PROTEIN-PROTEIN DOCKING USING GEOMETRICAL ARGUMENTS

Israel Cabeza de vaca Lopez
israel.cabeza@bsc.es

30-08-2009

Abstract

This project aims to develop nobels algorithms to model protein-protein complexes, a very important aspect in biophysics. The algorithm presented based only on geometrical arguments, is intended to be a first and fast approach to get the most probable configurations. The algorithm finds the best positions producing only a small number of solutions (over 250 solutions). The method is based on 2D FFT (fast fourier transform) and orthographic projections of the proteins. The method allows us to find solutions around 15 Å of C_α root mean square deviation for proteins with low electrostatic interactions.

Keywords: protein-protein docking, geometrical predictive docking, protein recognition, orthographic projection coefficient

1 Introduction

The 3D structures of protein complexes are important to understand molecular systems. The prediction of the final protein-protein complex using computer algorithms is complicated because there are many factors to take into account such as hydrophobicity, electrostatics, Van der Waals forces, etc. Most of the docking methods aim to reproduce these physics interactions with successful results in only a few cases[(1)].
The standard method of computer algorithms for protein-protein docking are based on shape complementarity[(2)]. Some studies are based on matching surface[(3)] while others focus on matching the position of surface spheres and surface normals[(4)]. The shape complementarity is measured using scoring functions which sometimes include electrostatics interactions[(5)] and hydrophobic effects are included.
Some methods include solutions to take into account the flexibility problem. The unbounded and bounded proteins have different conformations because the shape changes minimize the global energy of the complex. This produces enormous problems because variations in geometry implies changes in the solution. Some authors define grids with two surfaces where the external surface is less important than the internal one to allow a small superposition simulating flexibility[(6)].

Critical Assessment of Prediction of Interactions (CAPRI)[(7)] is a community-wide experiment for protein protein docking. Research groups around the word tries to find the new experimental complex proposed by CAPRI with his algorithms. The average percentage of acceptable solutions (RMSD<10) to predictor groups of the last nine proteins used in rounds 14, 15, 16, 17, 18 and 19 has been 7,4 %.

The method proposed here studies the rigid body case without any flexibility. This algorithm uses a grid discretization combined with surface recognition using a new coefficient called orthographic projection coefficient (OP). This reduces the number of solutions proposed and shows the regions geometrically favorable for protein interactions.
All protein protein docking methods are composed of two parts: Global and local search. The algorithm proposed here is focalized to global search generating a few structures using OP coefficient. The refinement part produces a local search but it could be improved with other known methods. Local methods work well when the initial structure is close to the experimental solution and spend a lot of time per solution. This algorithm tries to find a small set of solutions per protein with a group of solutions close to the experimental complex.

2 Algorithm

2.1 Identify Surfaces

The algorithm developed to identify surfaces consists of producing collisions between the protein and a spherical bullet which simulates the action of water molecules. The bullets are shot from many

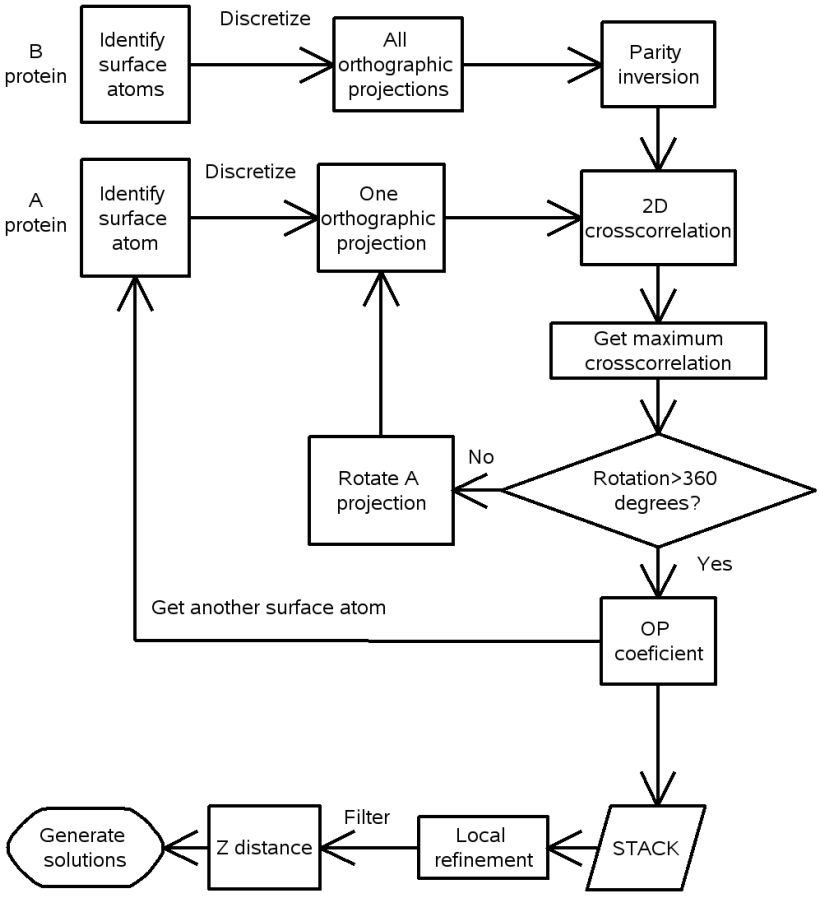

Figure 1: Scheme of the main algorithm. The algorithm is divided in two different parts. First part computes the better OP coefficients and the second part produces a refinement to generate the final complex.

points in six space directions with a linear trajectory from a far distance to detect the first collition. Then, the atoms detected with the collisions are considered the surface atoms of the protein. The size chosen for the bullet is the size of the water molecule 1.4 Å.
This method can not detect every surface atom specially when there are deep holes into the protein because the linear trajectory does not explore everything. But the number of solutions lost are very small and it is a fast and useful method.

2.2 Surface complementarity using crosscorrelation

The main algorithm is developed to work in complexes of two proteins identified by A and B (normally B is the smaller one). The real docking problem has to take into account two or more proteins at the same time but the algorithm is designed for the simplest system - two elements.

The first part consists of finding the surface atoms of A and B proteins. Then, to each atom of the surface it computes a discrete orthographic projection using a grid size per cell of 1 Å. The orthographic projection corresponds to a plane with a orthogonal vector generated between the geometrical center (GC) of the protein and the surface atom chosen to this projection. The space point of the plane is the position of GC and the values put into the grid matrix are the orthogonal distance of the atoms to the plane. The B protein has to be inverted changing the sign in all grid elements (parity inversion) because crosscorrelation computes the similarity between them. To find the best docking complementarity between both grids the algorithm includes a 2D rotation of A grid to score more orientations with these two projections. Then, the highest crosscorrelation result obtained from this process with all the B projections are saved in a file with the information to reproduce the complex found. The last step is to divide the best result by the autocorrelation of A grid to scale the coefficient. It allows us to compare between grids with different number of elements (see figure 1).
Crosscorrelation is computed with discrete fast fourier transform (DFT) using the convolution theorem[8]. Using it the speed of the algorithm in-

creases as $nlog(n)$, where n is the size of the grid. The number of DFT that the algorithm has to do follows the next equation

$$N_{FFT} = N_A N_B * \left[\left(\frac{360}{\alpha} \right) \right] \quad (1)$$

Where N_A and N_B are the number of atoms at the surface A and B, respectively. α is the angle of rotation chosen to rotate the grid and we have used $\alpha = 15$ in these simulations.

2.3 Reconstruction of the solution

The second part of the process consists of generating the 3D structure associated with these results and comparing it with the right solution using root mean square deviation (RMSD). The crosscorrelation algorithm fits the best position giving us an optimum bidimensional position in the plane of the orthographic projections but the real structure is tridimensional. The reconstruction algorithm computes the closest separation between A and B proteins computing the optimal radial distance avoiding overlap by fixing some parameters.
The reconstruction algorithm puts the complex (A and B proteins) into the GC. It rotates the proteins putting the surface atoms with the best crosscorrelation found with the previous algorithm in the opposite position in the same axis (180 degrees of rotation). The next step, is to separate the proteins a large distance and then apply the X-Y shift found previously. The last step consists of reducing the separation in Z axis progressively to get a minimum distance between a pair of atoms of 5 Å.

The structure generated by the reconstruction algorithm has an small error associated with the resolution of the grid and rotation angle. To minimize it the reconstruction algorithm produces a small exploration of the solution around the result given by surface recognition to find the deeper solution with a minimum distance between a pair of atoms of 5 Å. The method to explore consist in rotations of 15 degrees around the line between geometrical centers of both proteins and translations of 5 Å. This method is slow, it should be applied when there are a reduced number of solutions.

2.4 Computing RMSD

The root mean square deviation is a usefull measurement in the protein-protein docking problems[9]. The value is a measure of the proximity to the real result when comparing to an experimental structure. As usual in this type of calculations, the RMSD is computed only with α-carbons of the B proteins when experimental complex and solution proposed are superposed. The aminoacid side-chains have conformational changes when the proteins are bound and to avoid masking good results they are excluded from RMSD computation.

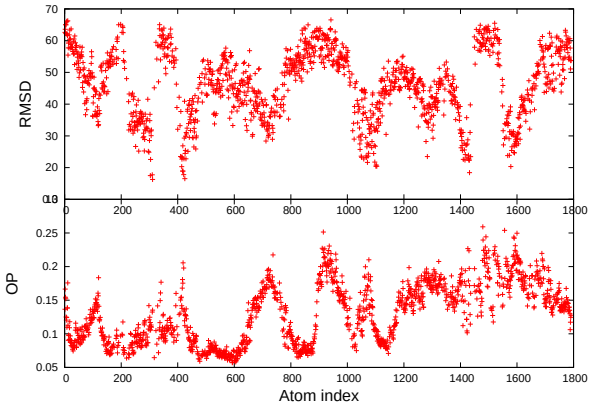

Figure 2: Complex 1CGI. Six local maximums of OP coefficient has coincidences with 6 local minimums of RMSD. Four local maximums goes to the right space region.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} (\vec{r}_{1i} - \vec{r}_{2i})}{n}} \quad (2)$$

Where n is the number of α-carbons and 1 means reference and 2 means solution proposed.

3 Results

The algorithm has been checked with seven complex. Chymotrypsinogen and trypsin, Bowman-Birk inhibitor and trypsin, MT-SP1/matriptase, beta-trypsin and CMTI-I, uracil-DNA glycosylase and uracil glycosylase inhibitor, eglin-c-subtilisin Carlsberg and CI-2-subtilisin Novo, colicin E7 and Im7 protein called in protein data bank as 1CGI, 1D6R, 1EAW, 1PPE, 1UDI, 2SNI and 7CEI, respectively. These proteins have been chosen from a data set of known proteins with bound and unbound structures found experimentally.
The number of outputs of the program is equal to the number of atoms at the surface of A protein. Standard methods in protein-protein docking show results using rankings of RMSD vs solutions proposed. All programs produce many solutions and these are sorted using scoring functions (normally based on total energy). The proposed algorithm uses OP coefficient (see equation 3) to study the solutions and reduce its number. We have found a relation between the increase of this coefficient and the decrease of RMSD (see figure 2). It means that one can reduce the number of solutions proposed without losing the best solutions. The method used to reduce the number of solutions consists of finding local maximums of the OP coefficient and storing only a range of solutions around the maximum value. The mean number of local maximums per protein is over 12.
The reduced number of results allows to use an improved version of the reconstruction algorithm re-

Complex	N_{lm}	N_{fm}	N_{right}
1CGI	11	6	4
1D6R	12	5	2
1EAW	14	6	4
1PPE	12	7	5
1UDI	8	5	4
2SNI	12	6	3
7CEI	6	0	0

Table 1: N_{lm} is number of local maximums of the OP coefficient. N_{fm} is number of local maximums with a RMSD local minimum associated. N_{right} is the number of maximums that produce solutions with the lowest RMSD.

ducing the errors associated to this new list of results. We have observed an increased of the number of solutions close to the right solution with this modification.
OP coefficient is a method to normalize crosscorrelations between different projections and it is defined by equation 3.

$$OP_{ij} = \frac{C(A_i, B_j)}{C(A_i, A_i)} \tag{3}$$

Where A_i, B_j means grid projection of i, j serial atom of the surface respectively.

The algorithm has been checked with 7 proteins and the results are similar in six cases. The values depicted in the table 2 show results around 15 Å RMSD for the best solution. 7CEI protein has a large electrostatic interaction and this may be the reason for less accurated results closest to the reference (see table 1 and discussions). Table 2 shows the value of OP coefficient to the bound and unbound reference in column 1 and 2, respectively. The third column in table 2 shows the OP result obtained from the algorithm (see figure 2) for the solution with lowest RMSD. The forth column shows the OP result for refined solution. Column 5 is the number of solutions found with a RMSD less than 20 Å after refinement obtained applying the algorithm of the reference to the best result obtained. Column 6 and 7 show the RMSD found with the bounded and unbounded structures, respectively. The last column shows the rank position of the lowest RMSD solution found in the final list.

The local maximums of OP where you have best solutions are less than the global maximum because the best geometrical docking is not the real docking. In some cases, the value between bounded and unbounded is similar because the deformation produced in the process is small.

Figure 2 shows the relation found between OP and RMSD for 1CGI protein as well as the fluctuation of RMSD and OP with the atom index. The spreading of OP and RMSD is produced by the error associated with the discretization of rotations and positions but the global behaviour gives us the important information. Table 1 shows a summary of results found with the seven proteins tested. The best protein is 1PPE with 5 coincidences and table 1 also depicts the lower RMSD in this protein.

4 Discussions and Conclusions

The docking method presented is only based on geometrical parameters. The method tries to find the best docking position using a FFT analisis of the surface shape with orthographic projections. The main objective is to produce solutions close to the experimental solution identifying the better regions of complementary quickly.

The common accepted values for RMSD in a docking program is less than 10 Å. This method produces results close to be accepted and the proportion with low RMSD allows us to combine it with other slower docking methods. The application of the algorithm to one protein produces a number of solutions close to 2000 as seen in figure 2. This figure indicates a substantial correlation with the OP score and the local RMSD minima. Thus, the OP coefficient allows us reduce the number of solutions around 250 without loosing the best results. This new coefficient reduces on average 10 times the number of solutions and gives us the information of the better region for the A protein to attach B protein using only geometrical information. This coefficient can be combined with other established methods to rank and refine these results. The reconstruction of the solution produces a refinement using a small translation and rotation to check the deepest position between the proteins.

One method to get more solutions is modifying the main algorithm to accept a range of higher crosscorrelation results per atom index and not only the maximum crosscorrelation. This variation of the method produces an improvement of 2 Å in RMSD in the best solutions but increases a lot the number of solutions proposed (data not shown). The number of solutions is directly proportional to the number of higher croscorrelations accepted.

In six of 7 complexes we do find the region in the conformational 3D space where the proteins dock. This is accomplished in less than 2 hours of CPU. Thus OP coefficient can be used to find the geometrical docking regions and reduce the number of solutions proposed by some methods. As seen in table 2, however, the OP coefficient is not a good scoring function. The results of the experimental complex are usually worst that many different other. Fur-

Complex	$OP_{ref\,B}$	$OP_{ref\,UB}$	$OP_{best_{org.}}$	OP_{best}	N. RMSD< 20	RMSD_B	RMSD_{UB}	Rank
1CGI	0.025	0.024	0.112	0.022	23	14.75	15.02	215
1D6R	0.011	0.012	0.35	0.046	13	16.28	17.39	9
1EAW	0.041	0.037	0.19	0.026	7	16.61	16.76	146
1PPE	0.039	0.013	0.11	0.010	54	10.87	10.68	191
1UDI	0.0041	0.009	0.012	0.017	6	18.82	19.47	178
2SNI	0.016	0.015	0.23	0.010	25	15.2	15.38	28
7CEI	0.100	0.07	0.28	0.104	0	20.02	20.07	98

Table 2: $OP_{ref\,B}$ is the value of the OP coefficient in the experimental bounded reference. $OP_{ref\,UB}$ is the value of the OP in the unbounded proteins superposed in the experimental reference. $OP_{best_{org.}}$ is the OP coefficient of the best solution found before local refinement. OP_{best} is the OP coefficient of the best complex after refinement. N. RMSD < 20 is the number of solutions obtained with this condition. RMSD_B and RMSD_{UB} is the best RMSD obtained with bounded and unbounded, respectively. Rank means the position in the number of solutions obtained sorted by OP.

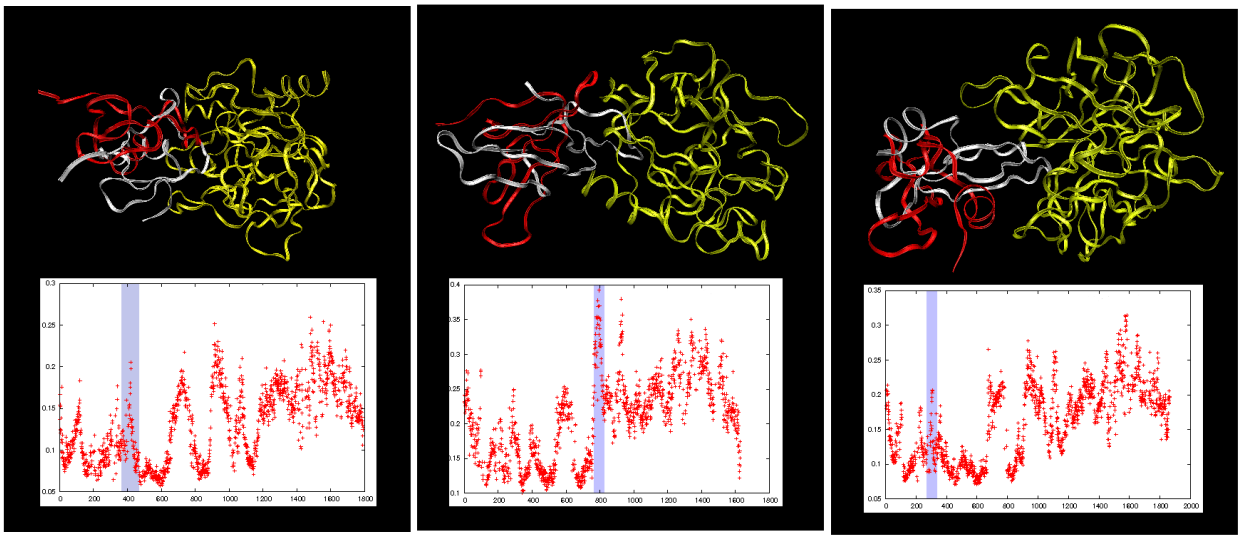

Figure 3: The representation structure of the best solutions found for 1CGI (left), 1D6R (center) and 1EAW (right) complexes. Yellow and white are the chain A and B of the bounded reference, respectively. Red chain is the best solution found for the B chain starting from the unbounded complex. Blue region in OP coefficient plot show us which maximum produces these solutions.

thermore, the best RMSD is ranked usually far from the top 10. To improve this ranking, one can try to develop another coefficient based on OP including other interactions.

The experimental X-ray complex should be the global minimum energy. An ideal accurate scoring function would always rank it as the top pose. For this purpose, we should add into our scoring function more energy terms describing the real interaction physics. For this reason, some researchers use electrostatics and hydrophobic interactions to improve the scoring eliminating bad energetics results[10],[11] and improve the scoring. Our next step will be to combine the geometrical results of OP coefficient with electrostatics methods. One way will be try to add electrostatics information to the grids to improve the OP coefficient.

As seen in figure 3, the results of the algorithm show in six proteins some solutions with an overlapping between experimental and proposed B protein. It implies good prediction for docking position but not good orientation in the three Euler angles. Thus, future improvement will include Monte Carlo rotations to find the complex with a major number of contacts between A and B proteins. Defining one contact between proteins as a pair distance of atoms between 10 Å and 4 Å. These distances define a range where there is the minimum of Lennard-Jones potential to pairs of atoms. The number of contacts is proportional to the stability of the protein complex.

5 Technical information

The complete package consists of approximately 7500 lines written in C++ with object oriented programming, bash and PERL. The code is designed to run in a linux machine. The code uses the MIT library FFTW 2.1.5[12] written in C to compute the

crosscorrelation. The parallelization has been implemented with MPICH2 library and the speedup is equal to the number of processors.
The time to compute a process depends of the size of the protein. The average time to compute a set of OP coefcients is 20 minutes in 8 processors PowerPC 970 2300 MHz (9.2 GFlops). The time to generate the solution with a refinement in a single processor is 2 hours.

6 Acknowledge

Thanks to lifescience group (Barcelona Supercomputing Center) for the opportunity to use Marenostrum and Juan Fernandez Recio group for provinding the test set. I'm great grateful to Victor Guallar for his helpful discussions and direction.

References

[1] S. Grosdidier, C. Pons, A. Solernou, J. Fernndez-Recio PROTEINS: Structure, Function, and Bioinformatics, Volume 69 Issue 4, Pages 852 - 858 (2007)

[2] Shoichet, B. K.,Kuntz, I. D. Predicting the structure of protein complexes: a step in the right direction. Chem. Biol. 3, 151 - 156 (1996)

[3] Jiang, F., Kim, S. Soft docking matching of molecular surface cubes. J. Mol. Biol. 219, 79 - 102 (1991).

[4] Shoichet, B. K., Kuntz, I. D. Protein docking and complementarity. J. Mol. Biol. 221, 327 - 346 (1991).

[5] Walls, P. H., Sternberg, M. J. E. New algorithm to model protein-protein recognition based on surface complementarity. J. Mol. Biol. 228, 277 - 297 (1992).

[6] Rong, C., Zhiping, W. "A novel shape complementarity scoring function for prot-prot docking", PROTEINS: Structure, Function, and genetics 51:397-408 (2003).

[7] Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ . "CAPRI: a Critical Assessment of PRedicted Interactions". Proteins 52 (1): 29 (2003)

[8] Bracewell, R. "Pentagram Notation for Cross Correlation." The Fourier Transform and Its Applications. New York: McGraw-Hill, pp. 46 and 243 (1965).

[9] J.Fernandez-Recio,"Optimal docking area",PROTEINS: Structure, Function, and bioinformatics 58:134-143 (2005).

[10] Henry A. Gabb, Richard M. Jackson and Michael J. E. Sternberg, "Modelling Protein Docking using Shape Complementarity, Electrostatics and Biochemical Information", J. Mol. Biol. (1997) 272, 106-120.

[11] Vakser, I., Aflalo, C. "Hydrophobic docking: A proposed enhancement to molecular recognition techniques", PROTEINS: Structure, Function, and genetics 20:320-329 (1994).

[12] Matteo Frigo and Steven G. Johnson, "The Design and Implementation of FFTW3," Proceedings of the IEEE 93 (2), 216231 (2005).