Scientific Paper

FAST ESTIMATION OF SURFACE COMPLEMENTARITY IN PROTEIN COMPLEXES

Giacomo Franzot[a] and Oliviero Carugo[b]

[a] *International School for Advanced Studies, via Beirut 4, 34014 Trieste, Italy and Sincrotrone Trieste, Strada Statale 14 - km 163,5 in AREA Science Park 34012 Basovizza, Trieste, Italy*
[b] *Department of General Chemistry, Pavia University, viale Taramelli 12, 27100 Pavia, Italy and TASC-INFM National Laboratory, Strada Statale 14 - km 163,5 in AREA Science Park 34012 Basovizza, Trieste, Italy*

Received 22-12-2003

Abstract

A novel measure of protein surface complementarity, sc_pride, is proposed. Each surface patch is represented by the distribution of the inter-atomic distances and the degree of similarity between two surface patches is estimated via a contingency table analysis of their two inter-atomic distance distributions. Such a low resolution surface representation allows very fast complementarity estimations that could find applications in protein-protein interaction prediction. The performance of sc_pride is compared to that of other surface complementarity measures with a very large set of protein-protein complexes obtained with docking simulations and the ability of sc_pride to recognize the surface complementarity is tested on a non-redundant set of experimentally determined crystal structures of protein-protein complexes.

Key words: protein structure, protein-protein interaction, protein surface

Introduction

The interactions between proteins, with consequent formation of interaction networks, are of fundamental importance in modern molecular biology.[1-4] The life depends in fact on the ability of each protein to correctly recognize its partners, which in turn recognize other proteins. The shape complementarity is a mandatory requirement in protein recognition. Several algorithms for estimating the surface complementarity have so far been developed.[2] The oldest were based on very detailed stereochemical descriptions of the protein surfaces. Later on, an increasing attention was devoted to the inclusion of the intrinsic molecular flexibility into the description of surface geometry.[5,6] In other words, a low resolution portrayal of the protein surface might allow one to overcome the problem of predicting the conformational rearrangements consequent to the inter-molecular association.

The major drawback of the methods for estimating the surface complementarity depends on the fact they are usually associated with docking simulations. These computational procedures, aimed to predict the stereochemistry of protein complexes, are very slow because of the comlex task of analysisng the immense conformational space that includes both the relative orientation of the interactiong partners and their stereochemical flexibility.[5,6] Moreover, within a docking simulation, the surface complementarity can be estimated only if the relative position of the interacting partners has been hypothesized.

In the present paper we present a new description of the protein surface geometry based on the distribution of the inter-atomic distances, an approach reminiscent of the extremety fast fold comparison procedure implemented in PRIDE.[7,8] The two surface patches that must be compared are represented by the two distributions of their inter-atomic distances which are then compared through a contingency table analysis,[9] resulting in the surface complementarity score sc_pride (see Experimental section for details). Such a surface representation is intrinsically a low resolution description of the surface geometry because the possible atomic displacements due to the complex formation modify some of the inter-atomic distances but do not influence to a great extent the distance distribution. Moreover, in such a representation, the geometric description becomes independent of the 3D structure of the protein-protein complex. Each single monomeric protein surface can be described independently of the position of the other protein partner. Such a surface geometry description can thus be used in computational approaches that partition the protein surface into adjacent patches, like for example PUZZLE.[10]

Given its computational simplicity and speed and given that it does not need any assumption on the relative position of the interacting partners, such a novel procedure could therefore be of extreme importance in large scale virtual screening studies.

Results and discussion

Comparison between sc_pride and other measures of surface complementarity

A very large data set of protein-protein complex 3D structures was obtained by the computational docking simulations summarized in Table 1. Each of the theoretical

models was assigned the sc_pride values together with the FADE, SC, and scscore values.

FADE (Fast Atomic Density Evaluator) values measure the shape complementarity for docked complexes.[11] Each surface is described as a series of contiguous grooves and protruding regions through a fractal atomic density index.[12] The latter is the slope of the relationship between log(N) and log(r), where N is the number of atomic centers within a sphere of radius r centered on a dot of a Connolly molecular accessible surface.[13] High indices are associated with deep grooves and low values are associated with protruding regions, remembering alternative definitions of protrusion at the protein surface.[14] The complementarity of the protrusion degree of neighbors surface patches result in a FADE value that is inversely proportional to the protein-protein surface complementarity.

SC values are an alternative measure of surface complementarity.[15] They depend on the relative orientation of two unit vectors, one outwardly oriented and normal to the molecular accessible surface of a protein, and the other, inwardly oriented and normal to the surface of the other protein. The first unit vector originates from any point P of the surface of the first protein. The second vector starts at the point of the surface of the second protein that is closest to P. If the two surfaces are parallel around P, the two vectors are also parallel and their scalar product reaches its maximum possible value. The 50^{th} percentile of these scalar products that span all the points P of each surface is assumed to measure the surface complementarity at the protein-protein interface. Large SC values are associated with highly complementary surface patches.

While both FADE and SC values depend on the molecular accessible surfaces, the computational docking software suite 3D-Dock[16] provides an alternative definition of surface complementarity. One of the two proteins, the complexation of which is simulated, is roto-translated around the other through the algorithm of Katchalski-

Katzir[17] and the surface complementarity is computed, after each roto-translation, by grid discretisation of the molecules. Core overlaps between grids are penalized while surface overlaps represent a positive contribution to the protein-protein recognition. The resulting scscore values are proportional to the degree of complementarity.

1,000 theoretical models were randomly selected from each of the 16 docking simulations. Each of the 16,000 protein-protein complexes was given the FADE, SC, scscore, and sc_pride values. Table 2 shows the linear correlation coefficients between

Table 1. Protein-protein complexes used in docking simulations. For each protein in each docking simulation the following information is provided: the PDB identification code (Idcode), the chain identifier (Chain), the protein name (Protein), and the biological source (Source).

Idcode	Chain	Protein	Source
1a0o	A	Chea	Escherichia Coli
1a0o	B	Chey	*Escherichia Coli*
1a2k	A	Nuclear Transport Factor 2	*Rattus nurvegicus*
1a2k	D	Ran, Gsp1P	*Canis familiaris*
1a4y	A	Angiogenin	*Homo sapiens*
1a4y	B	Ribonuclease inhibitor	*Homo sapiens*
1an1	E	Trypsin	*Sus scrofa*
1an1	I	Trypsin inhibitor	*Hirudo medicinalis*
1b2s	A	Barnase	*Bacillus amyloliquefaciens*
1b2s	D	Barstar	*Bacillus amyloliquefaciens*
1c1y	A	Ras binding	Homo sapiens
1c1y	B	Rap-1°	*Homo sapiens*
1clv	A	α-Amylase	*Tenebrio molitor*
1clv	I	α-Amlylase inhibitor	*Amaranthys Hypochondriacus*
1dpj	A	Proteinase A	*Saccharomyces cerevisiae*
1dpj	B	Proteinase inhibitor	*Saccharomyces cerevisiae*
1fc2	D	Immunoglobulin Fc	*Staphylococcus aureus*
1fc2	C	Fragment B of protein A	*Homo sapiens*
1fle	E	Elastase	*Sus scrofa*
1fle	I	Elafin	Homo sapiens
1jat	A	Ubiquitin Conjugating enzyme E2	*Saccharomyces cerevisiae*
1jat	B	Ubiquitin Conjugating enzyme Mms2	*Saccharomyces cerevisiae*
1jhl	H	Antibody D11.15	Mus musculus
1jhl	A	Lysozyme	*Phasianus colchicus*
1mee	A	Serine proteinase	*Bacillus pumilus*
1mee	I	Eglin C	*Hirudo medicinalis*
1tx4	A	Rho gap	*Homo sapiens*
1tx4	B	Rho a	*Homo sapiens*
1ugh	E	Uracil-DNA Glycosylase	*Homo sapiens*
1ugh	I	Glycosylase Inhibitor	*Bacteriophage PBS2*
2jel	H	Jel42 Fab Fragment	*Mus musculus*
2jel	P	His-Containing Protein	*Escherichia coli*

Table 2. Linear correlation coefficients between various surface complementarity scores. The average values, with standard deviations in parentheses, were computed on 16 sets of 1,000 theoretical models obtained with the 3D-Dock software suite.

	FADE	SC	scscore
SC	-0.249 (0.017)		
scscore	-0.166 (0.018)	0.035 (0.010)	
sc_pride	-0.159 (0.018)	0.066 (0.018)	0.036 (0.012)

these four measures of surface complementarity. Figure 1 shows the dependence on sc_pride of the FADE, SC, and scscore values. The correlation coefficients are very small, though statistically different from zero. As expected, the FADE values are inversely proportional to the three other scores and the latter ones are all positively correlated. The sc_pride values correlate with the other scores as it must be expected. They increase as the SC and scscore values increase and the decrease as the FADE values increases. The discrepancy between various shape scoring functions is quite surprising and has never been described and commented previously. It must nevertheless be observed that the protein-protein complexes examined here are produced by rigid body docking simulations. Conformational rearrangements, caused by the complexation, are thus not considered. This might account also for the fact that the SC values computed over the 16,000 three-dimensional models are relatively smaller (Figure 1) than those reported for real protein-protein complexes, which are around 0.6 or higher.

Dependence of sc_pride on the interface dimension

In order to compute sc_pride values, protein surface patches are described by the distributions of their inter-atomic distances. The information provided by these distributions is obviously dependent on the dimension of the surface patch. For example, the smallest patches containing only one or two atoms would be identical to any other patch. At the other extreme, a very large patch containing many atoms could be associated with inter-atomic distances uniformly distributed and thus it would be impossible to discriminate similar from dissimilar pairs of surface moieties. Sc_pride values were computed for the ensemble of surface patches of the proteins listed in Table 3 and shown in Figure 2. These were selected because they are very different one from each other. 1bz6 is a classical compact globin fold, 1cdm and 4cln are calmodulins but while 1cdm is in the bent conformation, adopted in the presence of the substrate (not shown in the figure), 4cln is in the extended conformation, 4aah is a beta-propellor, and 1qsa is a U-shaped alpha-super-helical domain.

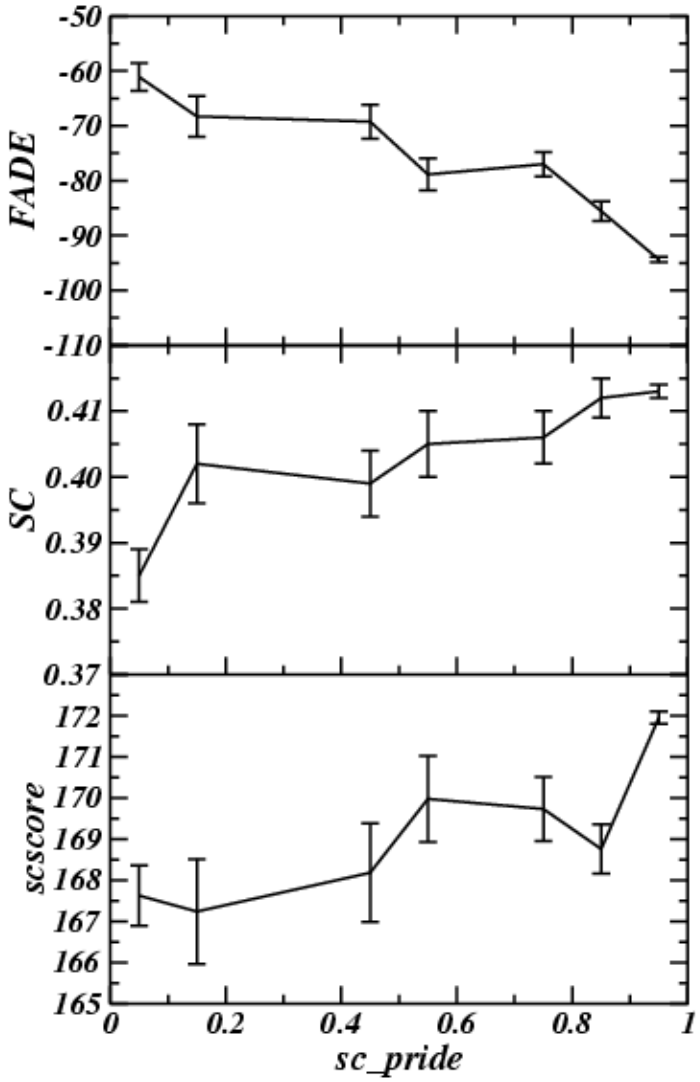

Figure 1. Dependence on sc_pride of three different measures of surface complementarity (FADE, SC, and scscore). Standard deviations are indicated by vertical bars. The data were obtained from a set of 16,000 protein-protein complex structures simulated by computational docking.

Table 3. Protein structures used to analyze the dependence of the sc_pride values on the dimension of the surface patches that are compared. For each protein the following information is provided: the PDB identification code (Idcode), the chain identifier (Chain), the protein name (Protein), and the biological source (Source).

Idcode	Chain	Protein	Source
1bz6	A	Myoglobin	*Physter catodon*
1cdm	A	Calmodulin	*Bos taurus*
1qsa	A	Transglycosylase Slt70	*Escherichia coli*
4aah	A	Methanol dehydrogenase	*Methylophilus methylotrophus W3A1*
4cln		Calmodulin	Drosophila melanogaster

For each protein, *n* surface patches containing the *m* (20<*m*<90) solvent exposed atoms closest to each of the *n* exposed atoms were built. Smaller ensembles of solvent exposed atoms were disregarded because of their little practical relevance to the problem of protein-protein interaction. Each surface patch of each protein was compared with each surface patch of all the other proteins. A total of 8,512,472 comparisons were performed. The average sc_pride values are plotted in Figure 3 against the patch dimension. It appears that for large patches, the sc_pride values tend, on average, to approach their maximal value (1.0). Lower values are observed, on average, for smaller patches. This observation suggests that the shape of large surface patches cannot be monitored effectively by the sc_pride values. The reason of this performance is presumably related to the fact that when a patch contains many atoms there are so many inter-atomic distances that their distribution becomes rather similar to that of any other large surface patch, independently of the real shape difference. This might also account for the small correlation coefficients that are obtained by comparing sc_pride with other measures of surface complementarity.

It must be concluded therefore that the use of sc_pride must be limited to the analysis of surface patches not larger then approximately 40 atoms.

Discrimination between interface and non-interface surface patches

The possibility to use sc_pride in order to distinguish surface patches that are at the protein-protein interface, and therefore have really complementary shapes, from patches outside the recognition sites, has been investigated by analyzing the protein-protein complexes shown in Table 4. An interesting feature of these complexes is that the three-dimensional structures of both the uncomplexed and the complexed proteins are available. For each protein, the surface patches containing the 20 atoms closest to each solvent exposed atom were built and classified according to the fraction of interface atoms they contain. The dependence of the sc_pride values on the type of patch pair in shown in Figure 4. It appears that higher sc_pride values are observed, on average, when the two patches that are compared contain a large fraction of the atoms that are actually at the protein-protein interface.

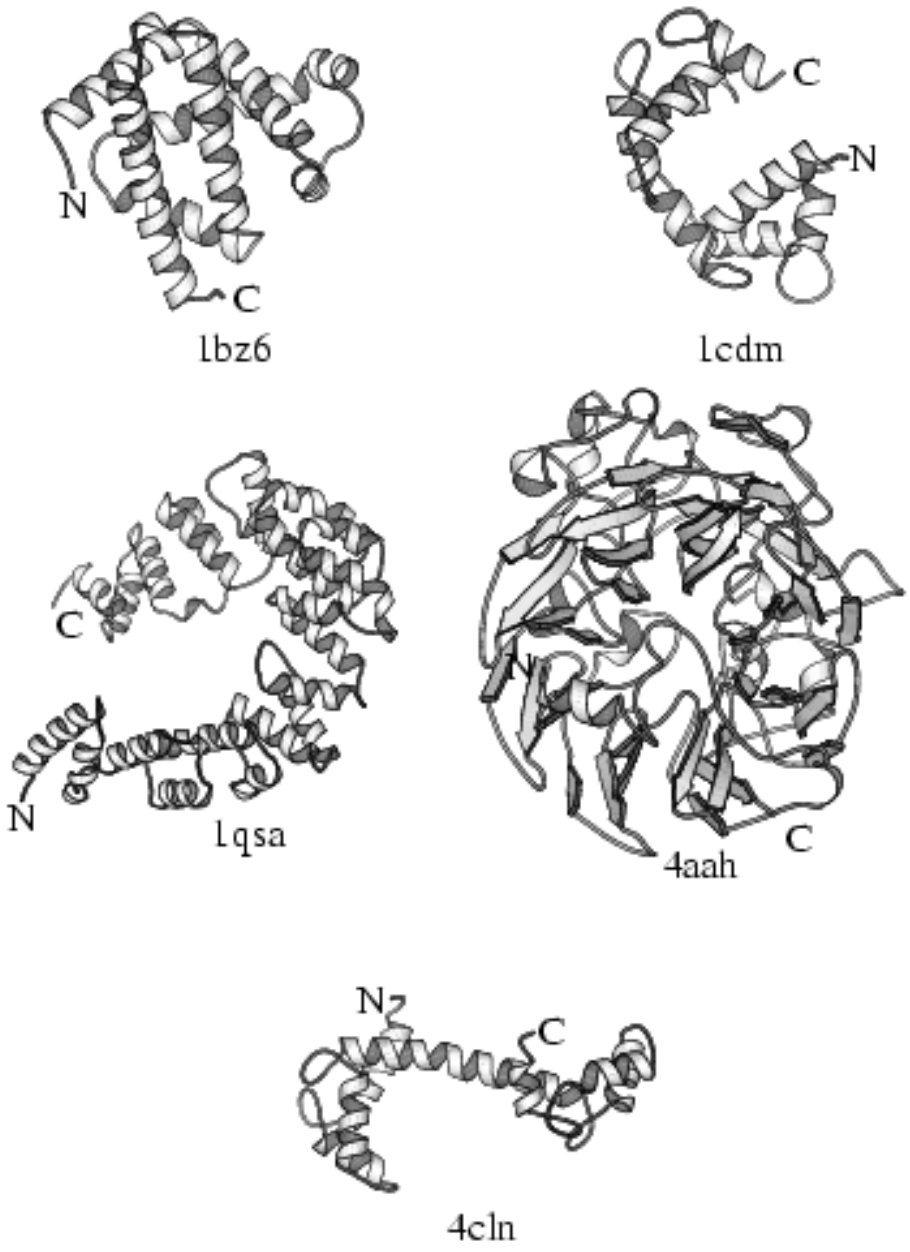

Figure 2. Molscript[20] views of the protein structures used to analyze the dependence of the sc_pride values on the dimension of the surface patches that are compared.

This does not depend on the fact that the patches that are compared are taken from the structures of the two bound or unbound partners, indicating that sc_pride is rather unaffected by small conformational rearrangements that may take place at the protein surface as a consequence of the complexation.

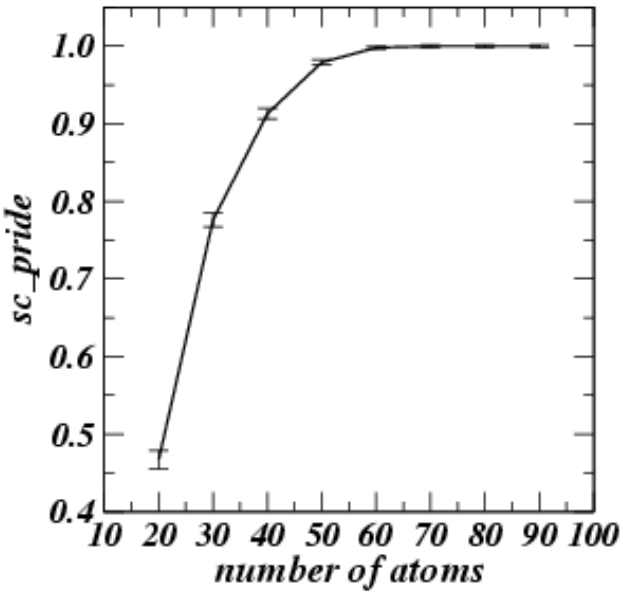

Figure 3. Dependence of the average sc_pride values on the dimension of the surface patches that are compared. Standard deviations of the mean are indicated by vertical bars.

Conclusions

A new measure of surface complementarity, sc_pride, has been designed. It agrees with other complementarity definitions and detects the geometric complementarity in stable, real complexes. Its main advantage is its computational speed. Like in the fold comparison procedure PRIDE,[7,8] it is based on a representation of the protein surface with the distribution of the inter-atomic distances and two surfaces are compared by contingency table analysis[9] of their inter-atomic distributions. Such a procedure is obviously simpler and faster than those that require the construction of molecular accessible surfaces around each of the interacting proteins. It is for example possible in only one second, with a 1GHz processor, to build the histograms and make about 38,000 comparisons of surface patches containing 40 atoms. A further advantage of this approach is that the shape of an ensemble of surface atoms within a protein can be described independently of the location of the second protein. It is thus possible to compare pairs of surface patches of different proteins without reorienting the two molecules in such a way that the two surface patches form a inter-molecular interface. In this way, complex and slow conformational search procedures, like for example the popular Katchalski-Katzir algorithm,[17] might be avoided and alternative approaches,

Table 4. Protein structures used to analyze the dependence of the sc_pride values on the fraction of atoms that are actually found at the protein-protein interface. Each protein pair can be found in the complexed and in uncomplexed form. In each case, the PDB identification code and the chain identifier are indicated. Data taken from reference 21.

Structues of pairs of proteins taken from the same PDB file				Structues of pairs of proteins taken from different PDB files			
Idcode	Chain	Idcode	Chain	Idcode	Chain	Idcode	Chain
1acb	E	1acb	I	5cha	A	1cse	I
1avw	A	1avw	B	2ptn		1ba7	A
1bcr	E	1bcr	I	1bra		1aap	A
1brs	A	1brs	D	1a2p	B	1a19	A
1cgi	E	1cgi	I	1chg		1hpt	
1cho	E	1cho	I	5cha	A	2ovo	
1cse	E	1cse	I	1scd		1acb	I
1dfj	E	1dfj	I	2bnh		7rsa	
1mah	A	1mah	F	1mma	B	1fsc	
1tgs	Z	1tgs	I	2ptn		1hpt	
1ugh	E	1udh	I	1akz		1ugi	A
2kai	A	2kai	I	2pka	XY	6pti	
2ptc	E	2ptc	I	2ptn		6pti	
2sic	E	2sic	I	1sup		3ssi	

like for example that implemented in PUZZLE,[10] where the protein surface is arbitrarily partitioned in small subunits, could gain further importance. It must, eventually, be observed that the representation of the protein surface geometry by means of the distribution of the inter-atomic distances intrinsically allows some conformational flexibility. Minor atomic displacements due to the inter-molecular association, like for example side-chain reorientations, have a minor impact on the distribution of the inter-atomic distances (Figure 4). Such an approach for measuring the surface complementarity should thus be able to implicitly treat the conformational flexibility of the protein surface.

Experimental

A set of 16,000 theoretical models taken from 16 docking simulations was examined. All experimental crystal structures were taken from the Protein Data Bank.[17] Computational simulations were performed with the software suite 3D-Dock.[16] The atomic solvent accessible area values were computed with naccess[19] with a probe radius

of 1.4 square Å. An atom was considered to be at the protein-protein interface if its solvent accessible area was different in the complexed and in the un-complexed structure.

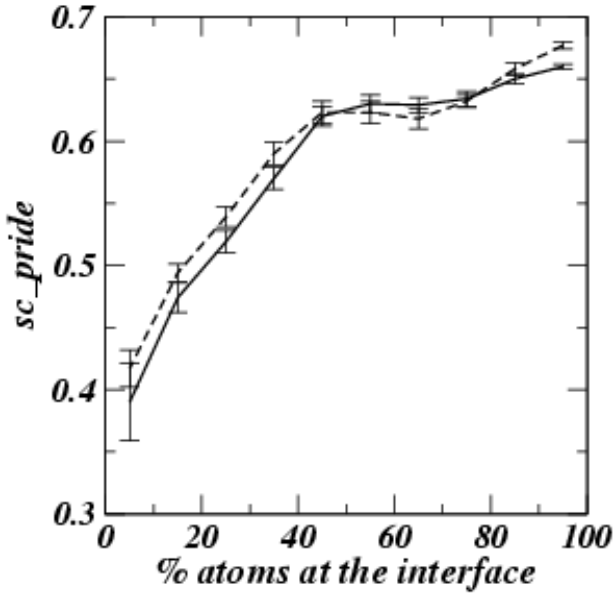

Figure 4. Dependence of the sc_pride values on the fraction of atoms that actually are involved in complexation. Standard deviations of the mean are shown by vertical bars. The continuous line indicates the comparison between the uncomplexed proteins and the broken line indicates the comparison between the proteins in the complexed conformation.

The shape of an ensemble of solvent exposed atoms was represented by a histogram of 100 bins, each 0.5 Angstroms large, describing the frequency of the distances between all atom pairs. Distances smaller than 3.5 Å were disregarded because they reflect obvious van der Waals contacts or atoms close to each other because of the covalent bonds.

The similarity between the shapes of two ensembles of atoms was estimated by contingency table analysis.[9] Given two histograms of n intervals written as obs(1,1), obs(1,2), obs(1,3)…obs(1,n) for the first surface patch and as obs(2,1), obs(2,2), obs(2,3)…obs(2,n) for the second surface patch, it is possible to compute the expected value for each observations as

$$\exp(i,j) = \frac{obs(x,i)obs(j,x)}{obs(x,x)}$$

where obs(x,i) is the sum of the observations in patch i (row sum), obs(j,x) is the sum of the observations of the variable j in the two histograms (column sum), and obs(x,x) is the sum of the observations in both histograms. Given that the observations are represented as percentages, obs(x,i) and obs(x,x) are equal to 100 and 200, respectively. Care was taken that none of the histogram bins contained less than 5% of the observations, by appropriate bin merging into *m* bins, like that described by Carugo.[7,8] The following χ^2 value can then be calculated

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{m}\frac{[obs(i,j) - \exp(i,j)]^2}{\exp(i,j)}$$

and the probability that the two distributions are identical can be deduced from the χ^2 distribution with *m*-1 degrees of freedom. The resulting probability of identity between the two frequency distributions (sc_pride) indicates if two surface patches are identical (sc_pride = 1) or totally different (sc_pride = 0).

Three other surface complementarity measures were used to validate the sc_pride values. Two of them (FADE[11] and SC[15]) are based on the Connolly molecular accessible surface[13] but are totally different in the criteria that define the similarity degree between two surface patches. The third (scscore) is based on a grid discretisation of the region between the two interacting molecules and is used in the 3D-Dock software suite.[16]

Acknowledgements

Kristina Djinovic (Sincrotrone Elettra Trieste) is gratefully acknowledged for the hospitality given to GF. A program, written in C language, is available from the authors on request.

References

1. A. V. Veselovsky, Y. D. Ivanov, A. S. Ivanov, A. I. Archakov, P. Lewi, P.Janssen, *J. Mol. Recognit.* **2002**, *15*, 405–422.
2. I. Halperin, B. Ma, H. Wolfson, R. Nussinov, *Proteins* **2002**, *47*, 409–443.
3. S. Vajda, I. A. Vasker, M. J. E. Sternberg, J. Janin, *Proteins* **2002**, *47*, 444–446.
4. M. J. Betts, M. J.E. Sternberg, *Protein Eng.* **1999**, *12*, 271–283.
5. C. J. Camacho, S. Vajda, *Curr. Opin. Struct. Biol.* **2002**, *12*, 36–40.
6. A. Heifetz, M. Eisenstein, *Protein Eng.* **2003**, *16*, 179–185
7. O. Carugo, S. Pongor, *J. Mol. Biol.* **2002**, *315*, 887–898.
8. O. Carugo, S. Pongor, *Curr. Protein Pept. Sci.* **2002**, *3*, 441–449.
9. S. Dowdy, S. Wearden, *Statistics for Research*, John Wiley & Sons, New York, N.Y., 1991.

10. M. Helmer-Citterich, A. Tramontano, *J. Mol. Biol.* **1994**, *235*, 1021–1031.
11. J. C. Mitchell, R. Kerr, L. F. Ten Eyck, *J. Mol. Graph. Model.* **2001**, *9*, 324–329.
12. L. A. Kuhn, M. A. Siani, M. E. Pique, C. L. Fisher, E. D. Getzoff, J. A. Tainer, *J. Mol. Biol.* **1992**, *228*, 13–22.
13. M. L. Connolly, *Science* **1982**, *221*, 709–713.
14. A. Pintar, O. Carugo, S. Pongor, *Bioinformatics* **2002**, *18*, 980–984.
15. M. C. Lawrence, P. M. Colman, *J. Mol. Biol.* **1992**, *234*, 946–950.
16. H. A. Gabb, R. M. Jackson, M. J. E. Sternberg, *J. Mol. Biol.* 1997, 272, 106–120.
17. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, S. J. Wodak, *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 2195–2199.
18. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bath, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.
19. S. J. Hubbard, J. M. Thornton, *Naccess Version 2.1.,* Department of biochemistry and Molecular Biology, University College, London, U.K., 1993.
20. P. J. Kraulis, *J. Appl. Crist.* **1991**, *24*, 946–950.
21. R. Chen, J. Mintseris, J. Janin, Z. Weng, *Proteins* **2003**, *53*, 88–91.

Povzetek

Predlagamo novo merilo za oceno komplementarnosti površin proteinov "sc_pride". Vsak del površine proteina predstavimo z razporeditvijo razdalj med atomi in s pomočjo kontingenčne analize teh razporeditev za dve površini ocenimo stopnjo podobnosti med njima. Tak način predstavitve površin pri nizki ločljivosti nam omogoča zelo hitro oceno komplementarnosti, kar bi lahko uporabili za napovedovanje interakcij med proteini. "sc_pride" smo primerjali z drugimi merili za oceno komplementarnosti površin na velikem številu podatkov za komplekse med proteini, ki smo jih dobili s simulacijo prileganja površin ("docking"). Na dovolj velikem številu eksperimentalnih podatkov, dobljenih z določanjem kristalne strukture kompleksov med proteini, smo preverili uporabnost "sc_pride" za prepoznavanje komplementarnost dveh površin.