Biochemical and Biophysical Research Communications 436 (2013) 725–729

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Atom depth analysis delineates mechanisms of protein intermolecular interactions

Davide Alocci [a], Andrea Bernini [a], Neri Niccolai [a,b,*]

[a] Department of Biotechnology, Chemistry and Pharmacy, University of Siena, via A. Fiorentina 1, 53100 Siena, Italy
[b] SienaBioGrafix Srl, via A. Fiorentina 1, 53100 Siena, Italy

ARTICLE INFO

Article history:
Received 30 May 2013
Available online 17 June 2013

Keywords:
Protein–protein interactions
Structural layers
3D atom depth
Protein surface charge
Protein–protein network

ABSTRACT

The systematic analysis of amino acid distribution, performed inside a large set of resolved protein structures, sheds light on possible mechanisms driving non random protein–protein approaches. Protein Data Bank entries have been selected using as filters a series of restrictions ensuring that the shape of protein surface is not modified by interactions with large or small ligands. 3D atom depth has been evaluated for all the atoms of the 2,410 selected structures. The amino acid relative population in each of the structural layers formed by grouping atoms on the basis of their calculated depths, has been evaluated. We have identified seven structural layers, the inner ones reproducing the core of proteins and the outer one incorporating their most protruding moieties. Quantitative analysis of amino acid contents of structural layers identified, as expected, different behaviors. Atoms of Q, R, K, N, D residues are increasingly more abundant in going from core to surfaces. An opposite trend is observed for V, I, L, A, C, and G. An intermediate behavior is exhibited by P, S, T, M, W, H, F and Y. The outer structural layer hosts predominantly E and K residues whose charged moieties, protruding from outer regions of the protein surface, reorient free from steric hindrances, determining specific electrodynamics maps. This feature may represent a protein signature for long distance effects, driving the formation of encounter complexes and the eventual short distance approaches that are required for protein–protein functional interactions.

1. Introduction

Interactions among the molecular components of Life determine a huge variety of biochemical events hosted by Nature. Nowadays, the available structural information of biomolecules is large enough to contain already many relevant clues for deciphering, at the atomic level, mechanisms of biological processes. Indeed, all the structural features that are required for stabilizing protein adducts with nucleic acids, small molecules or other proteins, are already well known and updated by the continuous growth of information stored in the Protein Data Bank, PDB [1].

At the end of last century, a System Biology perspective has been added to Structural Biology and molecular mechanisms of protein–protein interactions, PPI, have been investigated with experimental and computational approaches. Thus, the formation of encounter complexes has been proposed as a preliminary step for PPI [2–5] also under conditions of macromolecular crowding [6].

It is apparent that, in the molecular crowd typical of biological fluids, the formation of encounter complexes cannot be driven simply by self-diffusion processes, suggesting that a protein–protein networking must be present. Electrostatic assistance through long range interactions to drive protein associations has been proposed for hub proteins [2], that is for proteins exhibiting multiple PPI, and the presence of high surface charge has been suggested as the main source of their enhanced social activity [7].

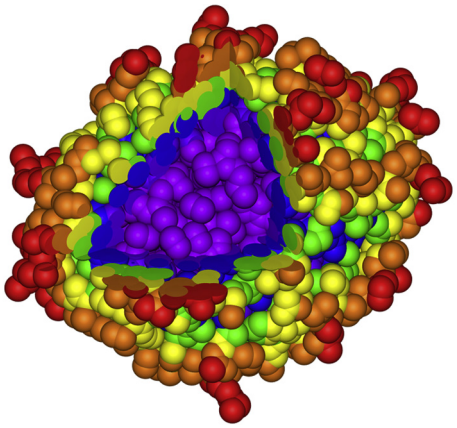
In order to find possible mechanisms which determine non-random translations between proteins and their eventual ligands, in the present report we have performed a systematic analysis on the population and distribution of amino acid residues on the surface of unbound proteins. Thus, to remove possible biases induced by protein–ligands complexation from all the available PDB protein structures, a reduced dataset of only protein singles, DOOPS, has been built.

To have a quantitative assessment of amino acid composition of protein surfaces, we used our original 3D atom depth analysis, similarly to what has been recently done to define protein cores [8]. A parameter dubbed depth index, D_i, has been used as a tool to classify DOOPS atoms into different structural layers. Recurrent presence of specific amino acids on protruding moieties of protein surface hosted by the outer structural layer has been here analyzed

* Corresponding author at: Department of Biotechnology, Chemistry and Pharmacy, University of Siena, via A. Fiorentina 1, I-53100 Siena, Italy. Fax: +39 (0) 577 234903.

E-mail addresses: alodavide@gmail.com (D. Alocci), andrea.bernini@unisi.it (A. Bernini), neri.niccolai@unisi.it (N. Niccolai).

0006-291X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2013.06.024



L_n	L_0	L_1	L_2	L_3	L_4	L_5	L_6
D_i	< 0.2	0.2 - 0.4	0.4 -0.6	0.6 – 0.8	0.8 – 1.0	1.0 – 1.2	> 1.2
% atoms	17,9	17,1	17,7	17,4	14,4	9,4	5,7
color	violet	indigo	blue	green	yellow	orange	red

Fig. 1. Spacefill representation of protein structural layers: rainbow coloring of the seven atom layers defined on the basis of atom depth indexes for the cyanobacterial bicarbonate transport protein, CmpA, ID PDB code 2I49, and the used D_i limits are shown in the inset. Atom percent presence in each structural layer of DOOPS proteins are also given. The same octants of L_1–L_6 structural layers have been removed to show L_0. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

to explore specific mechanisms of long range protein–protein interactions.

2. Materials and methods

2.1. The DOOPS and sDOOPS datasets

The entire content of the Protein Data Bank has been scanned with a final updating on May 8, 2013 when 90,424 structures were available. In order to take into account only proteins with well defined secondary structure elements, only crystal structures have been selected with the following limitations provided by the PDB advanced search interface: (i) homologue removal at 95% identity, (ii) only one chain in the asymmetric unit, (iii) only proteins with monomeric biological assembly and (iv) only one entity in the asymmetric unit. Among all these restrictions, the last one considerably reduces the number of proteins in the dataset, as most of protein PDB structures refer to protein–ligand complexes. Thus, only 2410 PDB structures are included in the dataset of only protein singles, DOOPS, where only proteins whose surface shape is minimally influenced by molecular neighbors different than water should be present. In Fig. S1, we show a comparison between the content of DOOPS and the whole 2013_05 release of Swiss-Prot. In spite of the huge difference in the number of objects contained in the two datasets, they report very similar amino acid populations and protein chain length distributions, suggesting that data derived from DOOPS are statistically significant. In order to obtain a subset of DOOPS containing only proteins with a relevant involvement in PPI, DOOPS has been further refined by choosing in the PDB advanced search interface a PubMed selection with the following syntax: protein–protein interaction AND structure AND X-ray. Thus, 321 protein structures were obtained; their relevance in PPI has been manually checked through the STRING server at the URL: http://string-db.org. It turned out that 84% of the selected proteins are actually involved in PPI and 29% with more than 10 other protein partners. Then, these 321 structures define a DOOPS subset of social proteins, sDOOPS.

2.2. Atom depth calculations

Atom depth for all the 4,657,574 atoms contained in DOOPS have been calculated with the SADIC algorithm [9], by using the freely downloadable software at http://www.sbl.unisi.it. Atom depths are calculated as depth indexes, D_i, defined as:

$$D_i = 2V_i/V_0 \quad (1)$$

where V_i is the exposed volume of a sphere of radius r_0 centered on atom i and V_0 is the exposed volume of the same sphere when centered on an isolated atom. As shown in Fig. S1, DOOPS protein se-

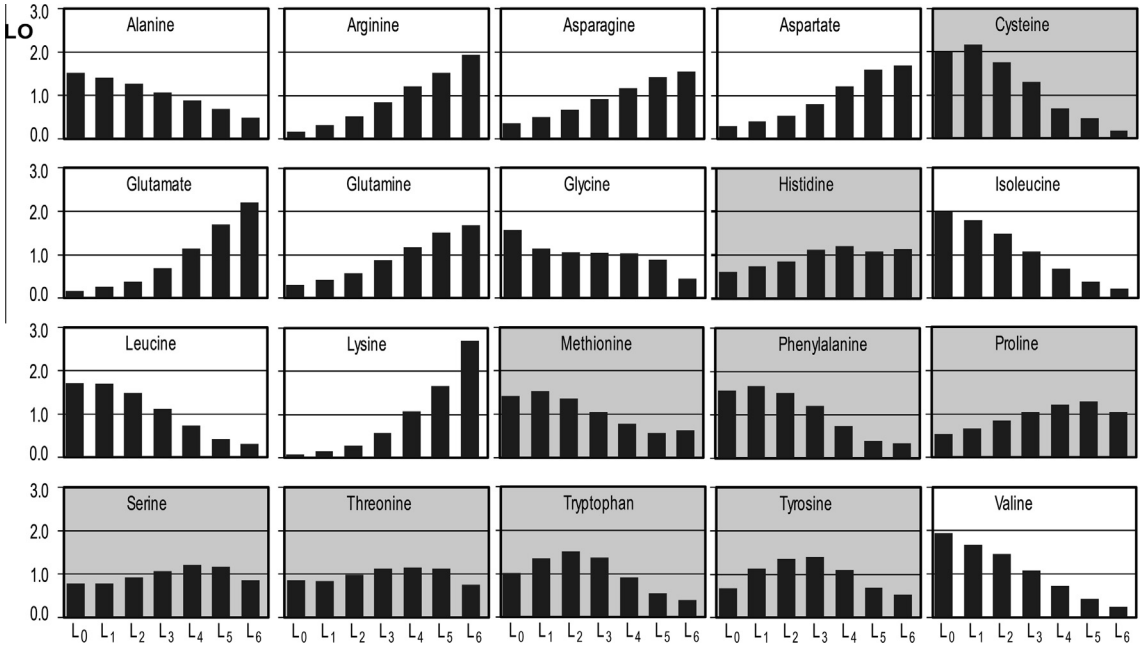


Fig. 2. Amino acid occupancy of structural layers: structural layer occupancy, LO, of the twenty natural amino acids normalized by their overall occurrence in DOOPS. Pale grey backgrounds highlight amino acids not having monotonic content changes at increasing L_n indexes.

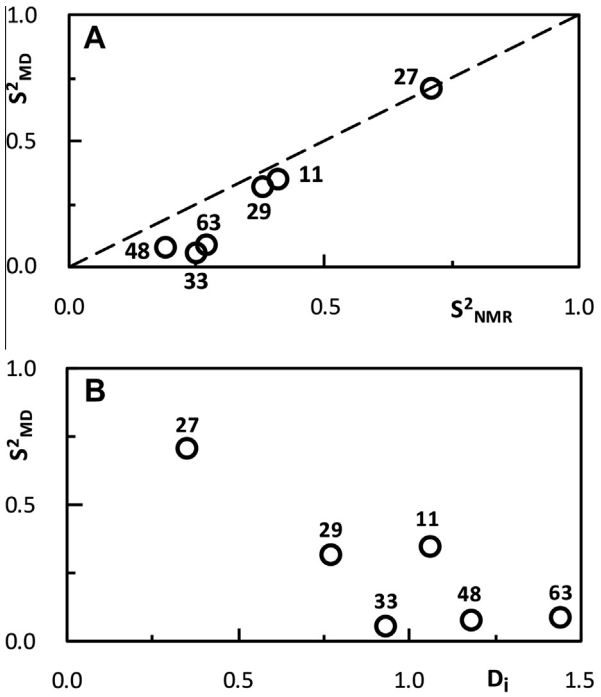


Fig. 3. Structural dynamics and D_i: in panel A, NMR derived order parameters, S^2_{NMR}, of ubiquitin NZ atoms of lysines from reference 14 are compared with S^2 values obtained from a 100 ns MD trajectory, S^2_{MD}. In panel B, the same S^2_{MD}'s are compared with the corresponding D_is.

quences largely differ in length and, accordingly, different radii of SADIC probing sphere, r_0, were needed for optimal D_i. calculations [9]. Thus, r_0 ranged from 7 to 15 Å with the largest value required

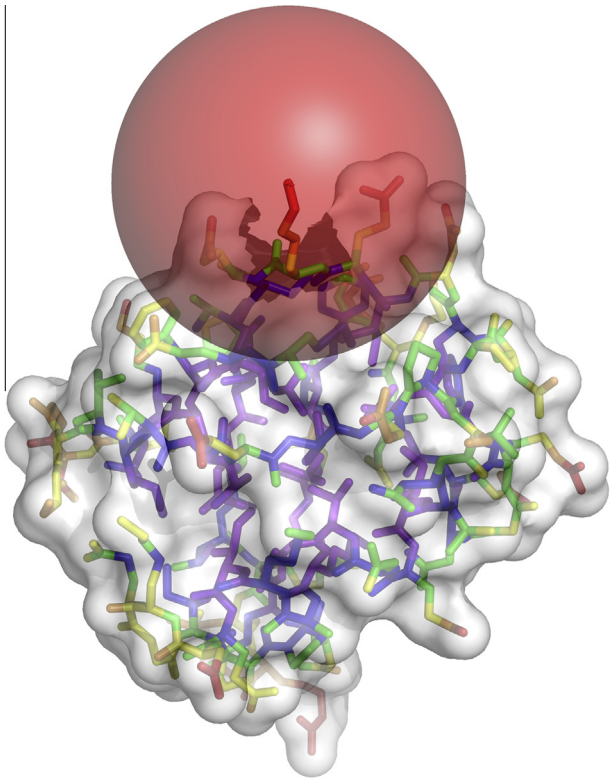


Fig. 4. Surface amino acid open space: ubiquitin K63 side chain is shown in sticks inside the 9 Å radius probing sphere used for D_i calculation of the NZ atom.

only by three DOOPS proteins, *i.e.* the ones with the PDB ID codes 2I49, 1OGM and 4FNV. The fact that these three proteins are very different in sequence length, 429, 574 and 702 residues respectively, suggests, as confirmed by Fig. S2, that at a given r_0 the smaller is the number of residues the more spherical is the overall protein shape. For each of the 583,147 amino acids of the dataset, the obtained highest D_i value, D_{imax}, has been chosen as a suitable parameter to label the most peripheral atoms. It can be noted that residues exhibiting D_{imax} lower than 0.2 correspond to the ones located by ProCoCoA in the protein core [8]. Moreover, all the residues bearing atoms with D_{imax} higher than 1.2 are found in the most protruding moieties of the selected protein structures.

2.3. Protein structure partitioning into onion-like layers

D_i values, calculated by using SADIC for all DOOPS protein atoms, were automatically stored in the B-factor field column of modified PDB files and used to assign each atom to one of the onion-like layers forming the overall protein tertiary structure. Dividing the protein structure in seven layers, L_0–L_6, equally spaced of 0.2 D_i units, is particularly useful, as the inner L_0 layer almost overlaps with the recently D_i defined protein core [8] and the outer layer L_6 includes all the most protruding protein moieties. Then, B-factors color coding provided by standard molecular visualization software can produce images which may be very informative on protein surface distribution of protrusions and pockets, see Fig. 1.

The number of atoms contained in each layer of DOOPS proteins, suggests the consistency of the used criteria for protein structure partitioning, see the inset of Fig. 1. The average atom content of the seven layers reflects how the protein shape, with its protrusions and pockets, influences the number of atoms that are present in each layer. Only the inner L_0–L_3 layers, indeed, incorporate a constant fraction of atoms regardless to the protein surface shape, as increasing layer volumes are counterbalanced by decreasing atom densities. The outer L_4–L_6 layers, indeed, hosting the entire roughness of protein surfaces, accommodate progressively less atoms.

2.4. Amino acid assignment to the structural layers

The above defined seven structural layers are differently occupied by the twenty natural amino acids. In order to derive amino acid percent occupancy in each of the seven onion-like layers, for all the atom of the dataset D_i's have been computed and D_{imax} has been selected for each residue as the representative amino acid flag. From manual inspection on a reduced set of proteins, it has been noticed that less ordered protein fragments, such the ones frequently observed at amino- or carboxy-termini, are characterized by a contiguous series of three or more residues exhibiting D_{imax} values >1.5. In order to avoid contributions to L_6 content from the latter fragments, whenever for any n residue the condition $(D_{imax})_n > 1.5$; $(D_{imax})_{n+1} > 1.5$; $(D_{imax})_{n+2} > 1.5$ is met, the corresponding residues are automatically excluded by further analysis. Accordingly, from the total 589,383 analyzed residues, 6236 were removed.

2.5. Molecular dynamics simulations

Time dependent evolution of the surface shape of ubiquitin has been investigated through Molecular Dynamics, MD, simulations. Trajectory obtained from a 100 ns MD run in explicit solvent has been obtained by using GROMACS and the AMBER force field on the PDB file with ID code 1UBQ. Details of the used procedures are given elsewhere [10].

Table 1
DSSP secondary structure assignment of K_{L6} and E_{L6} residues in their most frequent residue environments.

DSSP*	helix	extended	loop	N
K_{L6} #	5267	1504	4134	2592
E_{L6} #	5273	785	3790	2078
K_{L6} %	39,03	11,14	30,63	19,2
E_{L6} %	44,22	6,59	31,78	17,42
EK_{L6}	49,96	6,19	29,95	13,89
LK_{L6}	43,36	9,82	24,33	22,49
GK_{L6}	14,76	20,88	35,98	28,38
KK_{L6}	46,1	7,8	29,9	16,2
$K_{L6}E$	54,2	6,37	26,08	13,34
$K_{L6}L$	46,7	12,92	20,32	20,05
$K_{L6}G$	13,67	9,93	59,26	17,14
$K_{L6}D$	36,6	7,1	43,32	12,98
EE_{L6}	56,88	4,56	27,25	11,3
LE_{L6}	49,03	6,76	23,65	20,57
PE_{L6}	41,28	1,72	42,5	14,5
KE_{L6}	48,63	5,82	31,89	13,67
$E_{L6}E$	60,27	3,2	23,91	12,62
$E_{L6}K$	52,67	3,98	31,24	12,12
$E_{L6}G$	16,06	8,49	61,11	14,34
$E_{L6}L$	56,59	8,37	19,19	15,85

[a]Number of proteins in DOOPS: 2410; number of protein without DSSP output: 16. Helix, β strand and loop contents, respectively, refer to the sum of (H + G + I), (B + E) and (T + S) in the DSSP nomenclature, where H = α helix, B = residue in isolated β bridge, E = extended strand, *G* = 3/10 helix, I = π helix, T = hydrogen bonded turn, S = bend, N = not determined. Dipeptides containing two charged amino acids are highlighted in grey.

3. Results and discussion

The dataset of 2410 protein structures, obtained from the PDB by including the above mentioned limits, has been used to explore the amino acid composition at different molecular depths. Among the available algorithms to calculate amino acid depths in a large array of structures [9,11,12], we chose the home made SADIC algorithm to obtain individual depths of all the 4,657,574 atoms contained in DOOPS. SADIC describes the depth of atoms on the basis of the exposed volume of a probing sphere, thus emphasizing exposures of atoms belonging to protruding moieties, the main object of the present investigation.

Amino acid distribution among the above defined structural layers has been monitored by selecting, for each of the 589,383 residues of DOOPS, the highest D_i value, D_{imax}. Then D_{imax} has been used to classify all protein residues to the competent structural layer and their occupancy, normalized by the overall amino acid presence in DOOPS, are given as LO in Fig. 2. Numerical data of L_n amino acid contents of DOOPS and sDOOPS proteins are also reported in Tables SI and SII. As almost identical L_n distributions were obtained for the two datasets, no further distinction between social and non social proteins will be made. Three different evolution patterns of layer occupancy, LO, from the core to the surface can be seen: (i) LO of Q, R, H, K, N, D and E constantly increase, (ii) LO of V, I, L, A, and G constantly decrease and (iii) in the case of C, W, F, M, P, S, T and Y to an initial increase, a decrease is found with the trend inversion centered at different structural layers. So, L_n occupancy inversion occurs within the first three layers in the case of C, W, F and M, at L_3 for Y and at L_4 in the case of P, S, and T.

Amino acid content of L_6 has been analyzed in detail, being the first quantitative analysis of the most protruding moieties of protein surfaces. As expected, charged amino acids, *i.e.* D, E, H, K and R, collectively represent an amino acid fraction of 54.30% with only a minor contribution, 2.46%, from H residues. From the present D_{imax} survey it is possible to confirm the abundance of electric charges on protein surfaces, but, even more important, it is also possible to identify structural features which have not been discussed before. In L_6, indeed, K and E residues are much more abundant, respectively 17.13% and 15.16%, than the equally charged R and D, respectively 9.67% and 9.88%. This finding is quite unexpected it cannot be attributed just to differences in chain length or size, deserving more careful considerations. First of all, it must be noted that D_{imax} values of all the 36,671 K and 40,030 E residues of DOOPS proteins always come respectively from NZ and OE atoms, in the PDB nomenclature.

Sequence neighbors of K and E residues in L_6, E_{L6} and K_{L6} respectively, have been analyzed in terms of preceding and subsequent amino acids. The occurrence of all the natural amino acids in the $n-1$ and $n+1$ position in respect to each E_{L6} and K_{L6} is reported in Table SIII. Being X a generic amino acid, among all the possible XK, KX, XE and EX dipeptides, the most frequent occurrences were EK_{L6}, LK_{L6}, GK_{L6}, KK_{L6}, $K_{L6}E$, $K_{L6}L$, $K_{L6}G$, $K_{L6}D$, EE_{L6}, LE_{L6}, PE_{L6}, KE_{L6}, $E_{L6}E$, $E_{L6}K$, $E_{L6}G$ and E_6L. Data of Table SIII indicate also that along DOOPS protein sequences frequently two consecutive amino acids bearing opposed or equal electric charges are present. An estimate on the reciprocal position of these electric charges can be obtained by assigning the secondary structure elements to all E_{L6} and K_{L6} of the dataset. Thus, the DSSP algorithm [13] has been used on all DOOPS proteins and the results are shown in Table 1. It is apparent that, unless P or G are present in the fragments, protruding K and E are predominantly located in α helices. The latter condition is always met by EE, EK, KE and KK at very similar extents, suggesting that protruding electric charges are often located in close proximity, yielding frequent occurrences of surface ion pairs.

The possibility that D_is correlate with dynamic aspects of charged side chains has been considered by comparing D_i values of K with the corresponding order parameter S^2 experimentally derived from ^{15}N nuclear relaxation, or calculated from molecular dynamics, MD, simulations. The recently performed ^{15}N nuclear relaxation studies on the dynamics of ubiquitin side chains [14] have been used as an experimental reference for discussing MD S^2 and D_i of K side chains of this protein. In Fig. 3A, comparison of the available S^2 data calculated from NMR studies, S^2_{NMR}, with order parameters obtained by the 100 ns MD simulation, S^2_{MD}, shows the expected good correlation of local mobility derived from the two independent procedures. In the case of ubiquitin, the good correlation between S^2_{MD} and D_i values of lysyl NZ atoms is also apparent, see Fig. 3B. In general, the dynamic nature of D_i values reflects the way SADIC algorithm works, as D_i values higher than 1.2 imply large exposed volumes of the probing sphere and, hence, lack of hindrance for side chain internal reorientation, see Fig. 4.

Thus, many K and E protruding side chains are positioned on protein surfaces, mainly exposed to the solvent by secondary structure elements, free to move rapidly their electric charges. Data reported in Table SI indicate high L_6 occupancies also for R and D residues, but at a much lesser extent. It must be noted that argi-

nine, frequently found in protein–protein interfaces, bears the positive charge on a rather sticky side chain [15] whose motion can be limited by transient interactions more than K side chains. Similarly, due the different side chain length of D and E, slower motion of the charged aspartate side chain than the one of glutamate can be predicted. Thus, the fact that less R and D than K and E occupy L_6, see Fig. 2 and Table SI, can be interpreted as the need to ensure that each protein present its surface not only with the required local electrostatics for optimal intermolecular adduct formation, but also an array of rapidly moving electric charges generating specific electromagnetic fields for long distance protein–ligand recognition.

It must be stressed, here, that the use of 3D atom depths calculated by SADIC algorithm allows the first quantitative analysis of amino acid occupancy at different structural layers of molecular structures [9]. In the case of proteins, amino acid compositions of cores have been already defined with enough resolution to observe fold specific amino acid contents [8]. Moving outwards in a protein structure, quantitative analyses of amino acid compositions of outer surface moieties reveal new features which can be fundamental in protein–ligand interactions. The abundance of K and E residues in the most external protein structural layer, previously suggested as typical of hub proteins [7], determines the exposure of many rapidly reorienting charged groups. Thus, these two residues become responsible of electromagnetic fields which favor all the preliminary non random approaches before protein–ligand encounters and complex formation.

As a final remarks, it must be underlined here that in the same way electrostatics has been extensively employed to understand the stability of complex molecules and their adducts, electrodynamics can now provide quantitative explanations of long distance messaging driving not only PPI, but also any other protein–ligand interactions.

Acknowledgments

This work has received financial support from the Istituto Toscano Tumori. Thanks are due to Alberto Toccafondi, Piero A. Temussi and Henriette Molinari for helpful discussions and to SienaBiografix Srl for technical assistance.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bbrc.2013.06.024.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, et al., The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[2] G. Schreiber, A.R. Fersht, Rapid, electrostatically assisted association of proteins, Nat. Struct. Biol. 3 (1996) 427–431.
[3] T.L. Blundell, J. Fernández-Recio, Cell biology: brief encounters bolster contacts, Nature 444 (2006) 279–280.
[4] T.K. Yu, Y.J. Yun, K.O. Lee, J.Y. Suh, Probing target search pathways during protein–protein association by rational mutations based on paramagnetic relaxation enhancement, Angew. Chem. Int. Ed. Engl. 52 (2013) 3384–3388.
[5] M. Ubbink, The courtship of proteins: understanding the encounter complex, FEBS Lett. 583 (2009) 1060–1066.
[6] S. Qin, L. Cai, H.X. Zhou, A method for computing association rate constants of atomistically represented proteins under macromolecular crowding, Phys. Biol. 9 (2012) 066008.
[7] A. Patil, H Nakamura, Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks, FEBS Lett. 580 (2006) 2041–2045.
[8] S. Bottini, A. Bernini, M. De Chiara, et al., ProCoCoA: a quantitative approach for analyzing protein core composition, Comput. Biol. Chem. 43 (2013) 29–34.
[9] D. Varrazzo, A. Bernini, O. Spiga, et al., Three-dimensional computation of atom depth in complex molecular structures, Bioinformatics 21 (2005) 2856–2860.
[10] A. Bernini, O. Spiga, R. Consonni, et al., Hydration studies on the archaeal protein Sso7d using NMR measurements and MD simulations, BMC Struct. Biol. 11 (2011) 44.
[11] A. Pintar, O. Carugo, S. Pongor, Atom depth in protein structure and function, Trends Biochem. Sci. 28 (2003) 593–597.
[12] S.W. Chen, J.L. Pellequer, Adepth: new representation and its implications for atomic depths of macromolecules, Nucleic Acids Res. (2013).
[13] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[14] A. Esadze, D.W. Li, T. Wang, R. Brüschweiler, J. Iwahara, Dynamics of lysine side-chain amino groups in a protein studied by heteronuclear ^{1}H–^{15}N NMR spectroscopy, J. Am. Chem. Soc. 133 (2011) 909–919.
[15] S. Jones, Computational and structural characterisation of protein associations, Adv. Exp. Med. Biol. 747 (2012) 42–54.