Hydrophilicity Matching – a Prerequisite for the Formation of Protein-Protein Complexes

Mario Hlevnjak[1], Gordan Zitkovic[2] & Bojan Zagrovic[1,3*]

[1]Laboratory of Computational Biophysics

Mediterranean Institute for Life Sciences

Mestrovicevo setaliste bb, 21000 Split, Croatia

[2]Department of Mathematics

University of Texas at Austin

2515 Speedway, 78712 Austin, TX, USA

[3]Department of Physics

Faculty of Science, University of Split

Teslina 12, 21000 Split, Croatia

*To whom correspondence may be addressed: Laboratory of Computational Biophysics,

Mediterranean Institute for Life Sciences, Mestrovicevo setaliste bb, 21000 Split, Croatia.

Phone: +385-99-555-6011, Fax: +385-21-555-605, E-mail: bojan.zagrovic@medils.hr

Running Title: Hydrophilicity Matching

Abstract

A binding event between two proteins typically consists of a diffusional search of binding partners for one another, followed by a specific recognition of the compatible binding sites resulting in the formation of the complex. However, it is unclear how binding partners find each other in the context of the crowded, constantly fluctuating, and interaction-rich cellular environment. Here we examine the non-specific component of protein-protein interactions, which refers to all physicochemical properties of the binding partners that are independent of the exact details of their binding sites, but which can affect their localization or diffusional search for one another. We show that, for a large set of high-resolution experimental 3D structures of binary, transient protein complexes taken from the DOCKGROUND database, the binding partners display a surprising, statistically significant similarity in terms of their total hydration free energies normalized by a size-dependent variable. We propose that colocalization of binding partners, even within individual cellular compartments such as the cytoplasm, may be influenced by their relative hydrophilicity in response to intra-compartmental hydrophilic gradients.

Introduction

From signal transduction cascades to enzymatic activation, from antibody-antigen recognition to cellular trafficking, direct noncovalent protein-protein interactions are the central pillar supporting most of biological functional activity on the molecular level [1]. Studies of such interactions usually focus on the specifics of the binding sites of the partners, while, at the same time, typically neglect their overall physicochemical properties, with a few notable exceptions at the protein aggregation frontier [2-4]. It is generally assumed that binding partners execute random-walk diffusion in a crowded, interaction-rich cellular environment prior to encounter [5-7]. However, specific interactions that underlie the binding-site recognition itself are all short range and could not serve the purpose of guiding this global, presumably non-specific search for the binding partner.

Moreover, it has been shown that, given low copy numbers and short life-spans of typical signalling proteins in crowded eukaryotic cells, it is imperative that binding partners in signalling cascades be colocalized in order to relay meaningful signals on reasonable time scales [8,9]. It is known that proteins colocalize due to segregation into different organelles or cellular compartments, sequestration via anchor and scaffold proteins, or sometimes even chemical modifications [1]. For example, interactions between two membrane proteins are greatly facilitated by both of them being colocalized in the 2D-membrane, which is easier to search by diffusion [9,10]. In this case, almost trivially, the finding of the binding partners is enabled by a non-specific element encoded in their respective structures – the hydrophobicity of their overall molecular surface. Importantly, this non-specific component may not be related to the specific features of the complementary binding sites of two

3

proteins, and still significantly influence the binding. However, little attention has been paid to such general mechanisms when it comes to cytoplasmic or nucleoplasmic proteins, with some notable exceptions [11-13]. For example, significant commonalities were found for isoelectric points of proteins assigned to different nuclear compartments [11,12], or for the pH of maximal stability of a complex and its monomers [14]. Nevertheless, the majority of these studies focused on the information encoded in the sequence of colocalized proteins, and not necessarily in their 3D-structure. Here we examine whether any signature of potential colocalization mechanisms for a large set of bound pairs is encoded in their 3D-structures by searching for commonalities between partners in the same complex.

As a source of 3D-structures of known cocrystallized interacting partners, we used the DOCKGROUND database of transient, binary protein complexes in their unbound form [15]. After performing additional short relaxation molecular dynamics (MD) simulations of each of the binding partners, we evaluated for each of them different geometric properties such as solvent-accessible surface area, radius of gyration, and volume, or different physicochemical properties such as total charge, isoelectric point, hydration free energy (HFE), and total electrostatic energy (EE). We quantified the degree of similarity of the binding partners by calculating intraclass correlation coefficients (ICCs) [16,17] for different properties, and evaluated the associated p-values via randomization tests.

Results and Discussion

Pairs of interacting partners were classified into different subsets based on their origin and the known site of complex formation in the cell or extracellular space, following the detailed

characterization of the entire set of 268 proteins (Table S1). We focus first on the subset

containing 118 eukaryotic proteins (59 pairs) interacting in the cytoplasm or nucleoplasm.

Similar results were obtained for a larger subset comprised of 162 proteins (81 pair),

including additionally also archeal and bacterial proteins, or for the complete set containing

268 proteins (134 pairs), including intra- or extracellular segments of transmembrane

proteins, as well as organellar and secreted proteins (Table 1, Fig. S3-S5).

If the known binding partners are compared with respect to the sequence length (N) of the

fragments found in cocrystallized complexes (Fig. 1A), they expectedly exhibit no similarity

whatsoever. The observed ICC of 0.462 and the associated p-value of 0.682075 mean that

the same degree of similarity occurs in 68 % of the cases where the pairs are chosen

completely at random from the studied subset. It is important to note that the majority of

cocrystallized proteins, including those that were examined herein, are fragments of larger

proteins. For example, within the subset of 118 eukaryotic proteins, the average

completeness of their structures is around 50 % (Table S1). Even so, one observes a

significantly higher similarity between the binding partners with respect to their radii of

gyration, which occurs by chance in only 3.3 % of the cases (Fig. 1B). Further discussion of

this finding is provided in Supporting Information. Comparison of the binding partners with

respect to their HFEs, as calculated by GB/SA methodology [18,19], does not reveal any

significant similarity between them (Fig. 1C). Surprisingly, when their HFEs are normalized by

either their respective sequence length (Fig. 1D), or volume (Fig. S3A), the binding partners

show highly significant similarity, which itself occurs by chance in a remarkable one out of

eleven thousand cases (p-value of 0.000089). Similarly, size-normalized electrostatic energy

also appears to be significantly matched between partners, while other calculated

properties, such as isoelectric point, volume or solvent-accessible surface area exhibit significantly lower levels of matching (Fig. 2, Table 1, Fig. S2A and Fig. S2B).

Interestingly, when expanding the set by including organellar and extracellular proteins, or even cytoplasmic fragments of transmembrane proteins, a similar trend of matching properties is observed (Table 1). For example, the HFE normalized by sequence length remains well matched, regardless of the set increasing in size from 118 to 162 or 268 proteins (Fig. 3A). For the complete data set, in fact, the statistical significance of intra-pair matching for HFE/N reaches a maximum with a p-value of 3×10^{-6} (Table 1). Even more surprisingly, when these newly included proteins are analyzed separately, one still observes significant matching in their size-normalized HFEs (Table 1).

Even though hydration free energies, when normalized by radii of gyration, do not seem to be significantly matched between the partners regardless of the data set used (Table 1, Fig. S3B), we noticed that for small proteins (where both of the partners have less than 130 residues), this ratio seems to be extremely well correlated (ICCs above 0.9) between the members of the pair (Fig. S6). This type of size-filtering resulted in either 24 complexes from the set of all intracellular proteins (subset of 81 pairs), or 28 complexes when extracted from the maximal set (134 pairs).

Overall, the necessity for normalizing the HFEs by a size-dependent term is further emphasized if one examines the behavior of HFE/N^α for a range of exponents α (Fig. 3A). Clearly, the most significant match between the partners is observed only for a narrow range of such exponents, surrounding 1 (cca. 0.5 – 1.2). The most obvious rationale for

normalization of HFE by a size-dependent term is to adjust for missing residues. However, if one looks at those complexes in our data set where both partners are complete (17 complexes in total), one sees no significant matching for size-normalized HFE, weakening this argument (Fig. S7). Future research should elucidate a rigorous physical basis for matching of size-normalized HFE. Interestingly, a similar strategy is used in prediction of protein retention times in hydrophobic interaction chromatography, where hydrophobicity is normalized by solvent-accessible surface area [20,21], also a size-dependent variable. Finally, analysis of isoelectric points and charges at neutral pH estimated from primary sequences did not reveal any statistically significant trends (Fig. S2A and Fig. S2B), except when charge is normalized by N (p-values of 0.01, Fig. S2C and Fig. S2D). We are not excluding the possibility that charge might be an additional factor which acts synergystically with hydrophilicity as an important guiding force in bringing the binding partners together, especially given the fact that electrostatic interactions are responsible for the dominant part of the hydration free energy, i. e. its polar term.

It is possible that the observed matched properties, such as the size-normalized HFE, are significantly influenced by the properties of the binding sites themselves, which in turn, almost by definition have certain properties in common, such as the solvent-accessible surface area. To exclude this possibility, we calculated the fraction of atoms that form the binding-site interface for each protein. Given the fact that the size of the interface for the majority of proteins used in this analysis is below 10 % of the total number of atoms (Fig. 4), we assume that the contribution of the interface itself to the calculated properties is not responsible for the correlations observed. On the other hand, it is possible that the observed matching is a consequence of the experimental treatment through which proteins were

isolated, purified, and crystallized. Currently, it is not possible to fully discount this possibility, exciting in its own right.

With a recent report showing that the localization of a bacterial protein is determined by a geometric factor [22,23], namely, membrane curvature, the importance of assessing potential contribution of other non-specific properties to protein localization *in vivo* is additionally emphasized. Proteins are physicochemical entities, and the fact that their localization and interactions are exclusively determined by the particulars of the specific binding sites (to either their partners or anchoring elements such as cytoskeleton), as typically assumed [24], needs to be rigorously tested. The results presented here indicate that binding partners in different transient functional complexes have certain general physicochemical properties in common, which could then be responsible for their colocalization or clustering on the microscopic level, and thus indirectly facilitate their binding. Our results suggest that size-normalized HFE may be one such property, and allow us to propose the hydrophilicity matching hypothesis, where hydrophilic gradients, almost as in chromatographic separation [20,21], serve as an organizing force for the localization of proteins, even within individual compartments such as the cytoplasm.

It is our belief that *protein ecology* – where a given protein is located, and who and for what reasons its molecular neighbours are, even within individual compartments – may be an important frontier to study. Should it really turn out that the non-specific component of protein-protein interactions is functionally relevant, and therefore also under evolutionary control, this would represent a major paradigm shift, and would carry important implications on how we view biological systems on the molecular level or try to affect them in practical

situations. For example, most drug design applications almost exclusively target the specific component of protein-protein interactions. Should the non-specific component prove to be relevant, it would also present itself as a completely novel, orthogonal pharmacological target.

Materials and Methods

Dataset. The DOCKGROUND database [15] used contains either experimentally determined structures of the binding partners in their unbound form (when available), or the ones that are computationally modelled based on bound complexes. The starting set of 151 binary complexes obtained by excluding all members of the database (release of 8th July 2008) with missing atoms anywhere in the backbone, was reduced to 134 after all non-physiological complexes (antibody-antigen complexes that do not exist *in vivo* or artificially created proteins) were excluded. The completeness of each of the partners was determined by taking the ratio of the number of residues of a given protein in the cocrystallized complex and the length of the native protein as reported in the UniProt database. Signal peptides, as defined within the UniProt database, were excluded when calculating the completeness in those cases where they were present. Structures were considered to be complete if 3 % of the residues or less were missing in the cocrystallized complex. Localization of proteins was determined by an exhaustive literature research combined with the information available in the UniProt database. Localization of proteins based on where the encounter with their respective partners takes place was found to differ in some cases from their general localization as reported in UniProt or other databases. Viral proteins were assigned origin and grouped based on the characteristics of their interacting partner.

Calculation of physicochemical properties. The volume of interacting partners (vol) was calculated using 3v: Voss Volume Voxelator [25], while the hydration free energy (HFE), electrostatic energy (EE), solvent-accessible surface area (SASA), and radius of gyration (R_{gyr}) were calculated using TINKER molecular modeling package [26], following additional short relaxation MD simulations of each protein (15 ps of simulated time per protein – average values of properties were obtained from an ensemble of 100 structures in the last 10 ps of the run). Simulations were run in implicit GB/SA solvent with Langevin dynamics at 300 K, using OPLSaa force field [27] with no cutoffs for electrostatics, and friction coefficient of γ = 91 ps^{-1}. HFE was calculated using GB/SA methodology [18,19] with ε_{water} = 81. Isoelectric point values and charges at neutral pH were estimated using web-based Protein Calculator v3.3 (http://www.scripps.edu/~cdputnam/protcalc.html). The fraction of the atoms that form the interface between the partners when in their bound state was determined by counting atoms of each of the partners whose distance was smaller than the sum of their respective van der Waals radii plus an arbitrary value of 0.5 Å. Van der Waals radii used are as follows: r_{vdW}(C) = 1.7 Å, r_{vdW}(N) = 1.55 Å, r_{vdW}(O) = 1.52 Å, r_{vdW}(S) = 1.8 Å [28], r_{vdW}(H) = 1.09 Å [29].

Statistics. Intraclass correlation coefficients (ICCs) for a particular property were determined as previously reported [16,17]. Intraclass correlation is a standard statistical test for quantifying the extent to which the members of a given group resemble each other in terms of a certain property. For paired data sets where there is no meaningful way of ordering members of a given pair (such as properties of twins, for instance), ICC represents a more natural measure of association than the Pearson correlation coefficient (R), which is typically

reserved for those cases where there is a clear distinction between dependent and independent variables. In order to illustrate this difference, average Pearson correlation coefficient estimates for various properties between binding partners in different analyzed subsets are additionally discussed in Supporting Information (Fig. S1 and Table S2).

For a paired data set comprised of N pairs,

$$\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ ... & ... \\ y_{N1} & y_{N2} \end{bmatrix}$$

the group mean $\overline{y}_i$, the total mean $\overline{y}$, the variance between the groups σ_b^2 and the variance within the groups σ_w^2 are given as

$$\overline{y}_i = \frac{1}{2}(y_{i1} + y_{i2}) \qquad i = 1,..., N \tag{1}$$

$$\overline{y} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{2} y_{ij} \tag{2}$$

$$\sigma_b^2 = \frac{1}{N} \sum_{i=1}^{N} (\overline{y}_i - \overline{y})^2 \tag{3}$$

$$\sigma_w^2 = \frac{1}{2N} \sum_{i=1}^{N} ((y_{i1} - \overline{y}_i)^2 + (y_{i2} - \overline{y}_i)^2) \tag{4}$$

respectively. Then, the corresponding ICC (η^2) is defined as:

$$\eta^2 = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \tag{5}$$

ICC captures the relation between the average variance within pairs and the total variance between pairs. The significance of the observed ICCs was assessed via randomization tests in which proteins within a given set were paired completely at random, to obtain a same-size, randomized set of pairs whose ICC value was then calculated. After 10^6 such randomizations, the associated p-values were estimated by calculating the fraction of times an ICC value that

is greater than or equal to the one for the native data set appeared in the distribution of ICCs for randomized sets.

Abbreviations

HFE – hydration free energy, EE – electrostatic energy, SASA – solvent-accessible surface area, R_{gyr} – radius of gyration, N – number of residues, GB/SA – generalized Born/surface area, MD – molecular dynamics, ICC – intraclass correlation coefficient

Author Contributions

BZ and MH designed and performed research; MH, GZ and BZ analyzed data; MH and BZ wrote the paper.

References

1. Berg JM, Stryer L, Tymoczko JL (2006) Biochemistry. New York: W. H. Freeman & Co.

2. Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F et al. (2008) Prediction of aggregation-prone regions in structured proteins. J Mol Biol 380: 425-436.

3. Vendruscolo M, Tartaglia GG (2008) Towards quantitative predictions in cell biology using chemical properties of proteins. Mol Biosyst 4: 1170-1175.

4. Niwa T, Ying BW, Saito K, Jin W, Takada S et al. (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. Proc Natl Acad Sci USA 106: 4201-4206.

5. Gabdoulline RR, Wade RC (2002) Biomolecular diffusional association. Curr Opin Struct Biol 12: 204-213.

6. Ridgway D, Broderick G, Lopez-Campistrous A, Ru'aini M, Winter P et al. (2008) Coarse-grained molecular simulation of diffusion and reaction kinetics in a crowded virtual cytoplasm. Biophys J 94: 3748-3759.

7. Wieczorek G, Zielenkiewicz P (2008) Influence of macromolecular crowding on protein-protein association rates – a Brownian dynamics study. Biophys J 95: 5030- 5036.

8. Batada NN, Shepp LA, Siegmund DO (2004) Stochastic model of protein-protein interaction: Why signaling proteins need to be colocalized. Proc Natl Acad Sci U S A 101: 6445-6449.

9. Batada NN, Shepp LA, Siegmund DO, Levitt M (2006) Spatial regulation and the rate of signal transduction activation. PLoS Comput Biol 2: e44.

10. Berg HC (1993) Random walks in biology. Princeton: Princeton University Press.

11. Bickmore WA, Sutherland HGE (2002) Addressing protein localization within the nucleus. EMBO J 21: 1248-1254.

12. Sutherland HGE, Mumford GK, Newton K, Ford LV, Farrall R et al. (2001) Large-scale identification of mammalian proteins localized to nuclear sub-compartments. Hum Mol Genet 10: 1995-2011.

13. Du PF, Li YD (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics 7, doi:10.1186/1471-2105-7-518.

14. Kundrotas PJ, Alexov E (2006) Electrostatic properties of protein-protein complexes. Biophys J 91: 1724-1736.

15. Gao Y, Douguet D, Tovchigrechko A, Vakser IA (2007) DOCKGROUND system of databases for protein recognition studies: Unbound structures for docking. Proteins 69: 845-851.

16. Kotz S, Johnson NL, Read CB, editors. (1983) Intraclass correlation coefficient. In: Encyclopedia of statistical sciences. New York: Wiley. vol. 4, pp. 212-217.

17. Shrout PE, Fleiss JL (1979) Intraclass correlations – uses in assessing rater reliability. Psychol Bull 86: 420-428.

18. Qiu D, Shenkin PS, Hollinger FP, Still WC (1997) The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. J Phys Chem A 101: 3005-3014.

19. Chen JH, Brooks CL, Khandogin J (2008) Recent advances in implicit solvent-based methods for biomolecular simulations. Curr Opin Struct Biol 18: 140-148.

20. Lienqueo ME, Mahn A, Asenjo JA (2002) Mathematical correlations for predicting protein retention times in hydrophobic interaction chromatography. J Chromatogr A 978: 71-79.

21. Mahn A, Lienqueo ME, Salgado JC (2009) Methods of calculating protein hydrophobicity and their application in developing correlations to predict hydrophobic interaction chromatography retention. J Chromatogr A 1216: 1838-1844.

22. Ramamurthi KS, Lecuyer S, Stone HA, Losick R (2009) Geometric cue for protein localization in a bacterium. Science 323: 1354-1357.

23. Ramamurthi KS, Losick R (2009) Negative membrane curvature as a cue for subcellular localization of a bacterial protein. Proc Natl Acad Sci USA 106: 13541-13545.

24. Agutter PS, Wheatley DN (2000) Random walks and cell size. Bioessays 22: 1018-1023.

25. Voss NR, Gerstein M, Steitz TA, Moore PB (2006) The geometry of the ribosomal polypeptide exit tunnel. J Mol Biol 360: 893-906.

26. Ponder JW, Case DA (2003) Force fields for protein simulations. Adv Protein Chem 66: 27-85.

27. Jorgensen WL, Maxwell DS, TiradoRives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118: 11225-11236.

28. Bondi A (1964) Van der Waals volumes + radii. J Phys Chem 68: 441-452.

29. Rowland RS, Taylor R (1996) Intermolecular nonbonded contact distances in organic crystal structures: Comparison with distances expected from van der Waals radii. J Phys Chem 100: 7384-7391.

30. Fink AL (2005) Natively unfolded proteins. Curr Opin Struct Biol 15: 35-41.

Figure legends

Figure 1. Comparison of ICCs calculated for naturally occuring binding partners and those obtained by a randomization procedure. The results are for a subset of 118 eukaryotic proteins (59 pairs) that interact in the cytoplasm or nucleoplasm. The ICC values were calculated for (A) the sequence length of the binding partners (N), (B) their radius of gyration (R_{gyr}), (C) hydration free energy (HFE), and (D) HFE normalized by sequence length (HFE/N). Red arrow denotes the value of the observed ICC for the known binding partners.

Figure 2. Summary of the calculated ICCs and their associated p-values for various properties. The results are for a subset of 118 eukaryotic proteins (59 pairs) that interact in the cytoplasm or nucleoplasm. We show the results for amino acid sequence length (N), volume (vol), radius of gyration (R_{gyr}), solvent-accessible surface area (SASA), hydration free energy (HFE), and electrostatic energy (EE), or selected ratios thereof.

Figure 3. ICCs and p-values for the HFE/N$^\alpha$ ratio for the maximal set of 268 proteins (134 pairs). This set includes organellar, secreted, and extra- or intracellular fragments of transmembrane proteins. The values are plotted as a function of (A) the subset size, where the pairs were ordered by their maximal length (maximal N within a given pair) for $\alpha = 1$, and (B) the exponent α, with the scan performed in steps of 1/6. Stars denote the p-values < 0.0001.

Figure 4. Fraction of atoms comprising the interface between each of the partners. Size of the interface as a function of the sequence length of partners (N) is shown for the maximal set of 268 proteins (134 pairs).
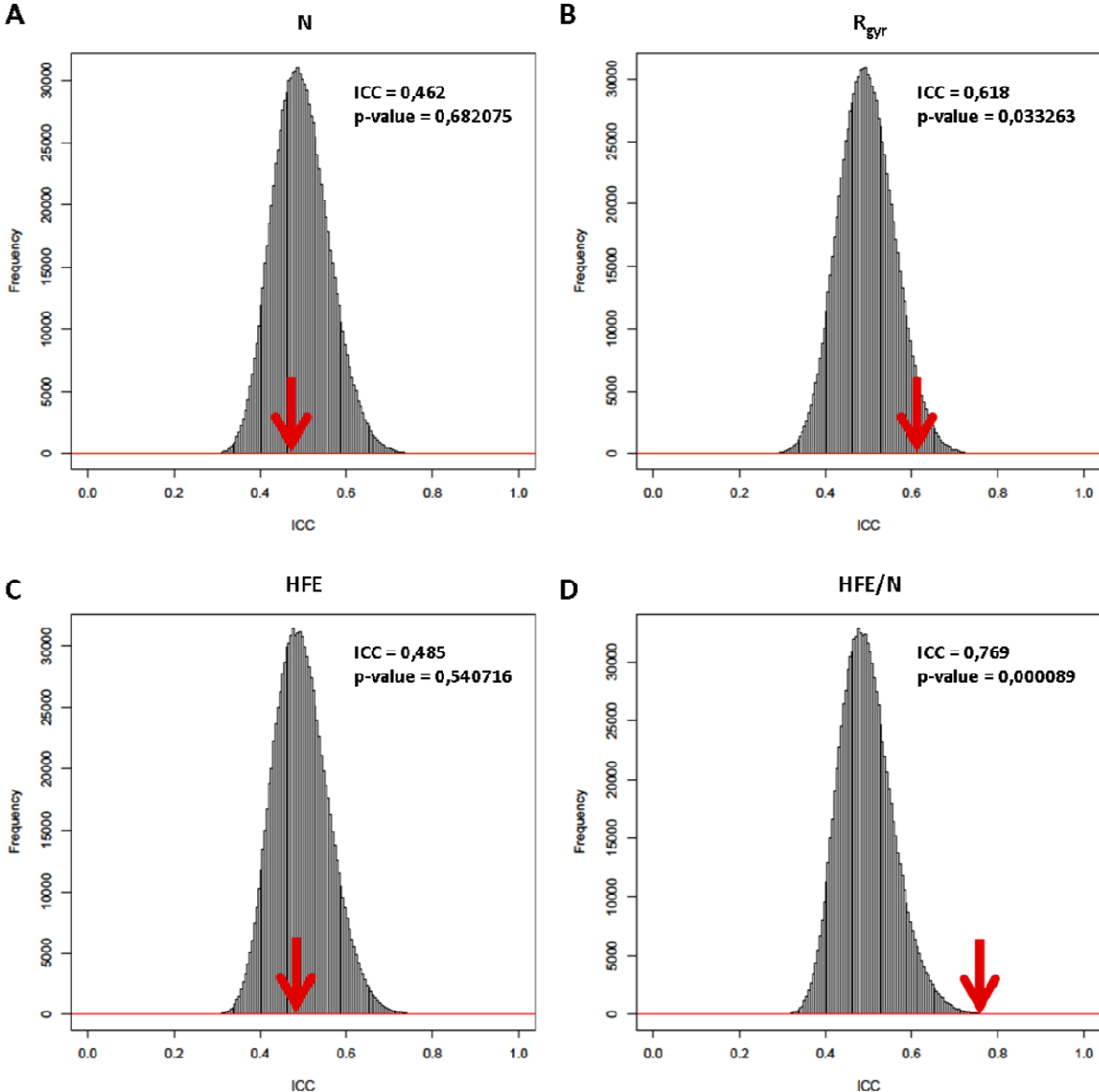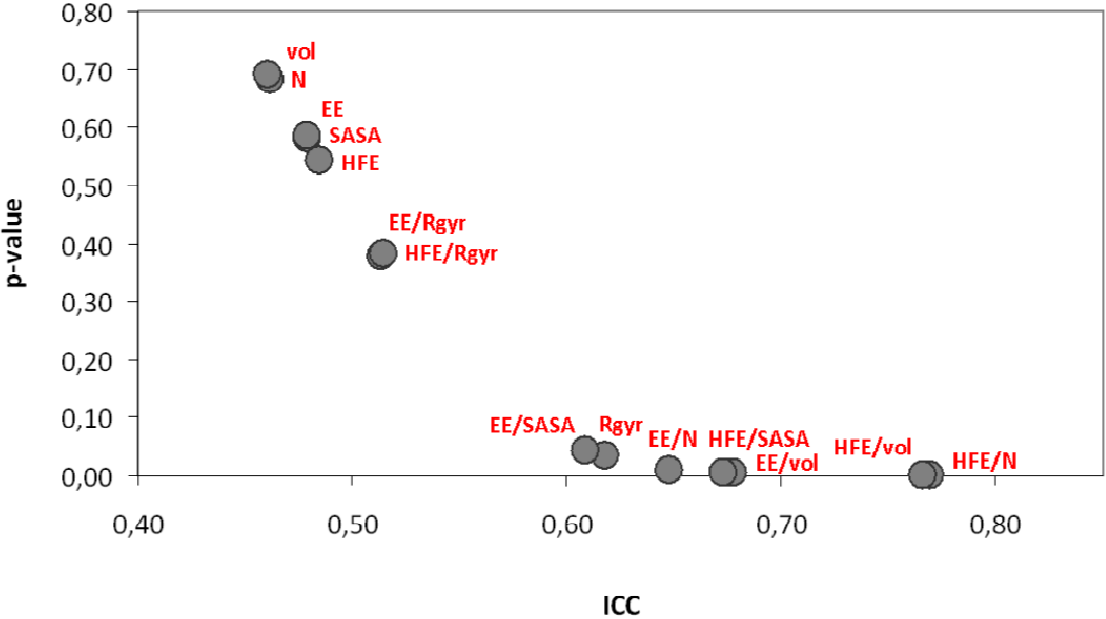
Figures

Figure 1.

A

N

ICC = 0,462
p-value = 0,682075

B

R$_{gyr}$

ICC = 0,618
p-value = 0,033263

C

HFE

ICC = 0,485
p-value = 0,540716

D

HFE/N

ICC = 0,769
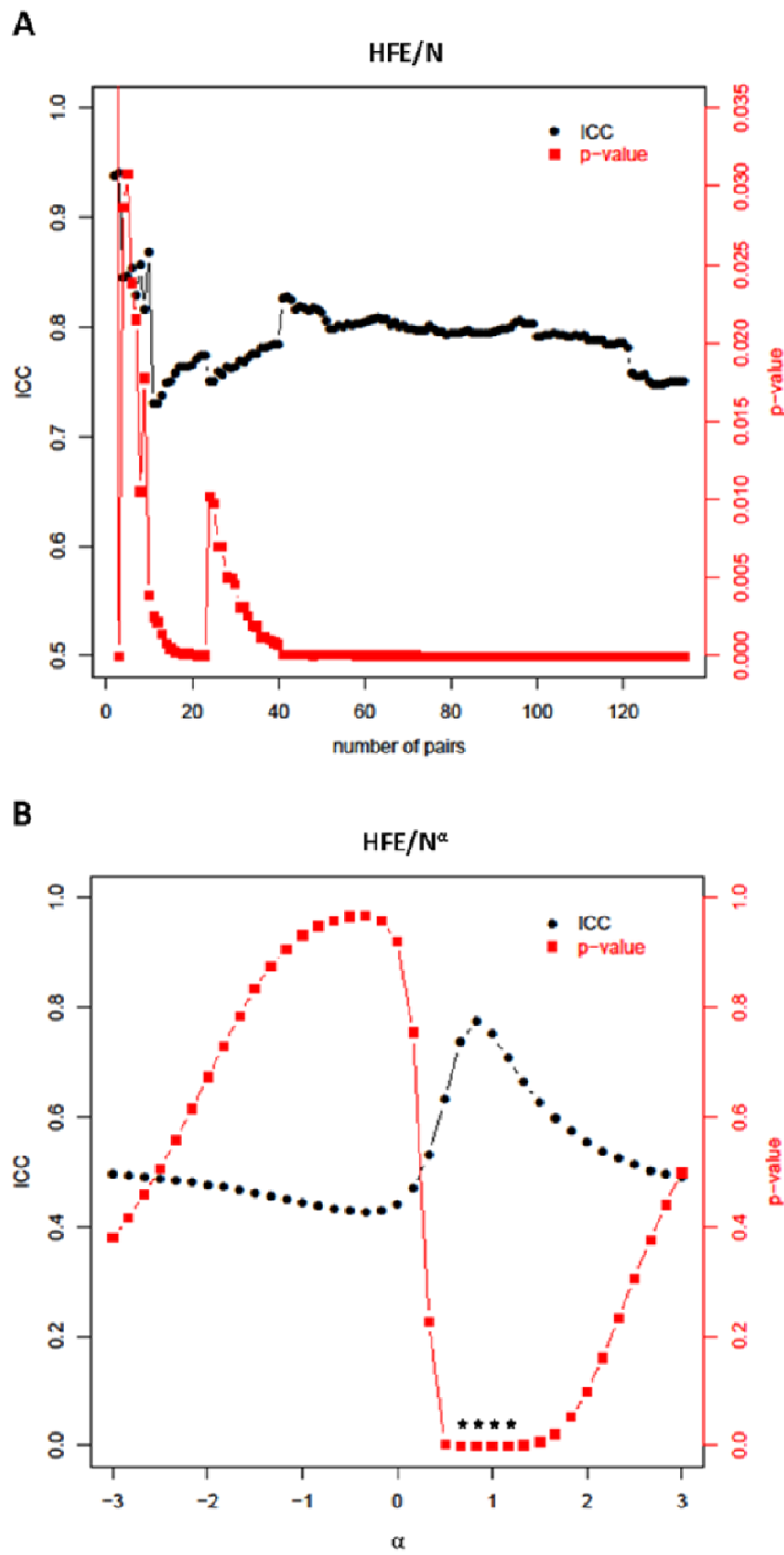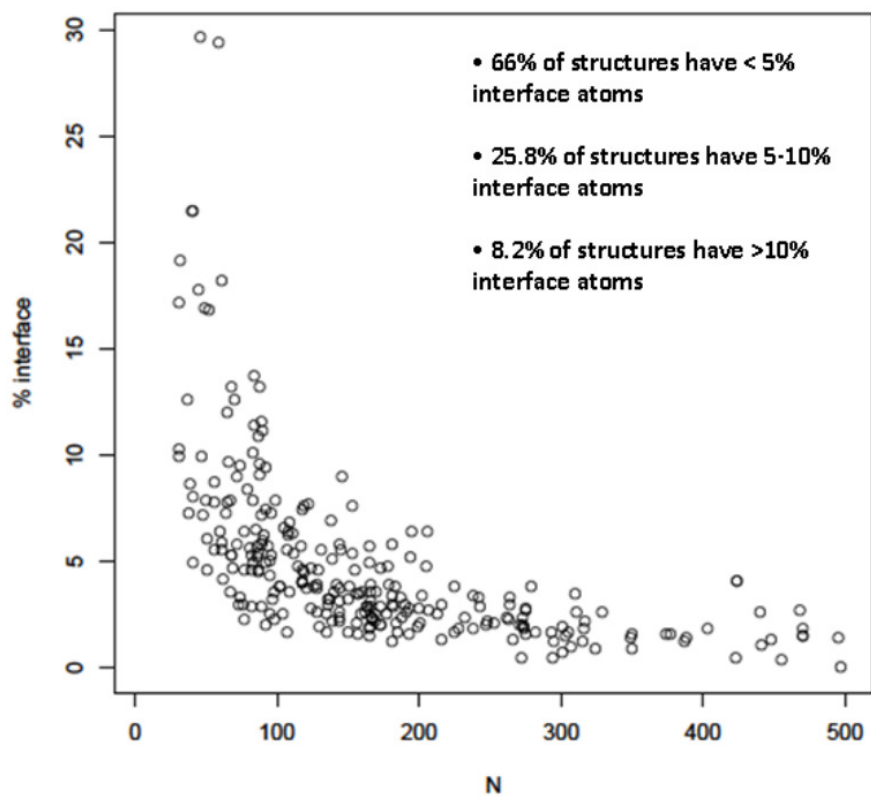p-value = 0,000089

Figure 2.

Figure 3.

Figure 4.

Tables

Table 1. Summarized results of the degree of similarity of known binding partners for various properties within different subsets.

compared property	59 pairs[*]		81 pair[†]		53 pairs[‡]		134 pairs[§]	
	ICC	p-value	ICC	p-value	ICC	p-value	ICC	p-value
N	0,462	0,682075	0,412	0,94576	0,418	0,877154	0,439	0,919747
R_{gyr}	0,618	0,033263	0,526	0,286742	0,405	0,912081	0,481	0,642724
SASA	0,48	0,581557	0,403	0,965165	0,417	0,883819	0,42	0,972538
vol	0,461	0,688965	0,409	0,952544	0,411	0,900043	0,431	0,948579
HFE	0,485	0,540716	0,452	0,788113	0,419	0,875998	0,439	0,921591
EE	0,479	0,585556	0,43	0,892154	0,46	0,687569	0,467	0,763852
HFE/N	**0,769**	**0,000089**	**0,731**	**0,000305**	0,678	0,005197	**0,751**	**0,000003**
HFE/R_{gyr}	0,513	0,378304	0,502	0,453216	0,477	0,60472	0,503	0,452854
HFE/SASA	0,678	0,004697	0,678	0,003423	0,563	0,159121	**0,679**	**0,000214**
HFE/vol	**0,766**	**0,000194**	0,714	0,001405	0,543	0,126554	0,666	0,004269
EE/N	0,648	0,008287	0,686	0,00213	0,678	0,006919	**0,684**	**0,000059**
EE/R_{gyr}	0,515	0,381677	0,488	0,563313	0,507	0,423722	0,515	0,344302
EE/SASA	0,609	0,040245	0,639	0,004634	0,562	0,165748	**0,631**	**0,000927**
EE/vol	0,673	0,002747	**0,705**	**0,000955**	0,572	0,033874	0,613	0,004257

[*]eukaryotic intracellular (nuclear and cytosolic) complexes, [†]archeal, bacterial and eukaryotic intracellular (nuclear and cytosolic) complexes (includes the entire subset of 59 binary complexes), [‡]archeal, bacterial and eukaryotic extracellular complexes, or intracellular complexes of organellar proteins or segments of transmembrane proteins, [§]maximal set comprised of [†] and [‡]; the p-values < 0.001 are bolded