International Journal of Genomics and Data Mining

Del Carpio CA and Ichiishi E. Int J Genom Data Min 01: 116.

Research Article

DOI: 10.29011/2577-0616.000116

Inference of Protein Multimeric Complex Dynamic Order of Formation: An Active Region Recognition Based Approach

Carlos A. Del Carpio[1,2*], Eichiro Ichiishi[3]

[1]Institute of Biological Diversity, International Pacific Institute of Indiana, Bloomington, USA

[2]Drosophila Genetic Resource Center. Kyoto Institute of Technology. Saga, Ippongi-cho, Ukyo-ku, Kyoto, Japan

[3]International University of Health and Welfare Hospital (IUHW Hospital) 537 Iguchi, Nasushiobara-city, Japan

[*]**Corresponding author:** Carlos A. Del Carpio, Institute of Biological Diversity. International Pacific Institute of Indiana, PO. Box 7304, Bloomington, IN 47401, USA. Tel: +15114456307; Email: carlos@dgrc.kit.ac.jp; delcmca@gmail.com

Citation: Del Carpio CA, Ichiishi E (2017) Inference of Protein Multimeric Complex Dynamic Order of Formation: An Active Region Recognition Based Approach. Int J Genom Data Min 01: 116. DOI: 10.29011/2577-0616.000116

Received Date: 16 November, 2017; **Accepted Date:** 23 November, 2017; **Published Date:** 30 November, 2017

Abstract

Prediction of bi-molecular protein associations or interactions has been the object of a gamut of computational studies, however extrapolation of these methodologies to compute the structure and function of multi-meric proteins faces several complications. One of them stems from the combinatorial aspect of the problem, since in many cases it requires the prediction of the dynamic order in which the subunits interact (the interaction path). A second, not less important, is the size of these molecules which account to the thousands of atoms, and thus require sophisticated computational platforms.

Catering to the need of predicting protein multi-meric configurations and thereby their dynamic order of formation we present here a genuine approach that requires the sole information of the isolated monomers' structures. The method is based on a protocol we have developed to recognize interaction sites on protein's surfaces. Hitherto attempts to solve this relevant problem in protein function elucidation have been limited to three body dockings using conventional docking algorithms. Here the aim is to infer complex configurations and dynamic orders of formation from the monomers known to constitute a multimeric complex unveiling active regions on the surfaces of the proteins and intermediate complexes. We present three case studies and show that important insights into the formation mechanisms of this type of multimeric complexes can be gained from the analysis of the surface characteristics of the interacting monomers which can facilitate, in a further stage, the docking and energy calculations involved in the prediction of the configurations of these complexes.

Keywords: Hydrophobic Cluster; Multimeric Protein Complexes; Protein-Protein Interaction

Introduction

Proteins and bio-macromolecules in general accomplish their biological function by aggregation, association, and interaction with other macromolecules within cells in living organisms [1-5]. The importance of protein multi-meric formation is epitomized by the diversity of physiological processes in which they play a critical role, some of them being transcription, folding, signal pathways, or cell attachment processes to name a few. In almost all of these processes molecular systems accommodate with each other so as to find the best interaction mode through a subtle molecular recognition mechanism of their interaction partners, which can either be small molecules, macromolecules, or macromolecular complexes. Associations of the latter type may respond to a cascade of events dictated by a particular undergoing cellular biochemical process [6]. In other words, the mechanism of formation of a complex of interest may respond to a sequential interaction process leading to formation of a large multi-meric complex. In these cases, formation of the intermediate may be a requirement for the third molecule to interact with the intermediate species. Thus, there is a dynamic order in the formation of the multimeric complex in which the interaction of the first two molecules lead to formation of the intermediate complex expressing the region where a third molecule is associated. On the other hand, the molecules may interact in a single step or interact simultaneously, in these cases formation of the intermediates does not influence the interaction of

any other pair and all the monomer independently interact among them leading to the formation of the final complex in a dynamic-order-of-formation free manner. In the latter case, interaction regions can be mapped on each molecule independently of any other monomer.

Studies dealing with docking two monomers have been summarized in several reviews on the matter [7-14]. Studying protein assemblies has been in vogue lately, and several reports in the literature [1-3,15,16] detail the intricacies of the problem. Many of these studies treat the multi-meric complex configuration prediction problem as an extension of the dimer complex prediction problem. The underlying idea leading the process being essentially a sequential process of interaction as mentioned before.

A seminal work in multimeric complex configuration prediction was reported several years ago where the objective was the three-body interaction of the human growth hormone (hGH) with its receptor (hGHr). This work reported by Hendrix et al. [17] led to interesting conclusions about interaction deriving in multimeric complexes. Namely, that, at least in the hormone-receptor trimer they studied, all the interactions are not equivalent, and that the formation of the complex follows a specific kinetic order. While intuitively it could be postulated that orders of interactions may indeed play a critical role in the formation of the specific conformation of a multimeric protein complex, as described earlier, this work poses the foundations for an assertive methodology leading to prediction of this type of formation mechanisms and thereby the configurations of this type of complexes.

Our group has been involved in the development of a computer system for interaction assessment of proteins and other macromolecules MIAX (bio-Macromolecular Interaction Assessment Computer System) [18-21] ; the main features of the system being the protein-complex configuration space search-engine, and the recognition of the interaction regions on proteins based on physicochemical characteristics of the atoms and amino acids constituting the molecular surface. We have shown that the algorithm performs outstandingly in recognizing hydrophobic clusters bearing high activity and being directly involved in the formation of the protein interface when interacting with another molecule [22].

With the aim of extending the reach of our original program towards the treatment of multimeric protein complexes, in this article we propose a genuine computational methodology to approach this important problem in molecular biology. The proposed methodology is fundamentally based on the recognition of active regions on the interacting monomer surfaces and their interplay in the complex formation process.

Interaction of macromolecules is intimately related to their three-dimensional structures which express interaction sites on their surfaces leading them to recognize their interaction partners or being recognized by them. These active sites naturally reside on the monomers surface or emerge as a result of particular biochemical transformations that affect the structure of the monomers and/or alter their physicochemical characteristics leading them to interact with other protein molecules [6]. The constituents of these interaction regions are atoms of amino acids distributed in particular arrangements in space according to the folding of the molecule. These amino acids seldom hold the sequential relationship dictated by the primary structures, and consequently, a sequence-to-function approach -a well-developed method for protein function prediction [23]- is limited in scope to treat the type of interactions aimed at in this paper; a rather elaborated structure-based analysis being required in order to gain insights into the underlying principles governing protein interaction and multi-meric complex formation in general.

This structure-based analysis in the present work consists in identifying the active sites of the monomers and intermediate dimers, using amino acid physicochemical parameters namely associated with the hydrophobicity and the electrostatic characteristics of the protein surface and performing an analysis of the complex formation assuming several hypothetical paths conducing to it.

We present three case studies to illustrate the analysis proposed here. We show that the active region recognition program is able to locate close-to native interaction regions and that the multimeric complex formation follows an order reflected in the emergence of interaction regions on the intermediate complexes that interact with a third molecule. When several non-intersecting interaction regions are expressed on the surfaces of the molecules the molecule interacts with different partners at the same time and the order of interactions is random.

The simplicity of the methodology is appealing, although it requires the intermediate pair wise docking of the structures in order to map the new interaction sites on the structures. However, a priori detection of the binding sites reduces the configuration space to be searched by a docking algorithm. This holds through the entire process of the multimeric configuration prediction. Moreover, with the emergence of structural genomics this process can constitute by itself an important methodology to assign structures levels of function that are not possible to assign with sequence-function based methodologies, or simple homology search in related data bases alone.

Methodology

Del Carpio et al . [18] [22] have proposed a methodology to recognize and map interaction regions on protein surfaces given the three dimensional structure of the polypeptide. These protein interaction sites -recognized by the interaction site prediction module in MIAX- are regions on the molecular surface characterized

by a set of geometrical features of the structure combined with physicochemical characteristics of a group of atoms on the surface that clearly discriminate them from the rest of the surface.

Since the algorithm has been reported elsewhere [18], here we briefly summarize its main aspects with a simple dimer example.

The algorithm to map the interaction sites of a protein 3D structure is a combination of an unsupervised learning algorithm based on a self organizing map (SOM) [24,25] with a filtering methodology based on a two dimensional fast Fourier Transform. For a particular protein structure, the first step is the computation of the solvent accessible surface points (SASP). A solvent accessible surface point is a point on the SAS that is not occluded by any other atom constituting the protein. Each atom being initially represented by a number of points uniformly distributed on a sphere of radius equivalent to the van der Waals radius of the particular atom plus the radius of a water molecule. At each of these points a physicochemical parameter is computed that depends on the atom to which the SASP belongs as well as its closest neighboring atoms. An additive property that can be easily computed to represent these effects is for example the hydrophobic potential, as introduced by Brasseur [26]. The molecular hydrophobic potential (MHP) at a SASP can be computed following the equation:

$$MHP = \sum E_{tr_i} e^{(r_i - d_i)} \quad \text{----------(1)}$$

where $\boldsymbol{E_{tri}}$ is the transfer energy parameter for atom $\boldsymbol{i}$, a value that depends on each atom type, $\boldsymbol{r_i}$ is the radius of the atom $\boldsymbol{i}$ and $\boldsymbol{d_i}$ is the distance between atom $\boldsymbol{i}$ and and the actual SASP on which the MHP is being computed. The value computed for the MHP on each SASP and the coordinates of the SASP itself can be regarded as a point in a four-dimensional space. Hence to project the 4-dimensional distribution of atomic hydrophobic potentials onto a two-dimensional space where the distribution of the MHP can be plotted and easily observed, we apply the SOM algorithm to the set of all the 4-dimensional points of the SAS of a protein. This enables the map of the distribution into a space of lower dimensionality. On the other hand, a quantitative measure of the asymmetry of the distribution of hydrophobic side chains in alpha helices and beta strands is given by a periodic property and expressed as the hydrophobic moment [27]. To extract regions of high MHP we have therefore figured out a filtering process that when applied to the SOM of the points representing the high dimensional distribution of the SASP results in a clustering of the points of the 2-dimensional map. The filtering process consists in applying a 2D Fourier transform to the set of neurons produced by the SOM and then eliminating all the points with higher frequency than a given threshold. The final hydrophobic clusters or hydrophobic patches are obtained by taking the inverse Fourier transform of the points in frequency space after the elimination of peaks of high frequency.

The methodology is illustrated in (Figure 1) for the monomers (a. receptor, b. ligand) composing the comple PDB:1CHO. The four steps sequentially in this figure show (step 1) the computation of the SASP's, (step 2) the SOM computation, (step 3) the Fourier transform of the raw SOM and the 4th step being the inverse transform showing the clusters of hydrophobic SASP's. These constitute the hydrophobic patches that lead the monomers and intermediate complex systems to interact or recognize the interaction partners.

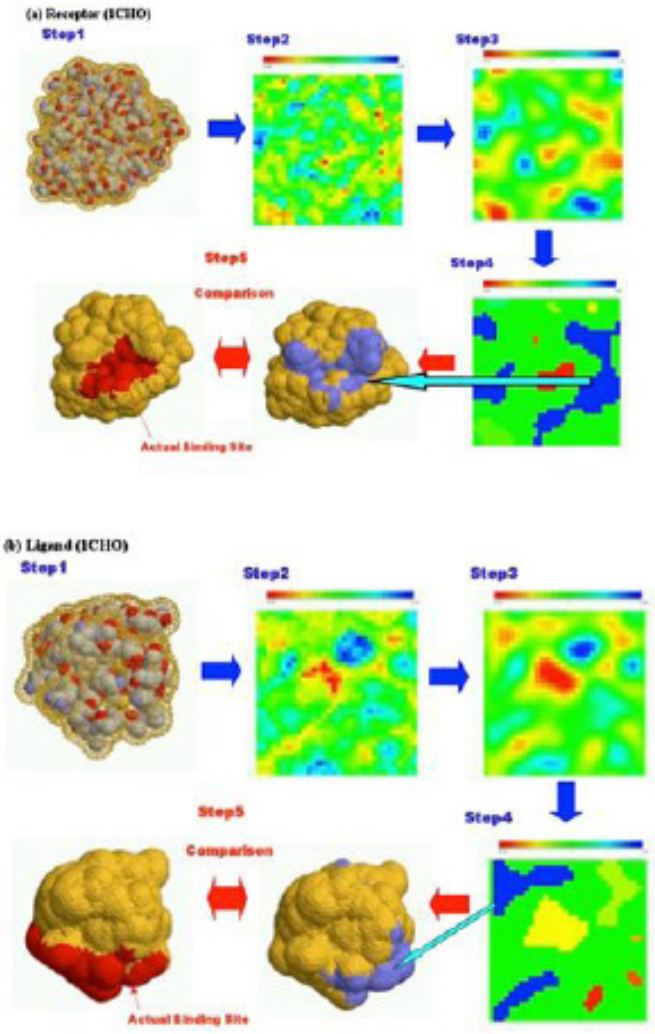

Figure 1 (a) (b): Steps in the SOM-FT protocol to predict hydrophobic patches for the receptor **(a)** and ligand **(b)** of the complex PDB: 1CHO.

The composition of each cluster is easily computed by the algorithm as shown in Figure 1, step 5, where for these two monomers the native interface is also plotted. It can be observed

that both the native interface and the computed patch coincide to a large extent, showing the validity of our methodology. The methodology has been in fact validated with several other complexes in PDB as reported elsewhere [18,22].

The computation of the hydrophobic clusters or patches by the SOM-FT technique is therefore applied in a stepwise strategy to the inference of the dynamic order of interactions of monomers forming a multimeric complex and thus its configuration.

Firstly, the hydrophobic clusters on all the monomers that form the target complex are mapped using the SOM-FT. In the case of a three-body docking, for example, if the monomers A, B, C form the complex ABC, then the order of interaction of these monomers can be studied mapping the hydrophobic clusters in monomers A, B, and C as well as the intermediate complexes AB, AC, and BC. The hydrophobic compatibility of the interfaces assists then in the inference of the order of interaction. For higher multimeric complexes the methodology is applied to all the intermediate complex structures.

Results

We have selected three case studies for validating our strategy in order to study multimeric complexes. In the first we show how the SOM-FT methodology enables the distinction of separate binding sites on the same monomer leading to its interaction with two different ligands. This leads to formation of a trimeric complex ABC the dimeric interfaces of which (AB and BC) are independent from each other. The second and third case studies show that predicting the configuration of the complex corresponds to the order in which the interaction leads to the final configuration, the methodology is thus employed to infer the order of interactions of three monomers leading to the configuration of a trimer structure. In order to perform a consistent SOM (i.e. the learning process leading to the same results) the number of steps in the learning process is proportional to 30 thousand times the number of amino acids of the subunit. This value has been obtained as result of the experiment with several protein complexes in our original paper [18].

Case Study 1

The first consists in the analysis of the trimeric complex constituted by the variable (Fv) fragment, Fv4155 from a murine immunoglobulin G1 [28] and a small organic ligand. The aim of the analysis is to automatically show that out of the six-heavy chain complemetarity determining regions (CDR's), CDR H3 accounts for most of the ligand binding and the formation of the interface between the heavy and light chains VL-VH and that the interaction sites of the VL and the steroid are relatively independent from each other. The SOM-FT is applied to analyze the heavy chain of Fv4155, the complexed structure which is reported in PDB with code 2BFV and is shown in (Figure 2a). Using a filter at level 5 (where frequencies higher than 5 are eliminated from the Fourier transform of the raw SOM) the map shown in (Figure 2b) is obtained. The proposed SOM-FT technique identifies two close but unambiguous and highly hydrophobic clusters on the V_H domain of Fv4155. Mapping back the 2-dimensional clusters onto the surface of the 3D structure of the heavy chain is shown in (Figure 2c). The composition of the clusters is shown in Table 1 together with the corresponding values for the molecular hydrophobic potential (MPH) and their contribution to the surface area. Comparing the results of Table 1 with those obtained by the crystallographic analysis of Trinh et al. [28] that are shown in Table 2, it can be observed that compositions of cluster 1 SOM-FT contains mainly amino acids that bind to the light chain while those of cluster 2 mostly coincide with the hormone binding region on the heavy chain of Fv4155.

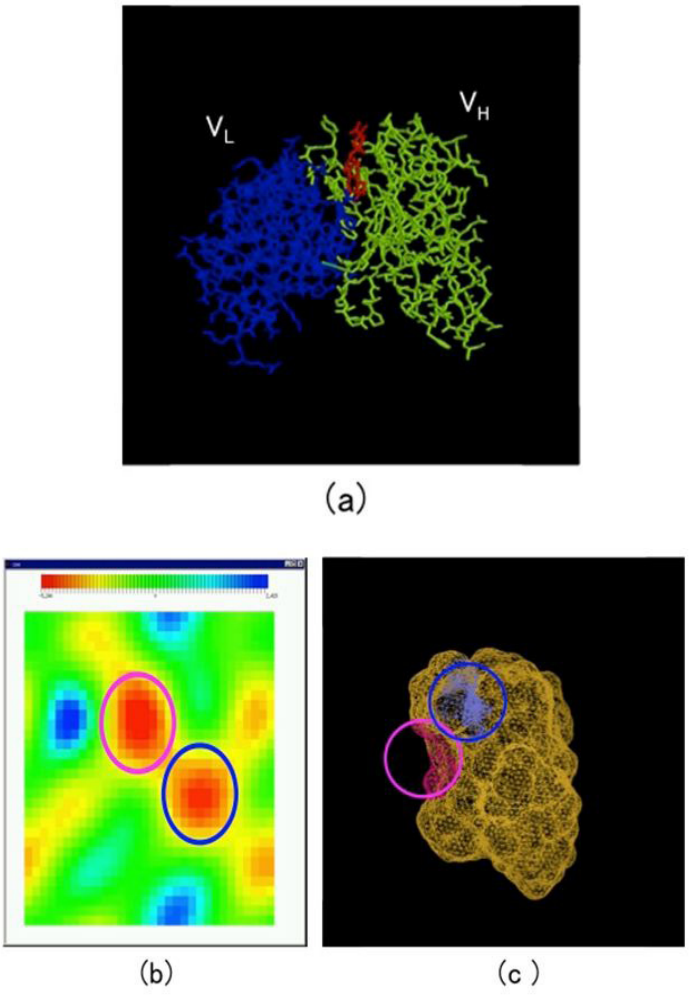

Figure 2(a-c): (a) Fv4155 fragment constituted of light chain (VL) (blue), heavy chain (VH) (green), and the steroid hormone (red). **(b)** Hydrophobic clusters predicted by the SOM-FT technique. (c) Map of the hydrophobic patches on the surface of VH.

Table 1. Cluster composition for antibody fragment Fv4155 (VH)

Cluster No	Amino Acid Composition (contributed area)	Surface (Å²)	Hydrophobic Potential (MPH)
1	I58 (25), Y59(14), L99(15), Y101(50), Y104(110)	214	-997.767
2	Y31(11), Y33(18) L47(49), Y100(26), G105(11) , M106 (17), W109(19)	151	-705.053

Table 1: Cluster Composition for antibody fragment Fv4155 (V_H).

Table 2. Composition of experimental V_H binding regions in Fv4155 (V_H) by Trinh et al (1997)

	Amino Acid Composition	Surface (Å²)	Hydrophobic Potential (MPH)
V_H-Steroid	T31(15), Y33(49), L47(18), I58(25),	263	-1540.444
Interface V_L –V_H	Y59(17), L99(18), Y101(57), Y104(52)		

Table 2: Composition of Experimental V_H binding regions in Fv4155 (V_H) by Trinh et al. (1977).

While there are some amino acids that experimentally are shown to be involved in binding of both the V_L domain and the hormone the SOM-FT technique clearly shows that the two hydrophobic clusters can accommodate two different ligands. Moreover, the V_H domain is the main contributor to the interfaces with both molecules, V_L interacts with the steroids via two hydrogen bonds, but the interactions can be attributed mostly to hydrophobic factors. The analysis using the SOM-FT technique enables to predict this behavior unambiguously and thus can be applied to reduce the search space of a docking algorithm. In addition, the technique can be used to predict the dynamic order in which molecules interact to form a multimeric protein complex as in the next case study.

Case Study 2

In the second case study we attempt to infer a dynamic order of interaction of the monomers by examining the formation of the hydrophobic clusters through the different interaction pathways that can be hypothesized from the isolated monomers. This case study is constituted by the complex 1JSU (a factor in the control of the cell life cycle [29], the formation of which occurs as result of interaction of three monomers: p27, Cdk2 and Cyclin A, the order of interaction being critical to the process [29] . In this case the order of interaction is not obvious as in the former example since independent hydrophobic patterns are difficult to discern for this complex system as shown in Figure 3, safe p27, for which only one cluster is found by the SOM-FT technique coinciding with high accuracy with the experimental or native interface.

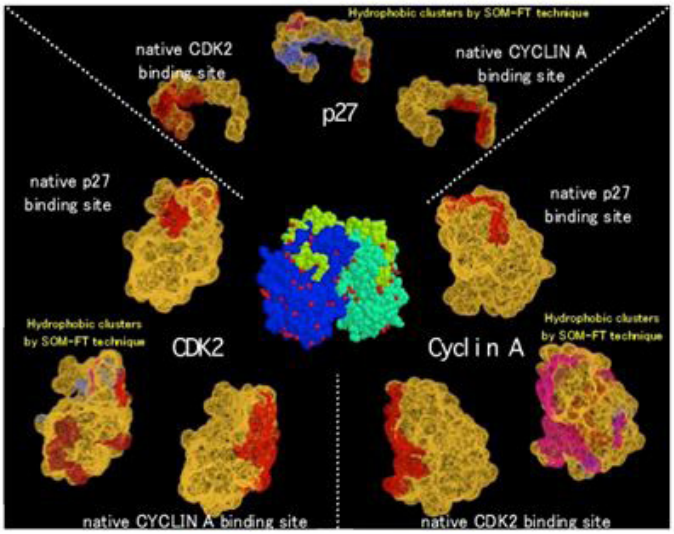

Figure 3: Computed binding sites and native interfaces for monomers: p27, Cdk2 and Cyclin A.

The clusters identified are ambiguous for Cdk2 and CyclinA, consequently different pathways leading to the formation of the complex can be hypothesized. These pathways can be generated by analyzing the intermediate complexes that would form when a particular binding order is assumed. The following three paths can be generated:

a) The first hypothetical path is shown in Figure 4 where the first interaction assumed is p27 with Cdk2. Assuming that the two monomers interact with an interface of composition similar to that found in the native trimer, identification of the cluster for interaction with the third monomer (Cyclin A) does not lead to a unique cluster similar to that of the interface found in the complex (as can be observed by visual inspection). While the native interface (shown in red on the right low part of Figure 4) is a well-defined region on the surface of the dimer, the computed interface is ambiguous and of difficult delimitation (shown in blue in the low-left part of Figure 4). Consequently, this path can be discarded from consideration.

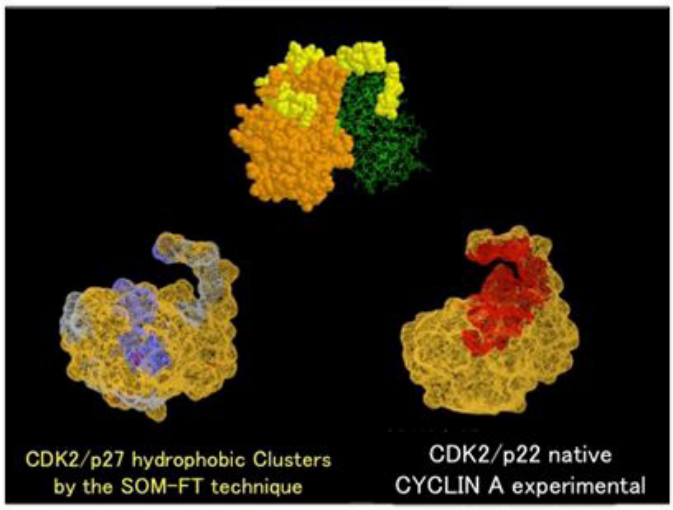

Figure 4: Computed and observed hydrophobic cluster for dimer p27 and Cdk2 (Indicating probable interface with Cyclin A).

b) A similar result is obtained when the first interaction in the path leading to the trimer is assumed to be that of the interaction between p27 and Cyclin A as illustrated in Figure 5. Here as supposed in (a) if one assumes an interaction between these two monomers with the observed interface in the trimer, there is no unambiguous recognition of the interaction region with Cdk2 by the SOM-FT technique

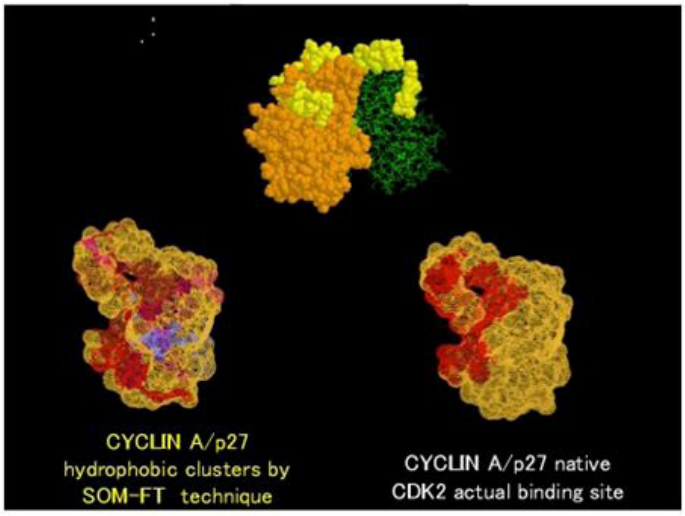

Figure 5: Computed hydrophobic cluster and native interface with Cdk2 for dimer p27 -Cyclin A.

c) A third path is illustrated in Figure 6 where the first two monomers allowed to interact are Cdk2 with Cyclin A at the interface observed in the crystal. This dimer taken as the receptor in the following step of the binding process and performing the calculation of the active site or hydrophobic cluster on its surface conduces to two independent hydrophobic clusters one of which is identical to that of the interface observed in the crystal. Furthermore, docking of the two molecules (the intermediate dimer with p27) generates a close to native structure with a high rank using any rigid body docking algorithm. All these instances would be indicating that among the three hypothetical paths proposed for the formation of the trimer 1JSU the third one would be the most feasible, and indeed, it agrees with an experimental report [29] that describes the binding of the human p27Kipl kinase inhibitory domain to the phosphorylated cyclin A-cyclin-dependent Kinase 2 (Cdk2) complex. Experimentally p27Kipl binds the complex as an extended structure interacting with both cyclin A and Cdk2, which is the path proposed by the analysis of the hydrophobic clusters in this case.

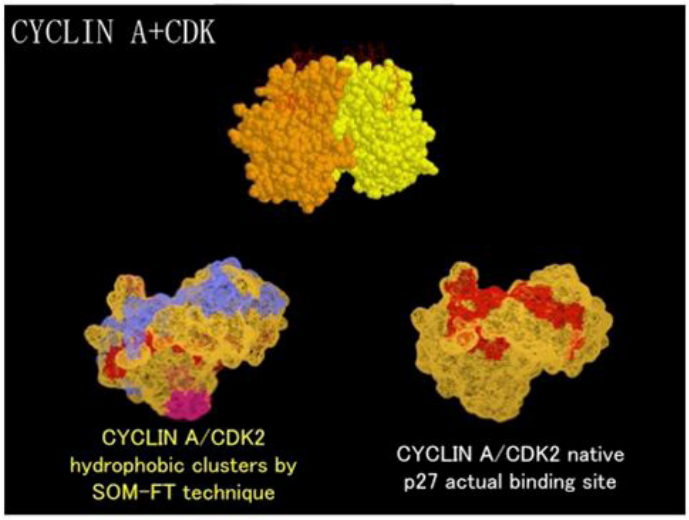

Figure 6: Computed hydrophobic clusters and actual binding site with p27 for dimer: Cyclin A - Cdk2.

This case study illustrates the importance of the shape of the hydrophobic clusters mapped by the SOM-FT technique which additionally to what was mentioned before can be used in determining the dynamic order of interaction in the case of multimeric complexes. This case study illustrates a straightforward example of a qualitative analysis of the hydrophobic clusters.

Case Study 3

A rather subtle problem is constituted by the third case study in which a qualitative analysis as described in the second case is not sufficient to make a consistent affirmation on the dynamic order of interaction of the monomers composing a multi-meric complex.

This third case study entails a detailed analysis for the complex of Adenylyl Cyclase C1A/C2A/Gs with PDB entry code 1AZS [30], and is illustrated in Figure 7.

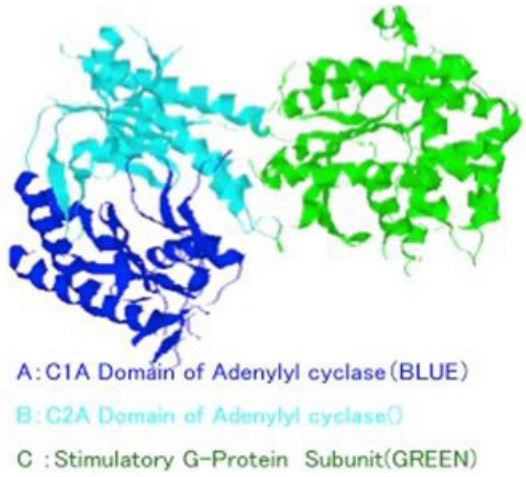

Figure 7: A ribbon model of the structure of complex 1ZAS.

Here, in order to cross-validate the SOM-FT derived results with the experimental structure the analysis based on the structure analysis of the complex has been simultaneously performed. Figure 8 shows the actual binding sites of each of the monomers composing the complex with every other monomer.

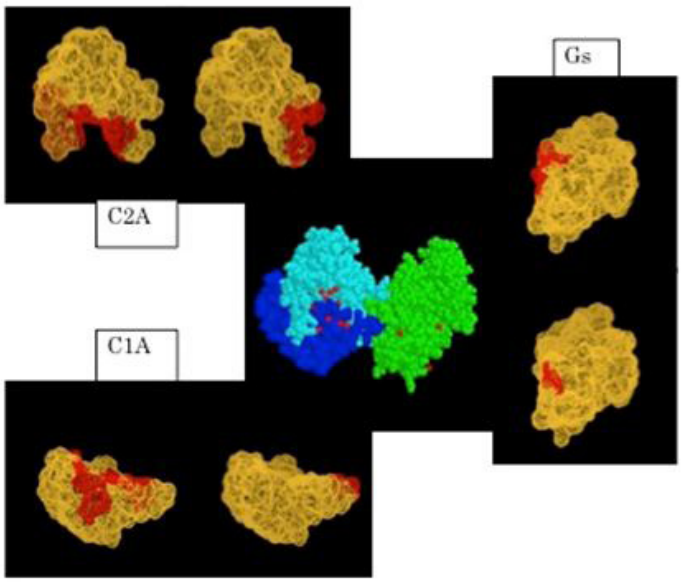

Figure 8: Binding sites of each monomer with the other two monomers composing 1ZAS.

Table 3 shows the composition of each interface (the number of SASP for each monomer and intermediate complex, the SASP per patch, the total and average MHP value, and the number of SASP corresponding to hydrophobic and hydrophilic amino acids). Moreover, the composition of the SAS in terms of the number of SASP belonging to the 20 amino acids is illustrated in Table 3, and these results are graphically presented in Figure 9, where the histograms on the right of each interface illustrate the character of each binding region in relation to the interaction partner.

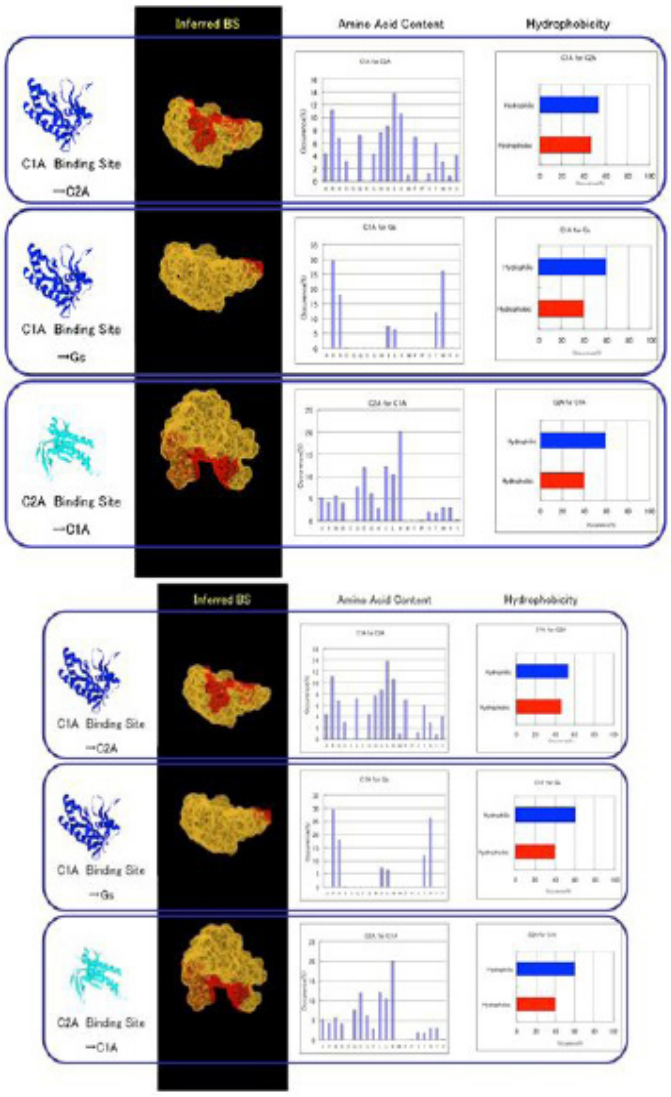

Figure 9 (a-b): Quantitative analysis of the hydrophobicity of each binding site (C1A -> Gs read: binding region C1A for Gs).

Figure 10 shows the same information but in this case for the following step in the interaction cascade, i.e. the binding regions and their characteristics assuming the formation of each of the hypothetical dimmers. It can be observed that binding regions with higher hydrophobic potential are present for only two species in the path of interaction to form complex 1AZS, and these are the clusters of monomer Gs for interaction with monomer C1A on one hand, while on the other is the cluster of the species C1A/C2A that interacts with monomer Gs.

Table 3. Characteristics of the actual interaction interfaces

			Total SASP	B.S. SASP	HMP	Av. HMP	Hphob SASP	Hphil SAS
1AZS	C1A	C2A	8111	1709	−2655.43	−1.55	797	912
		Gs	12770	233	148.04	0.64	93	140
	C2A	C1A	8223	1756	−2236.01	−1.27	704	1052
		Gs	12770	732	−792.94	−1.08	230	502
	Gs	C1A	8111	324	−1480.44	−4.57	305	19
		C2A	8223	782	−1620.31	−2.07	361	421
	C1A/C2A	Gs	12869	986	−2996.96	−3.04	608	378
	C2A/Gs	C1A	19479	1860	−1877.05	−1.01	751	1109
	C1A/Gs	C2A	20324	2308	−3278.23	−1.42	941	1367

			A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1AZS	C1A	C2A	74	190	115	52	0	123	0	74	130	147	235	181	16	118	0	19	102	50	14	69
		Gs	0	69	42	1	0	0	0	0	0	17	15	0	0	0	0	0	28	61	0	0
	C2A	C1A	92	72	99	69	0	134	210	109	49	213	183	352	0	0	1	35	31	51	51	5
		Gs	0	241	86	12	2	101	17	0	0	94	3	9	0	22	0	36	0	88	21	0
	Gs	C1A	0	0	0	1	0	0	0	0	18	0	6	0	178	127	0	0	0	0	0	0
		C2A	4	50	40	40	1	0	90	0	99	9	22	70	0	214	0	26	12	0	0	111
	C1A/C2A	Gs	3	26	40	39	1	0	65	0	104	9	14	70	175	295	0	26	12	0	0	111
	C2A/Gs	C1A	91	104	122	68	0	134	185	109	49	218	196	352	0	0	1	35	55	87	51	5
	C1A/Gs	C2A	74	416	182	64	2	224	17	74	117	229	238	190	13	94	0	55	102	113	35	69

Table 3: Characteristics of the actual interaction interfaces for the complex 1AZS (C1A-C2A-Gs).

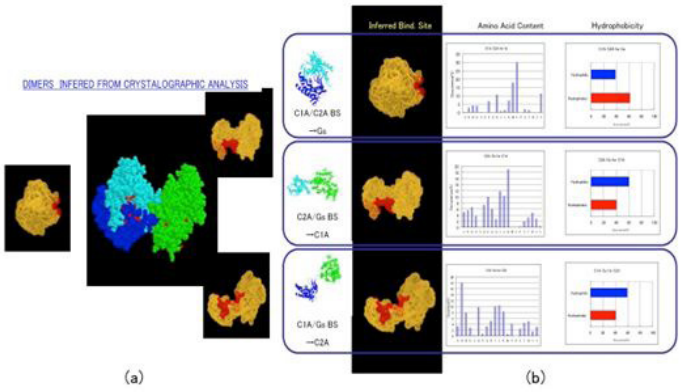

Figure 10: Identification of binding regions in dimeric species in the path of formation of complex 1AZS (C1A/Gs BS →C2A; read: intermediate C1A/Gs binding site for C2A).

A consistency test applied to the system would evidently lead to the conclusion that the most relevant step in the formation of the complex is the formation of the intermediate C1A/C2A which expresses a highly hydrophobic region which must be the driving force behind its association with monomer Gs.

The automatic procedure to discriminate the native interaction path leading to complex 1AZS must then have the ability of identifying the intermediate C1A/C2A as the first complex intermediate expressing its highly hydrophobic cluster.

We show that the SOM-FT methodology proposed here fulfills this requirement. The automatic prediction of the interaction path begins with the inference of the binding sites using the SOM-FT protocol. The calculations are summarized in Tables 4a, b, c and d.

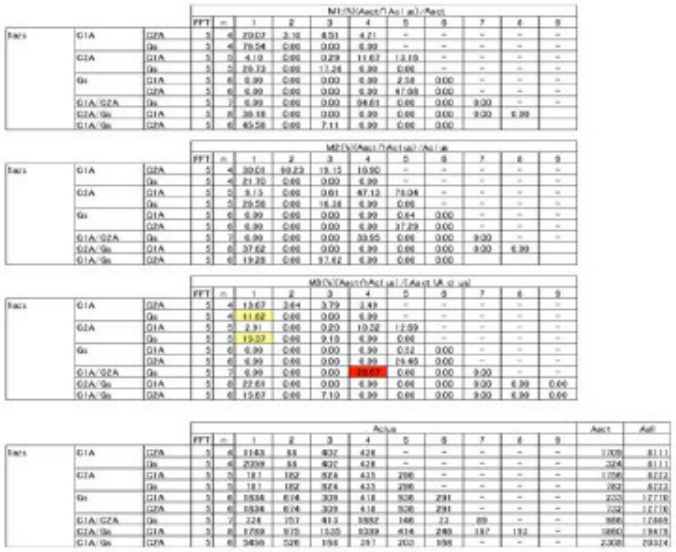

Table 4(a-d): Analysis of SOM-FT predicted clusters (Aclus) with actual interfaces (Aact) for the components of complex 1AZS (n =number of predicted clusters at level FFT).

One of the characteristics of the SOM-FT protocol is the exhaustive search for hydrophobic clusters it performs on protein surfaces, resulting many times in several clusters of variable size. Significant clusters are, however, only a fraction of them, the maximum number allowed for analysis here are nine clusters, taken at a filtering level of 5 [18]. To compare the predicted clusters with those in the experimental structure, three metrics: M1, M2 and M3 are computed as shown in tables 4a, 4b, and 4c respectively. Large values of M1 stand for predicted clusters larger than the experimental interface, and large values of M2 mean larger experimental regions than the predicted ones. Table 4d, shows the absolute number of SASP constituting each predicted cluster. Inspection of tables 4a and 4b reveals that many binding regions involved in the interaction path are predicted correctly. Thus, M1 and M2 values for clusters C1A→C2A (cluster 1), C1A→Gs (cluster 1), C2A→Gs (cluster 1) are both of high magnitude meaning that the predicted clusters overlap well with those of the actual crystal. Particularly significant are the values of M1 (64.81) and M2(33.85) for cluster 4 of the C1A/C2A→Gs intermediate which shows the preeminence of this cluster over any other in the intermediate step of the interaction path. Formation of the complex C1A/C2A on which a highly hydrophobic cluster is expressed is therefore correctly predicted by the system, as shown by a further calculation in table 4c, where the value of M3 for the cluster in the intermediate C1A/C2A underscores any other, including those for C1A→Gs and C12→Gs which, if higher would stand for a different pathway of interactions. The hydrophobic path identified on the C1A/C2A by the SOM-FT is shown in blue in Figure 11, where the real inter-

face is also shown in red for visual comparison. Consequently, the order of interaction of the monomers to form the 1AZS complex is shown in Figure 12.

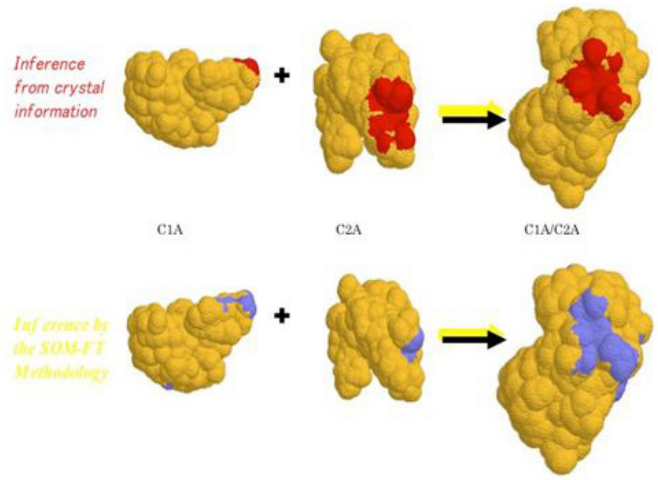

Figure 11: Comparison of the cluster on the intermediate complex C1A/C2A; red: native interface, blue: SOM-FT predicted hydrophobic cluster.

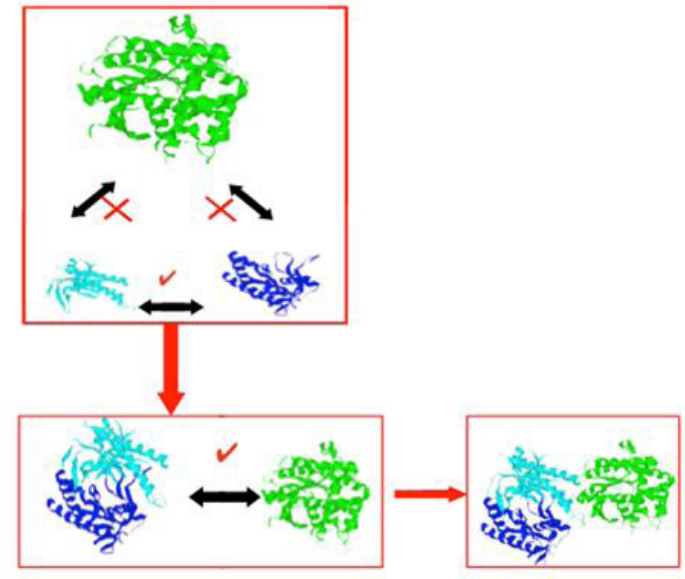

Figure 12: Inferred interaction path leading to complex 1AZS (× → not allowed, √ → allowed).

We have carried analysis of the dynamic order for the formation of several other complexes (data not shown), with the methodology proposed and using the SOM-FT analysis of active sites on protein surfaces. We have arrived at similar results as the discussed in the three precedent cases. Since the SOM-FT methodology intrinsically involves a geometrical factor of the interacting monomers, the active sites recognized by the program are complementary in shape and hydrophobic characteristics. This is a further factor that can be used in reducing the search space for docking algorithms, ranking decoys with interfaces close to the identified interaction regions on the monomers and dimmers over other orientations. One must note, however, that the automatically inferred hydrophobic clusters will not infallibly provide the complete, close-to-native information about an interaction site. This is because the patches it recognizes are hydrophobic and may not overlap entirely with the real interface. Nevertheless, these clusters undoubtedly provide enough clues for assessing the likelihood of a hypothetical path of dynamic interactions in the formation of dimmers and thereby the configuration of multimeric protein complexes.

Conclusions

The present article aims at the evaluation of pathways of interaction leading to multimeric complexes by analyzing the geometry and physicochemical characteristics of automatically recognized interaction regions on surfaces of the interacting protein species in hypothetical pathways also automatically generated. The technique underlying the recognition of these interaction regions is a genuine algorithm that combines unsupervised machine learning (SOM) and a spectrum analysis methodology based on Fourier transforms (FR) that has been proposed by the authors [18,22], the SOM-FT technique. This methodology for assessing the paths of formation as well as configuration of multimeric proteins identifies active sites (hydrophobic clusters) on the surfaces of the interacting proteins (monomers, and complex intermediates).

We have applied the technique to three interesting case studies. The common denominator to these cases, (and thereby the reason to use them as prototypes in this study) is the fact that experimental details on their formation have been reported, facilitating the conclusions of the computational results we present. The first is concerned with the analysis of an antibody structure that is reported bound to some steroids. Coinciding with the experimental findings [28] we have found that CDR H3 has prominent roles both in ligand binding and in formation of the VL-VH complex. For Fv4155, the CDR H3 alone accounts for 45.5% (~323Å2) of the buried surface area of the heavy chain upon Fv formation, the hydrophobic clusters found by the SOM-FT analysis predicts a somewhat larger area for both hydrophobic clusters covering the experimentally reported one, with the fundamental difference that the clusters infer a priori that the VH is able to host two ligands, independently. The second case study shows how the SOM-FT protocol can be used to predict the dynamic order of interaction of the monomers to form the multimeric complex. As a suitable model complex to illustrate this procedure we selected the formation of complex PDB: 1JSU from the monomers p27, Cdk2 and Cyclin A. The inference of the pathway of complex formation requires the

analysis of all the hypothetical pathways using the hydrophobic clusters generated by the SOM-FT technique at all the levels of those pathways. We show that in this case a qualitative analysis of the clusters suffices to infer the right dynamic order of interaction or pathway leading to the configuration of the complex.

A more detailed analysis is necessary in the third case study since the SOM-FT produces many more clusters for each species. Here, nevertheless, comparison of the automatically predicted interaction regions coincide with the experimental complexes interfaces. The second case study shows a qualitative way in which the information on the hydrophobic patches or clusters can be used to infer interaction pathways. However, treating other systems with clusters where the information is not unambiguous on the extend of hydrophobicity driving certain interactions requires a step by step computation of these values.

We have shown that the likelihood of a path of interaction over other hypothetical paths can be assessed using the information provided by the automatically recognized hydrophobic patches. The process has to be evidently complemented by a docking algorithm-within the system for macromolecular interaction assessment MIAX [18] for instance-however, the procedure we propose besides constituting by itself a putative methodology to search for starting positions in docking processes, optimizing thus configurational searching operations, it can be a powerful method when analyzing multi-meric protein complexes for which the interactions dynamic order is the main information required for analyzing biochemical or biophysical processes that are driven mainly by these type of interactions.

Finally, with the launching of structural genomic projects, the methodology presented here can be the basis for achieving the long goal of predicting protein function within the context of sequence-to-structure-to function, there where sequence-to-function methods are quite limited to achieve these goals.

References

1. Bonvin AMJJ, Karaca E, Melquiond ASJ, de Vries SJ, Kastritis PL (2010) Building Macromolecular Assemblies by Information-driven Docking. Molecular & Cellular Proteomics 9: 1784-1794.
2. Cheng TMK, Blundell TL, Fernandez-Recio J (2008) Structural assembly of two-domain proteins by rigid-body docking. BMC Bioinformatics 9: 441.
3. Chitale M, Hawkins T, Park C, Kihara D (2009) ESG: extended similarity group method for automated protein function prediction. Bioinformatics 25: 1739-1745.
4. Chuang GY, Kozakov D, Brenke R, Comeau SR, Vajda S (2008) DARS (Decoys as the Reference State) potentials for protein-protein docking. Biophys J 95: 4217-4227.
5. de Vries SJ, Bonvin AMJJ (2011) CPORT: A Consensus Interface Predictor and Its Performance in Prediction-Driven Docking with HADDOCK. PLoS One 6.
6. Scott JD, Pawson T (2009) Cell Signaling in Space and Time: Where Proteins Come Together and When They're Apart. Science 326: 1220-1224.
7. Arafat Y, Kamruzzaman J, Karmakar GC, Fernandez-Recio J (2009) Predicting protein-protein interfaces as clusters of optimal docking area points. Int J Data Min Bioinform 3: 55-67.
8. Bajaj C, Chowdhury R, Siddavanahalli V (2011) F2Dock: fast Fourier protein-protein docking. IEEE/ACM Trans Comput Biol Bioinform 8: 45-58.
9. Baker D, Schueler-Furman O, Wang C (2005) Progress in protein-protein docking: Atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. Proteins 60: 187-194.
10. Bonvin AMJJ, de Vries SJ, Melquiond ASJ, Kastritis PL, Karaca E, et al. (2010) JPGLM: Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. Proteins 78: 3242-3249.
11. Bonvin AMJJ, Kastritis PL (2010) Are Scoring Functions in Protein-Protein Docking Ready to Predict Interactomes? Clues from a Novel Binding Affinity Benchmark. Journal of Proteome Research 9: 2216-2225.
12. Bourquard T, Bernauer J, Aze J, Poupon A (2011) A collaborative filtering approach for protein-protein docking scoring functions. PLoS One 6: e18541.
13. Comeau SR, Vajda S, Camacho CJ (2005) Performance of the first protein docking server ClusPro in CAPRI rounds 3-5. Proteins 60: 239-244.
14. Marshall GR, Vakser IA (2005) Protein-Protein Docking Methods. Protein Reviews 3: 115-146.
15. Huang SY, Zou X (2007) Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. Proteins 66: 399-421.
16. Janin JI (2002) Welcome to CAPRI: A Critical Assessment of PRedicted Interactions. Proteins: Structure, Function, and Genetics 47: 257-257.
17. Hendrix DK, Klein ET, Kuntz ID (1999) Macromolecular docking of a three-body system: The recognition of human growth hormone by its receptor. Protein Science 8: 1010-1022.
18. Del Carpio CA, Ichiishi E, Yoshimori A, Yoshikawa T (2002) A new paradigm for modeling biomacromolecular interactions and complex formation in condensed pahses. Proteins: Structure, Function, and Genetics 48: 696-732.
19. Del Carpio CA, Iulian Florea M, Suzuki A, Tsuboi H, Hatakeyama N, et al. (2009) A graph theoretical approach for assessing bio-macromolecular complex structural stability. J Mol Model 15: 1349-1370.
20. Del Carpio CA, Qiang P, Ichiishi E, Tsuboi H, Koyama M, (2006) Robotic path planning and protein complex modeling considering low frequency intra-molecular loop and domain motions. Genome Inform 17: 270-278.
21. Del Carpio CA, Shaikh AR, Ichiishi E, Koyama M, Kubo M (2005) A Graph Theoretical Approach for Analysis of Protein Flexibility Change at Protein Complex Formation. Genome Informatics 16: 148-160.

22. Yoshimori A, Del Carpio Munoz CA (2001) Automatic Epitope Recognition in Proteins Oriented to the System for Macromolecular Interaction Assessment MIAX. Genome Informatics 12: 113-122.

23. Skolnick J, Fetrow JS (2000) From genes to protein structure and function: novel applications of computational approaches in the genomic era. Trends Biotechnol 18: 34-39.

24. Kohonen T (1999) Comparison of SOM point densities based on different criteria. Neural Comput 11: 2081-2095.

25. Kohonen T (1990) The self-organizing map. Proceedings of the IEEE 78: 1464-1480.

26. Brasseur R (1991) Differentiation of Lipid - associating Helices byUse of Three dimensiona Molecular Hydrophobicity Potential Calculations. The Journal of Biological Chemistry 266: 16120-16127.

27. Eisenberg D, Schwarz E, Komaromy M, Wall R (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. Journal of Molecular Biology 179: 125-142.

28. Trinh CH, Hemmington DS, Verhoeyen ME, Phillips SEV (1997) Antibody fragment Fv4155 bound to two closely related steroid hormones: the structural basis of fine specificity. Structure 5: 937-948.

29. Russo AA, Jeffrey D, Patten AK, Massague J, Pavletich NT (1996) Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. Nature 382: 325-331.

30. Tesmer JJ, Sunahara RK, Gilman AG, Sprang SR (1997) Crystal Structure of the Catalytic Domains of Adenylyl Cyclase in a Complex with GsazGTPgS. Science 278: 1907-1916.