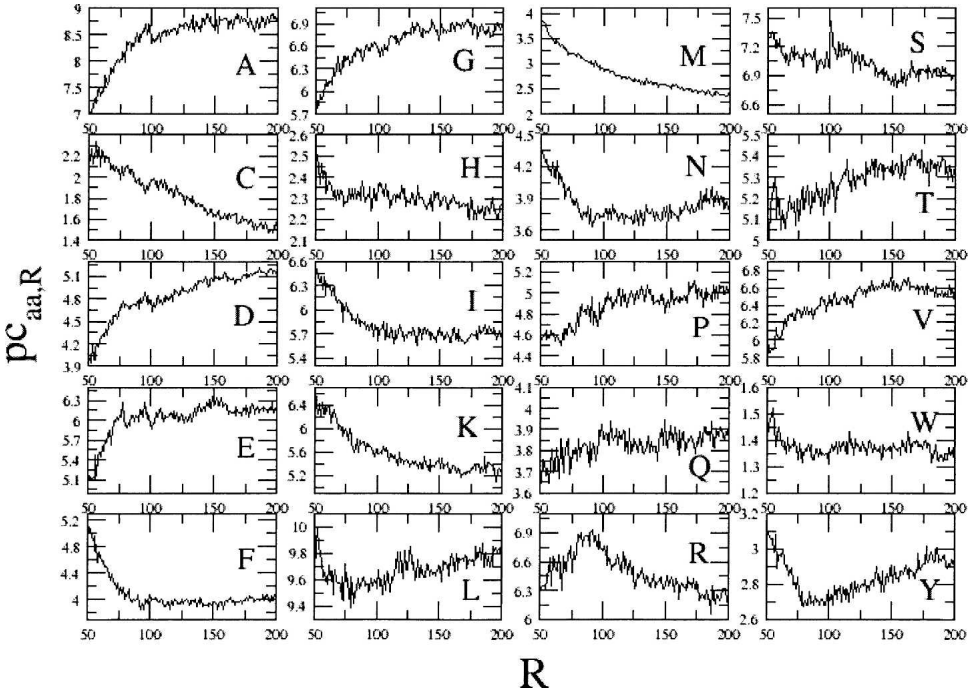

Figure 2. Dependence of $pc_{aa,R}$ on R for all the 20 types of amino acids. Standard deviations of the mean are not shown for clarity. Percentages are on the y-axis, and the number of residues are on the x-axis.

increases. Two residues (L and Y) show higher percentages for small and large proteins, with a minimum for middle-sized proteins. One residue, R, on the contrary, is observed more frequently in middle-sized proteins.

Some of the trends shown in Figure 2 can be understood on the basis of simple considerations. For example, the fact that cysteines are more commonly observed in small protein might depend on the fact that these

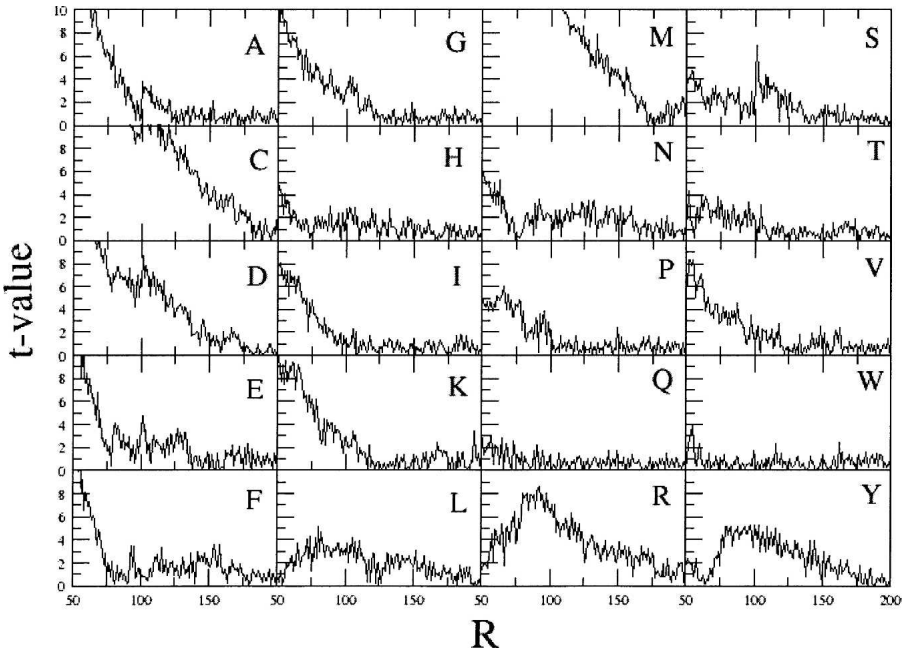

Figure 3. T-values computed by comparing the percentage of observation of a residue in proteins of different dimensions with that observed in proteins containing 200 residues. T-values are on the y-axis, and the number of residues are on the x-axis.

molecules are too small to form a hydrophobic core and often require disulfide bonds to stabilize their native fold (Carugo et al. 2001). Moreover, the fact that large aromatic amino acids (like phenylalanine) are frequently observed in small protein, more than small aliphatic residues (like alanine), might depend on the fact that the small hydrophobic core of small proteins is better stabilized by aromatic–aromatic interactions, like the parallel stacking, which are stronger than van der Waals interactions between aliphatic groups (Marsili et al. 2008).

Other trends of Figure 2 are, on the contrary, absolutely unexpected. For example, lysine and arginine—both positively charged residues—show different behaviors. While the first is more frequent in very small proteins, less frequent around $R = 80$, and more frequent if R increases, the frequency of arginine has a maximum around $R = 80$.

In order to verify that these results are not biased by some unexpected feature of the data present in the uniref50 fasta file, from which the protein sequences were taken, we randomly divided it into 10 subsets of 55,000 entries each and repeated all the computations 10 times, by using separately each subset. In this way, we expect to observe different dependencies of $pc_{aa,R}$ on R for different subsets of proteins if the uniref50 data set is biased. As an example, Figure 4 shows the dependence of $pc_{aa,R}$ on the number of residues for methionine. The 10 curves, each obtained by examining one of the 10 subsets, are clearly very similar and superposed to each other.

In order to monitor quantitatively the differences among different trends, we computed the Pearson correlation coefficients between all pairs of curves $pc_{aa,R}$ versus R, obtained on the 10 different subsets of data. This implies that for each of the 20 types of amino acids, 45 values of the correlation coefficient were computed.

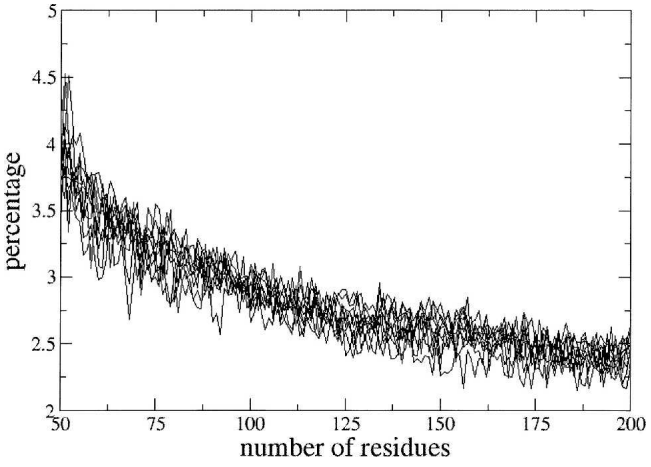

Figure 4. Dependence of $pc_{aa,R}$ on R for methionine. Ten very similar curves are shown, each obtained by examining a subset of the data.

They are actually very similar, ranging, on average, from 0.6 to 0.9, and smaller values are observed only for the residue types, the percentage of which is substantially independent of the protein dimension, demonstrating that the dependence of $pc_{aa,R}$ on R is actually unbiased and genuine.

The percentage of occurrence of the amino acids in proteins depends, at least for some of the residues, on the protein dimension. This is not really unexpected, though it cannot be easily understood. Further studies seem to be necessary. It might be interesting to examine if the dependence of the amino acid composition on the protein dimension is different for different living species/phyla or for different types of proteins (enzymes, cytoskeleton components, metal storage systems, etc.). It must, eventually, be mentioned that these results might be extremely helpful in improving gene-finding algorithms, by restraining the gene triplet composition variance.

Methods

Protein sequences were taken from the UniProt database (The UniProt Consortium 2007), by downloading the uniref50 fasta file, which does not contain pairs of proteins with sequence identity >50%. Sequences containing residues unresolved or different from the 20 types of natural amino acids were ignored. About 550,000 sequences were retained, and their amino acid composition was computed with locally written computer programs.

Acknowledgments

This work was financially supported by the Austrian GEN-AU funding agency (BIN-II networks) and by the NAR funds of the Pavia University. S. Kumar, S. Kirillova, B. Sjoeblom, and K. Djinovic are gratefully acknowledged for helpful discussions.

References

Carugo, O. 2007. A structural proteomics filter: Prediction of the quaternary structural type of hetero-oligomeric proteins on the basis of their sequences. *J. Appl. Crystallogr.* **40:** 986–989.

Carugo, O., Lu, S., Luo, J., Gu, X., Liang, S., Strobl, S., and Pongor, S. 2001. Structural analysis of the free and enzyme-bound amaranth α-amylase inhibitor: Classification within the knottin fold superfamily and analysis of the functional flexibility. *Protein Eng.* **14:** 639–646.

Chen, C., Tian, Y.X., Zou, X.Y., Pai, P.X., and Mo, J.Y. 2006. Using pseudo-amino acid composition and support vector machine to predict protein structural class. *J. Theor. Biol.* **243:** 444–448.

Chen, K., Kurgan, L., and Rahbari, M. 2007. Prediction of protein crystallization using collocation of amino acid pairs. *Biochem. Biophys. Res. Commun.* **355:** 764–769.

Coeytaux, K. and Poupon, A. 2005. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics* **21:** 1891–1900.

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

Fisher, H.F. 1964. A limiting law relating the size and shape of protein molecules to their composition. *Proc. Natl. Acad. Sci.* **51:** 1285–1291.

Krishnan, A., Giuliani, A., Zbilut, J.P., and Tomita, M. 2007. Network scaling invariants help to elucidate basic topological principles of proteins. *J. Proteome Res.* **9:** 3924–3934.

Marsili, S., Chelli, R., Schettino, V., and Procacci, P. 2008. Thermodynamics of stacking interactions in proteins. *Phys. Chem. Chem. Phys.* **10:** 2573–2585.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of protein database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH—a hierarchical classification of protein domain structures. *Structure* **5:** 1093–1108.

Rose, G.D. and Wetlaufer, D.B. 1977. The number of turns in globular proteins. *Nature* **268:** 769–770.

Shen, H. and Chou, K.-C. 2005. Using optimized evidence-theoretic K-nearest classifier and pseudo-amino acid composition to predict membrane protein types. *Biochem. Biophys. Res. Commun.* **334:** 288–292.

The UniProt Consortium. 2007. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **35:** D193–D197.

Yuan, Z. 2005. Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics* **6:** 248. doi: 10.1186/1471-2105-6-248.