 DOI: 10.17344/acsi.2015.1538 *Acta Chim. Slov.* **2015,** *62,* 574–581

Scientific paper

Amino Acid Correlation Functions in Protein Structures

Klemen Kržišnik and Tomaz Urbic*

University of Ljubljana, Faculty of Chemistry and Chemical Technology, Chair of Physical Chemistry, Večna pot 113, SI-1000 Ljubljana, Slovenia

** Corresponding author: E-mail: tomaz.urbic@fkkt.uni-lj.si*

Received: 17-03-2015

Dedicated to prof. Jože Koller on the occasion of his 70th birthday.

Abstract

Understanding the spatial folding of proteins from their amino acid sequences has an enormous potential in contemporary life sciences. The ability to predict secondary and tertiary structures from primary ones through the use of computers will enable a much faster and more efficient discovery of organic substances with therapeutic or otherwise bioactive potential, largely eliminating the need for synthesis and testing of large numbers of organic substances for physiological effects. Our manuscript presents an application of correlation function analysis, usually used to describe properties of liquids, to protein structures in order to elucidate statistically favored distances among amino acids. Pairwise distribution functions were calculated between C-alpha atoms of 20 amino acids in a large ensemble of Protein Data Bank structures. The correlation functions show characteristic distances in amino acid interactions. Different propensities for forming various secondary structure elements among all 210 possible amino acid pairs have been visualized and some have been interpreted. Notably, we found helices to be surprisingly common among certain pairs.

Keywords: Protein structure, radial distribution function, amino acids, correlation function

1. Introduction

Proteins are the basic building blocks of life. Chemically they are polymers made of 20 proteinogenic amino acids, some of which can be further modified post synthesis. Understanding the mechanisms of protein folding is of central importance in structural biology. The main problem of protein folding is the determination of a protein's native structure based on its amino acid sequence.[1] In cellular and other physiological solutions proteins assume a well-defined three dimensional structure, presumed to be of the lowest free energy state.[2] The structure is described on four levels. Primary structure is a one-dimensional string of amino acids, conventionally listed from N-terminal to C-terminal end.[3] Secondary structure signifies well defined local three dimensional motifs which commonly repeat themselves, most often alpha helices and beta sheets.[3] These motifs are mainly a consequence of hydrogen bond forming between amide groups of the peptide backbone; however, there are some interesting new findings which suggest that alpha helix formation is also aided by side-chain interactions in a large degree.[4,5] Tertiary structure is a list of spatial coordinates of every atom in the protein in its native fold, as precise as can be deduced from an experimental technique, usually x-ray crystallography or two-dimensional NMR spectroscopy. Quaternary structure represents an arrangement of multiple peptide chains joined by non-peptide bonds in a functional protein. Notably, not all proteins have quaternary structure, i.e. they are not composed of multiple polymeric chains.[3] The solution to the problem of protein folding is of outstanding value in molecular biophysics and biopharmacy, as the design of a new drug or a new vaccine will be increasingly dependent on our ability to construct molecular structures with very specific binding affinities. Starting from a linear sequence of amino acids (primary structure), a relatively small protein is believed to adopt its native conformation (of minimum free energy) through the interplay of intermolecular forces and thermal energy k_BT (k_B being Boltzmann's constant and T being absolute temperature).[1] The problem of protein folding has been studied with rather different approaches. On one hand, Ising-like models allow us to enumerate exhaustively all confor-

mations.[6] These models can be enriched with pairwise contact potentials.[7] On the other hand, structure prediction using atom-based force field molecular dynamics simulations requires a vast computational effort.[8–11] It would be thus desirable to combine the best of both approaches, i.e. simplicity and accuracy.

Knowledge-based potentials are commonly used as effective free energies or effective potentials to parametrize coarse-grained protein models. These types of potentials are obtained from databases of known protein structures and are successfully used in some of the best known protein structure prediction and fold modeling algorithms.[12–15] Nevertheless, the ability of these methods to predict existing protein structures and model novel ones is limited. The main reason is due to inaccuracies in their energy parametrization. This necessitates investigation into which assumptions used in these potentials are responsible for these limitations, and to what extent. A wide variety of knowledge-based potentials has been introduced, differing in levels of geometric resolution, in terms of contribution to the potential energy, in procedures of relating energies to the observed frequencies, in levels of applicability(some can be used for all proteins in general while some for only a specific protein family), and in their intended purposes, ranging from native fold recognition to protein stability and dynamics simulations,[16–22] including study of protein denaturation in dependence to temperature and pressure.[28]

Here, we calculated radial correlation functions between the 20 naturally occurring amino acids within a large ensemble of Protein Data Bank structures. Pairwise distribution functions have proven themselves to be a useful tool in the theory of liquids and we show they can also be of help in elucidating amino acid interactions. In the latter, they are somewhat curbed by the non-isotropic environment of a protein and a large border problem. Both of these arise from the fact that the radial pairwise distribution function is a tool meant originally for research of amorphous materials, mainly liquids, which are isotropic in nature.[23] As the inside of a native protein much more closely resembles a crystal than a solution, we can only assume that in a large number of structures, anisotropic effects cancel each each other out. However, this likely applies only up to a point, as structural motifs are well known to occur in similar instances throughout the PDB database.[24,25] Further, while in a system such as a solution of B in A, it can easily be assumed that every particle B will be surrounded on all sides with solvent particles A, we have amino acids that themselves span a significant portion of the total length in protein systems, and can have no residue-residue interactions on the outside borders of the system. Also, we do not have a system of two particles in a protein, but one of twenty – meaning that, while for each function we assume there are only AA1-AA2 interactions present, others nearby affect their spatial densities as well. Therefore, a next step of the presented analysis could be an application of the liquid theory's integral equation to extract effective statistical potentials between amino acids,[26,27] which ideally represent only the interactions between the two amino acids in question.

This paper consists of four sections. In Section II, the methods are outlined, with results and discussion following in the Section III. A short conclusion in Section IV completes the paper.

2. Methods

The radial distribution function or pair correlation function $g_{ij}(r)$ shows how relative density of particles *j* varies as a function of distance from a reference particle *i*. If average density of particles *j* is known

$$\rho_j = \frac{N_j}{V} \tag{1}$$

where *V* is volume of system and N_j number of particles *j* in the system then local density of particles *j* from particle *i* is

$$\rho_{ij}(r) = \rho_j g_{ij}(r) \tag{2}$$

Of course all these assumptions hold in an isotropic and homogenous system. In our case we assumed this and that interaction between amino acids is pairwise. In our analysis, we calculated distances between C-alpha atoms for all amino acids of 63890 protein structures from Protein Data Bank and we obtained a number of pairs $N_{ij}(r)$ as a function of distance. We obtained local density as an average over all particles *i* divided by volume of a spherical shell

$$\rho_{ij}(r) = \frac{N_{ij}(r)}{N_i 4\pi r^2 dr} \tag{3}$$

Where *dr* is the step at which distances were sampled. As r grows large enough, affinity between particles is reduced and local density approaches averaged global density

$$\rho_{ij}(r) \to \rho_j \tag{4}$$

In our case we noticed that local density becomes flat at distances higher than 12 Å, so we determined averaged density of particles *j* as average of local densities between distances 12 and 15 Å. When local and averaged bulk densities were known, pair distribution function was calculated as

$$g_{ij}(r) = \frac{\rho_{ij}(r)}{\rho_j} \tag{5}$$

3. Results and Discussion

In protein structures there are 20 different amino acids. This translates into 210 pair correlation functions for all their possible pairs. Although individual pairs have hardly any truly unique peaks, certain amino acids do have a preference to form definite peaks which correspond to secondary structural motifs. Due to a limitation of available space, only a few of the most interesting structures are presented and described in detail, while the rest are available as supplementary material. Those presented were chosen according to uniqueness of their shape, which means highly expressed peaks that are rare in most correlation pairs, or an unusual ratio among peak surfaces. As the research paper outlines, shape differences correspond to different inclinations of the pairs to forming various se-

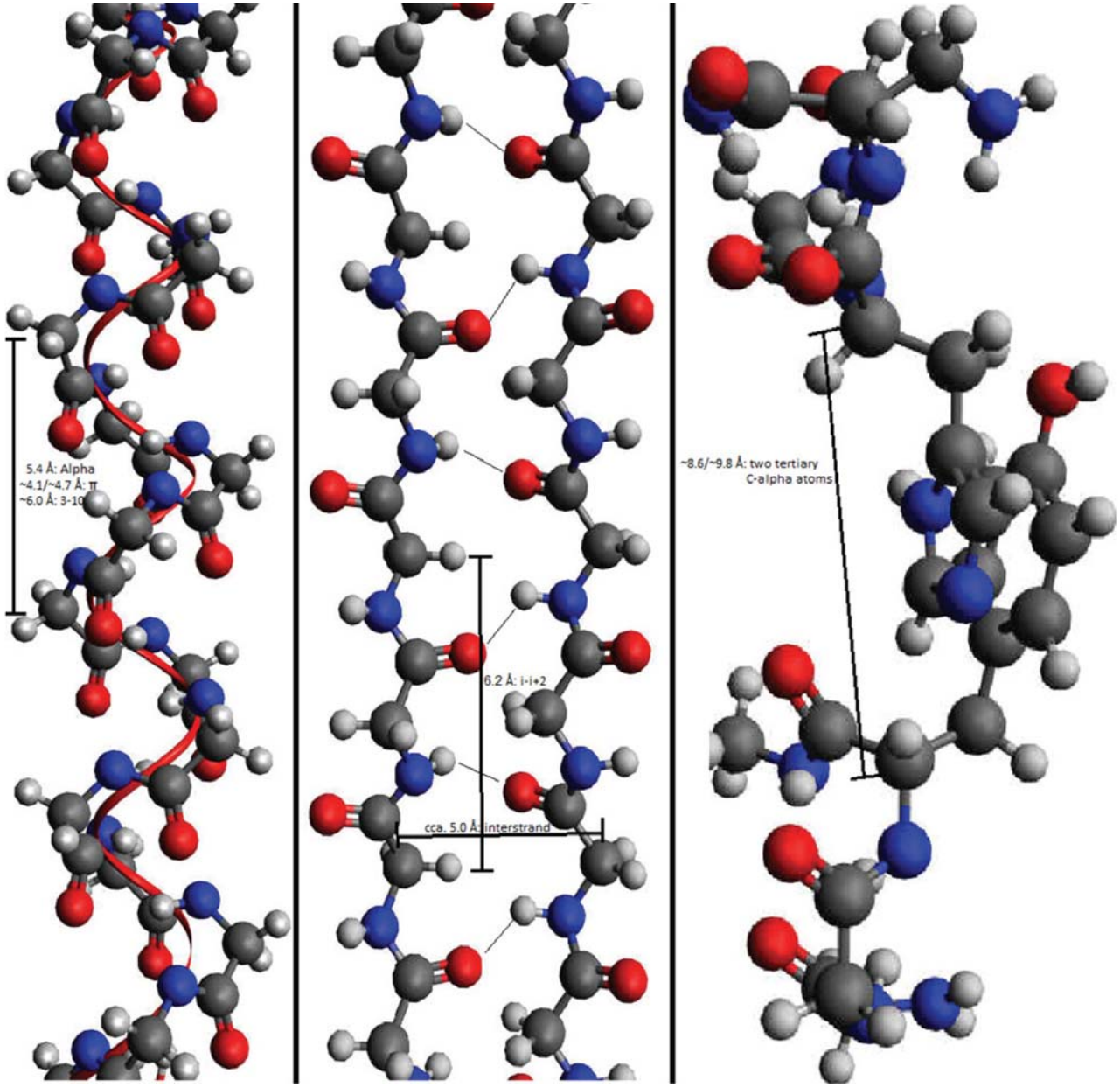

Figure1: Left – a polyglycine helical peptide with the distance between two C-alpha atoms likely responsible for , alpha and 3–10 peaks at around 4.1–4.7 Å, 5.4 Å, and 6.0 Å in our results, respectively, marked by a black line. Note that the molecular structure shown is actually an alpha helix – the other two helices have a bit different shape, besides a different pitch length. Middle – two polyglycine beta strands forming an antiparallel beta sheet, with hydrogen bonds marked by interrupted thin black lines. The horizontal black line highlights an example of an interstrand C-alpha-C-alpha distance likely responsible for peaks at around 5.1 Å. The vertical black line marks an example of an intrastrand i-i+2 C-alpha-C-alpha distance likely responsible for peaks at around 6.2 Å. Right – an example of two peptides related by a tertiary interaction between an imidazole group and a p-hydroxyphenyl group of a histidine and a tyrosine residue, respectively. It is an example of our structural interpretations for peaks at around 8.6 Å and 9.8 Å. The black line shows the distance between C-alpha atoms of the two residues.

Table 1. A list of typical peaks. Prefix ~ means that the exact distance of the peak may vary slightly from pair to pair.

Distance in Å	Structural interpretation
~4.3	Pitch of π helices
~4.7	Pitch of Symbol helices
5.1	Hydrogen bonded beta strands
5.4	Alpha helices
6.0	Pitch of 3–10 helices
6.2	Distance between i-th and i+2-th aa in beta sheets
~7.1	Shorter range tertiary interactions, or maybe a longer pitched coil
~8.6	Tertiary side chain interactions
~9.8	Tertiary side chain interactions

condary structure motifs. First we present a list of typical peaks and our interpretations of their structural meanings in Table 1 and in Figure 1 structural examples.

In Figure 2 we present pair correlation function for two alanine molecules in the entire distance range. A high peak at 3.8 Å can be seen from the figure, which corresponds to peptide bond distance and is present in all distribution functions. This is the distance between two successive C-alpha atoms in the peptide backbone. In all other histogram figures, the region before 4 Å is omitted from the rest of the function for clarity. The alanine-alanine curve has some interesting features. There are the highly expressed peaks at 5.1 Å and 6.2 Å. The peak at 6.2 Å is related to beta sheets, namely to the distance between C-alpha atoms of an i-th and i+2-th amino acids in one strand of the sheet. The typical distance between two H-bonded amino acids among two neighboring strands is around 5 Å, but this peak is usually hidden among the larger alpha helix curve. It is prudent to conclude that alanine is heavily presented in beta sheets because of the two clearly expressed peaks. Beta peak at 5.1 Å is actually likely to be present in the majority of histograms, but only in pairs markedly preferential for beta sheets can it define itself from the larger peak nearby at 5.4 Å, which corresponds to an alpha helix, the most often encountered secondary structure element. This is the distance of two C-alpha atoms between an i-th and an i+4-th amino acids in the helix, the amide groups of which are H-bonded one to another, nearly parallel with the length of the helix. Next two peaks can be found at 8.6 Å and 9.9 Å. Considering their distance, we speculate they are related to side chain interactions.

Next we present the correlation functions between alanine and several other amino acids – glycine, valine and histidine – in Figure 3 . Alanine-glycine histogram is one of two small amino acids, one hydrophobic and one ambivalent. The pair forms all the more typical structural elements, as shown by the alpha helix peak at 5.4 Å and beta peak at 6.2 Å, as well as tertiary interactions. A bit stronger peak at 7.1 Å stands out, perhaps indicating shorter hydrophobic side-chain interactions, which corresponds to smaller side chains of the two amino acids. As next on Figure 3 we present alanine-histidine correlation. The curve has an interesting shape in the beta structure area. It actually peaks at 6.0 Å, instead of the usual obvious 6.2 Å peak, and is much flatter, particularly when compared to the ala-ala curve. Our interpretation is that the peak is shared with the above averagely expressed 3-10 helices. Tertiary interactions are also quite strong, with well-defined peaks at 8.6 Å and 9.8 Å. The final correlation we plotted on Figure 3 was a distribution function for the alanine-valine pair. It has a noticeable peak at 4.3 Å, which is a bit rare generally and also the only pair on Figure 3 sporting this feature. We believe that it likely belongs to closer-pitched π helices. Before the ubiquitous alpha helix peak at 5.4 Å, there is another smaller one at 5.1 Å. For this one we believe it belongs to interstrand distances of beta sheets, indicating a higher-than-usual preference of the pair for this secondary structure type. The next peculiarity is the split peak with creases at 6.1 Å and 6.3 Å, in place of the typical sharply defined beta sheets peak at 6.2 Å. We consider two possibilities for this: it could be either a preference for beta structure at more bent sheets, or it could be related to tighter 3–10 helices.

We continued with a correlation between two glycine molecules and plotted the peculiar result of the smallest self-paired amino acid in Figure 4, amidst the pairs glycine-tryptophane and glycine-alanine. Particularly, peaks at 4.2 Å and 7.1 Å in the gly-gly curve both stand out starkly from the other two curves, indicating a likeliness for involvement of glycine in atypical π helices and short-range tertiary interactions. Beta structure peak at 6.2 Å is expressed a bit below average, indicating glycine's lesser participation in beta sheets, at least as a pair in close proximity. Peaks at 8.6 Å and 9.8 Å are nearly absent – particularly the latter one, confirming our hypothesis that these peaks are related to sidechain interactions, since glycine does not have a side chain. Also on Figure 4, glycine-tryptophane radial distribution function is presented, pairing the smallest amino acid with one of the largest. Glycine has a large freedom of rotation of φ and ψ angles, while tryptophane is hydrophobic due to the indole group, and more rotation-encumbered. The forming of alpha and beta secondary structure is clearly defined, as well as tertiary structure at region from 8.6 Å to approximately 10.0 Å. More unique is the lesser peak at 6.0 Å, which we believe is likely to be related to formation of rarer 3–10 helices, which have a pitch height of around 6 Å, showing a greater than usual tendency of the pair to forming this type of structure. While the peak at 6.2 Å clearly belongs to beta sheets, the indication of a peak at 6.6 Å is more of a question.. It is, however, poorly defined, and so it is our opinion that we cannot assign a particular secondary structure to it. An obvious curiosity of the curve is the peak at 7.7 Å, clearly not present in the other two glycine distribution functions on Figure 4. It could repre-

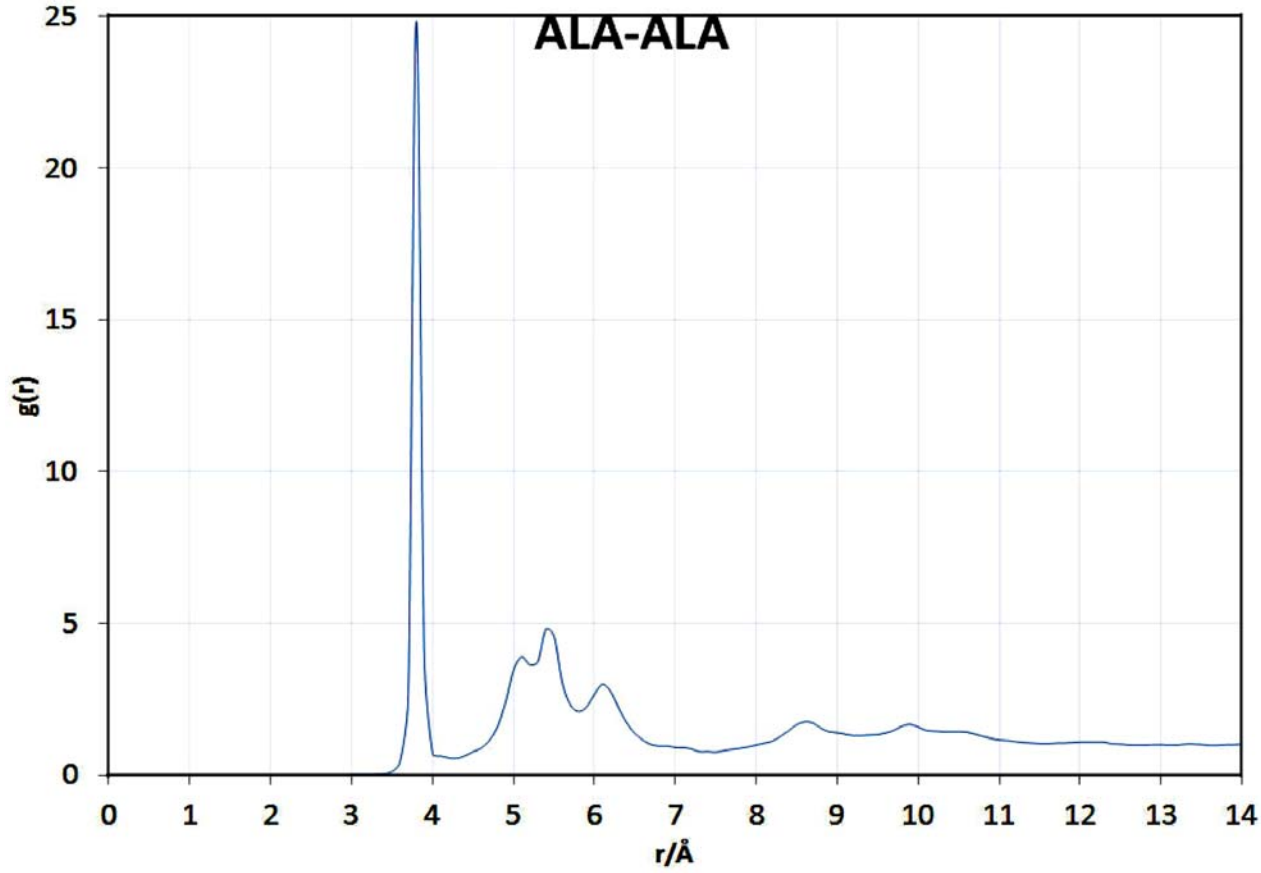

Figure 2. Radial distribution function for two alanine molecules.

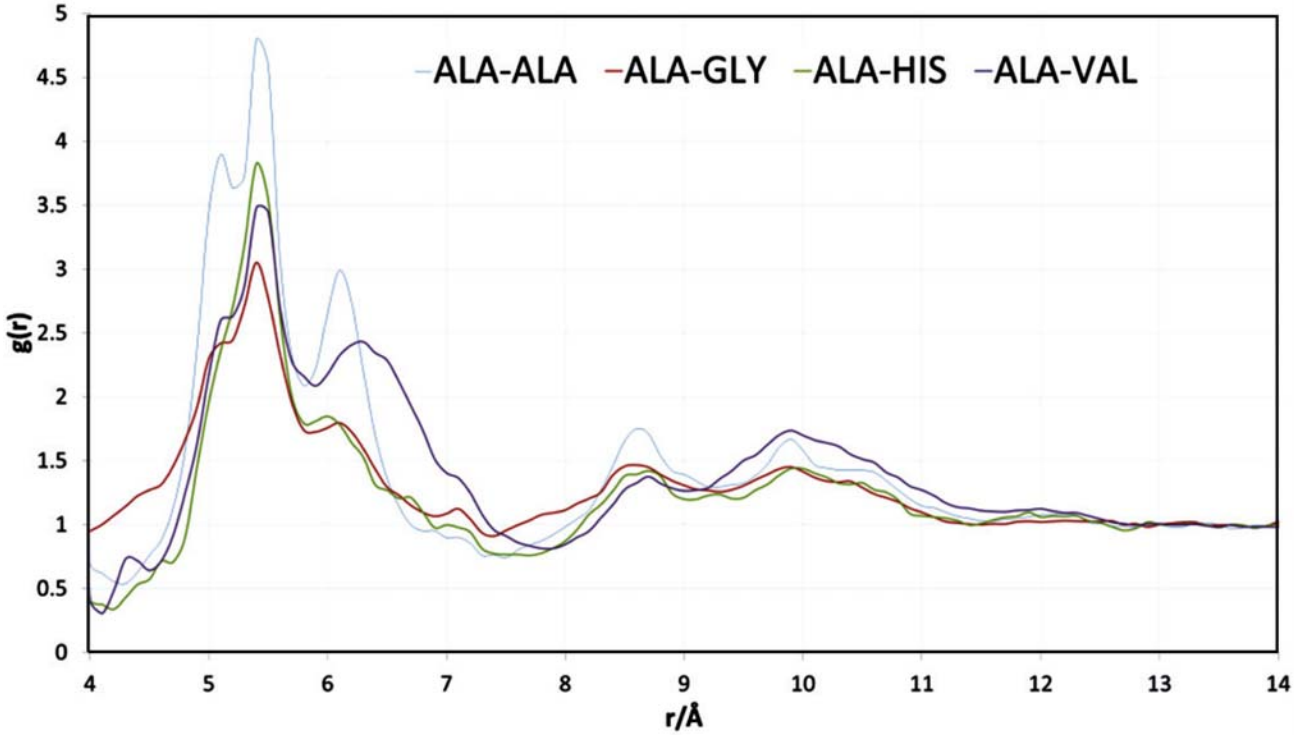

Figure 3. Radial distribution functions for pairs alanine-alanine, alanine-glycine, alanine-histidine and alanine-valine.

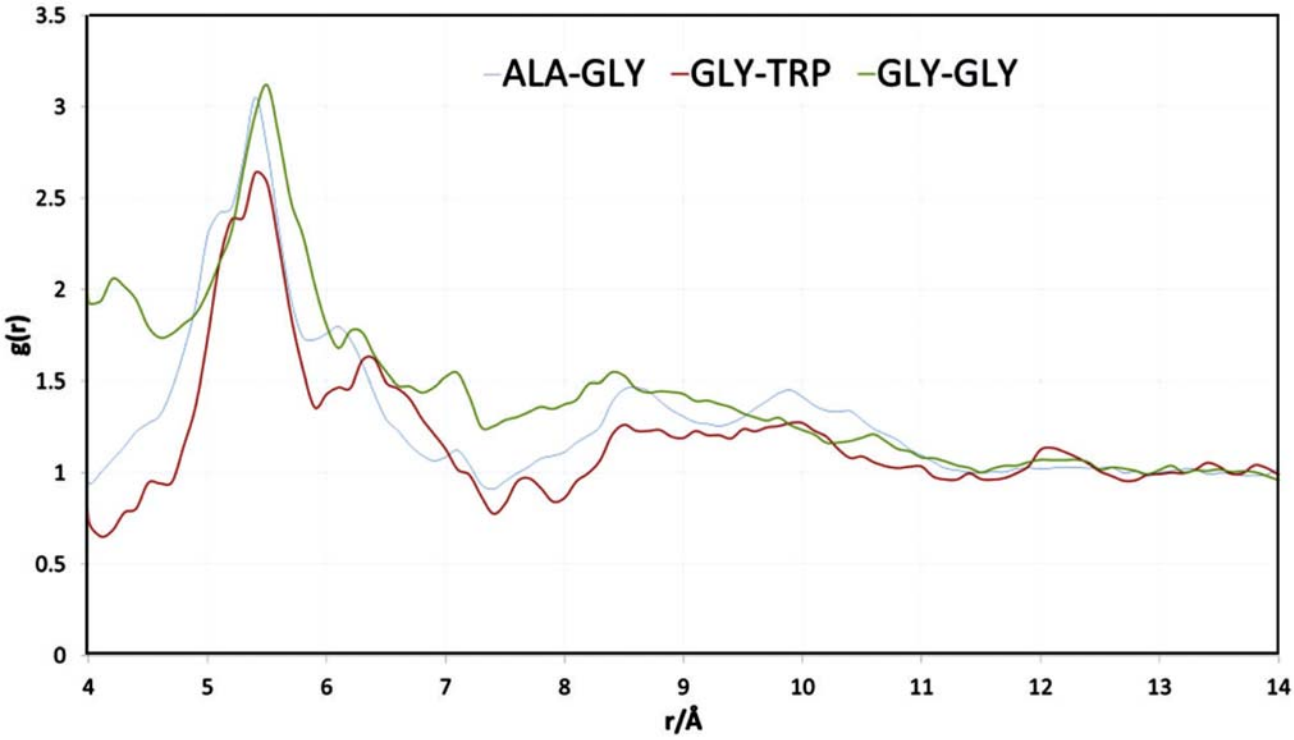

Figure 4. Radial distribution functions for pairs alanine-glycine, glycine-tryptophane, and glycine-glycine.

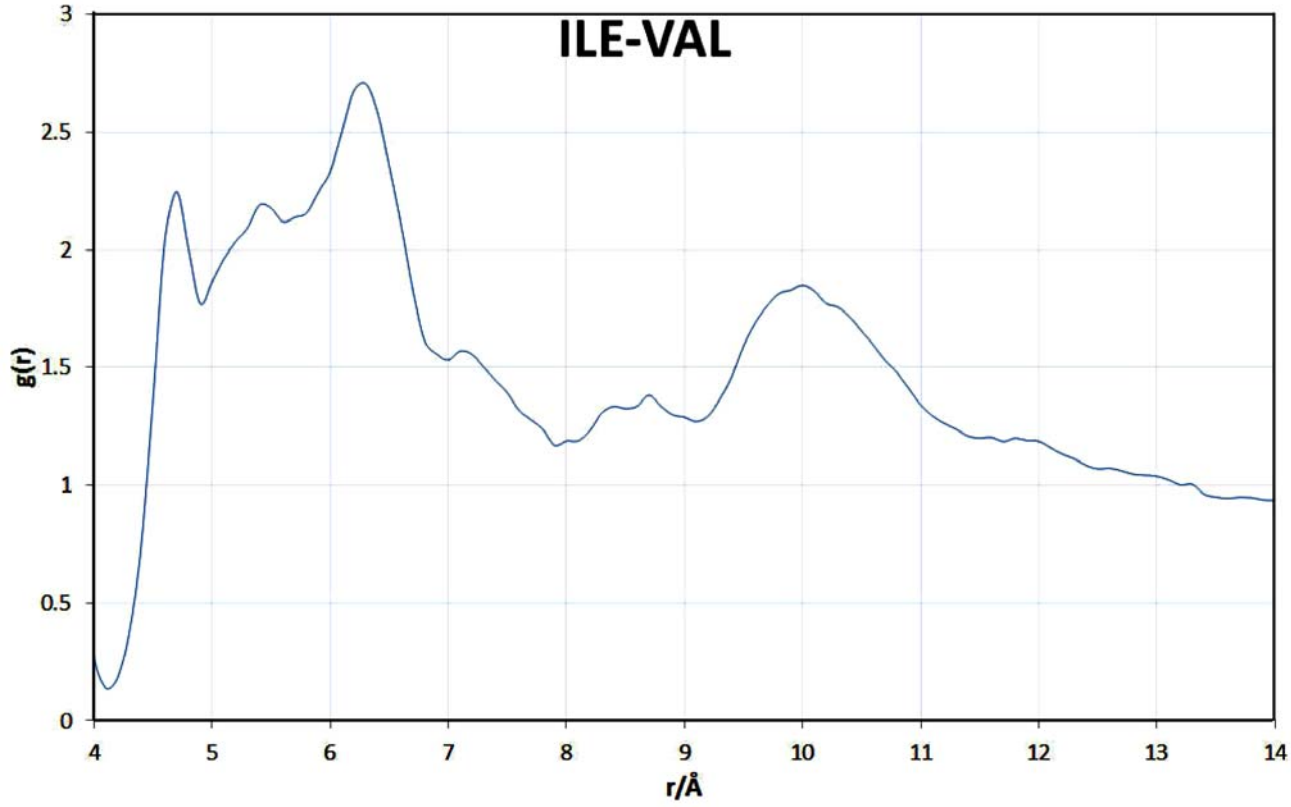

Figure 5. Radial distribution function for the pair isoleucine-valine.

sent shorter than usual tertiary interactions, an idea supported by the fact that glycine lacks any real side chain, as well as that both amino acids are non-polar.

In Figure 5, which is certainly one of the most unusual correlation functions, curve for the pair of two highly hydrophobic amino acids of the BCAA group isoleucine-valine exhibits an out-of-the-ordinary shape in the entire region. Perhaps most obviously, the alpha helix peak at 5.4 Å is actually lesser than the beta sheets peak at 6.2 Å, indicating an extremely high preference of the pair for the latter kind of secondary structure. Also highly unusual is a very tall and defined peak at 4.7 Å, which is likely attributed to more densely coiled π helices. In fact, these two amino acids seem to be more present as a pair in π helices than in otherwise much more prevalent alpha helices, which is certainly unique. Indeed, along with other histograms displaying peaks in this region, it could indicate that π helices are not as rare as has been generally assumed. We can see a defined peak at 7.1 Å as well, indicating what we believe to be a propensity of the pair to hydrophobic interactions at this distance, or maybe even a previously unidentified type of secondary structure, perhaps such as a coil with a pitch of this length. Also of interest is the combination of a little expressed peak in the area of around 8.5 Å, and a much larger, although undefined, peak at 10 Å; it so seems that the pair likes to form hydrophobic interactions at a larger distance, which may be between non-proximal beta strands, considering the high preference of the pair for this type of structure.

4. Conclusions

Pairwise distribution functions were calculated between C-alpha atoms of 20 amino acids in a large ensemble of Protein Data Bank structures. Results for 210 pairs were obtained and shown what can be learned from them and how to interpret them. We have determined that the implementation of pairwise distribution functions enables elucidation of useful information about preferential distances between amino acids in protein structures. We have found typical peaks in radiuses of secondary and tertiary interactions. These preferential distances are very typical, but different pairs can have different combinations of them. Of some interest, we found that certain pairs have a high propensity for π helices, which may be more common than generally thought. Distribution function by itself has certain problems in its use for this purpose, namely non-isotropicity, border problems and a more complex environment of 20 different amino acids all interacting with each other. Distribution function values could be improved by acknowledging sidechain orientation of the pair, thus more clearly separating backbone interactions from tertiary ones. Another step could be a determination of effective potentials.

5. Acknowledgments

Financial support from Slovenian Research Agency through grant P1 0103-0201 is appreciated. KK gratefully acknowledges fellowship by Decido, creative solutions d.o.o.

6. References

1. Protein Folding, edited by T. E. Creighton, **1992**, Freeman, New York.
2. E. Alm and D. Baker, *Proc. Natl. Acad. Sci. USA,* **1999,** *96*, 11305–11310. http://dx.doi.org/10.1073/pnas.96.20.11305
3. D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry 5th Ed*., **2008,** Freeman, New York.
4. R. Zangi, *Physical Review,* **2014,** *E 89* 012723
5. F. Mahmoudinobar, C. L. Dias, R. Zangi, *Physical Review,* **2015,** *E 91 032710*
6. E. Shakhnovich and A. Gutin*, J. Chem. Phys*., **1990,** *93*, 5967–5971. http://dx.doi.org/10.1063/1.459480
7. S. Miyazawa and R.L. Jernigan, *J. Mol. Biol*., **1996,** *256*, 623– 644. http://dx.doi.org/10.1006/jmbi.1996.0114
8. C. D. Snow, H. Nguyen, V. S. Pande, and M. Gruebele, *Nature*, **2002**, *420*, 102–106. http://dx.doi.org/10.1038/nature01160
9. T. Urbic and C. L. Dias. *J. Chem. Phys.,* **2014,** *140*, 165101. http://dx.doi.org/10.1063/1.4871663
10. C. L. Dias and H. S. Chan. *J. Phys. Chem. B*, 2014, *118* , 7488–7509. http://dx.doi.org/10.1021/jp501935f
11. Z. Su and C. L. Dias. *J. Phys. Chem. B,* **2014,** *118*, 10830–10836. http://dx.doi.org/10.1021/jp504798s
12. H. Zhou and Y. Zhou. *Protein Sci.,* **2002**, *11*, 2714–2726. http://dx.doi.org/10.1110/ps.0217002
13. R. Das, B. Qian, S. Raman, R. Vernon, J. Thompson, P. Bradley, S. Khare, M. D. Tyka, D. Bhat, D. Chivian, D. E. Kim, W. H. Sheffler, L. Malmstrom, A. M. Wollacott, C. Wang, I. Andre, D. Baker. *Proteins,* **2007**, *69*, 118–128. http://dx.doi.org/10.1002/prot.21636
14. H. Zhou, S. B. Pandit, S. Y. Lee, J. Borreguero, H. Chen, L. Wroblewska, J. Skolnick. *Proteins,* **2007**, *69*, 90–97. http://dx.doi.org/10.1002/prot.21649
15. Y. Zhang. *Proteins,* **2007**, *69*, 108–117. http://dx.doi.org/10.1002/prot.21702
16. M. L. Sippl, *J. Comput. Aided Mol. Des*., **1993**, *7*, 473–501. http://dx.doi.org/10.1007/BF02337562
17. J. Moult. *Curr. Opin. Struct. Biol*., **1997**, *7*, 194–199. http://dx.doi.org/10.1016/S0959-440X(97)80025-5
18. A. Rojnuckarin and S. Subramaniam. *Proteins*, **1999**, *36*, 54–67. http://dx.doi.org/10.1002/(SICI)1097-0134(19990701)36:1<54::AID-PROT5>3.0.CO;2-B
19. U. Bastolla, M. M. Vendruscolo and E. W. Knapp, *Proc. Natl. Acad. Sci. USA*, **2000**, *97*, 3977–3981. http://dx.doi.org/10.1073/pnas.97.8.3977

20. M. Ota, Y. Isogai and K. Nishikawa. *Protein Eng*., **2001**, *14*, 557–564. http://dx.doi.org/10.1093/protein/14.8.557
21. T. Lazaridis and M. Karplus. *Curr. Opin. Struct. Biol*., **2000**, *10*, 139–145. http://dx.doi.org/10.1016/S0959-440X(00)00063-4
22. M. R. Betancourt. *J. Phys. Chem. B*, **2008**,*112*, 5058-5069.23. B. J. Yoon, M. S. Jhon and H. Eyring, *Proc. Natl. Acad. Sci. USA*, **1981**, *78*, 6588–6591.
24. R.G. Parra, R. Espada, I. E. Sanchez, M. J. Sippl and D. U. Ferreiro, *J. Phys. Chem. B*, **2013**, *117*, 12887–12897. http://dx.doi.org/10.1021/jp402105j
25. A. Babajide, I. L. Hofacker, M. J. Sippl and P. F. Stadler, *Fold. Des.,* **1997**, *2*, 261–269. http://dx.doi.org/10.1016/S1359-0278(97)00037-0
26. A. Ben-Naim, *J. Chem. Phys*., **1997**, *107*, 3698–3706. http://dx.doi.org/10.1063/1.474725
27. W. A. Koppensteiner and M. J. Sippl, *Biochemistry*, **1998**, *63*, 247–252.
28. C. L. Dias, *Phys. Rev. Lett.,* **2012,** 109, 048104 http://dx.doi.org/10.1103/PhysRevLett.109.048104

Povzetek

Razumevanje prostorskega zvitja proteinov prek njihovih aminokislinskih zaporedij ima v današnjih vedah o življenju ogromen potencial. Sposobnostračunalniškega predvidevanja sekundarne in terciarne strukture iz primarne bo omogočilo mnogo hitrejše in učinkovitejše odkrivanje organskih snovi s terapevtskim ali drugače bioaktivnim potencialom, kar bo v veliki odpravilo potrebo po sintezi velikega števila organskih spojin in njihovega preizkušanja za fiziološke učinke. V našem članku predstavljamo uporabo analize s korelacijskimi funkcijami, matematičnega orodja, običajno namenjenega za opis lastnosti tekočin, za ugotavljanje statistično preferenčnih razdalj med aminokislinami v proteinskih strukturah. Izračunali smo parske porazdelitvene funkcije med C-alfa atomi dvajsetih aminokislin v veliki zbirki PDB struktur. Korelacijske funkcije pokažejo karakteristične razdalje v medsebojnih interakcijah aminokislin. Slikovno smo prikazali različna nagnjenja vseh 210-ih posameznih parov do tvorjenja raznih elementov sekundarne strukture, nekatera pa smo tudi interpretirali.