DockAnalyse: An Algorithm for Protein-Protein Interaction Analysis

Delicado P., Amela I., Gómez A., Querol E., Cedano J.

Prof. Pedro Delicado

Departament d'Estadística i Investigació Operativa

Universitat Politècnica de Catalunya. 08034-Barcelona. Spain.

Isaac Amela

PhD Student

Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular
Universitat Autònoma de Barcelona. 08193-Bellaterra, Barcelona. Spain.

Dr. Antonio Gómez

Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular
Universitat Autònoma de Barcelona. 08193-Bellaterra, Barcelona. Spain.

Prof. Enrique Querol
Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular
Universitat Autònoma de Barcelona. 08193-Bellaterra, Barcelona. Spain.

Dr. Juan Cedano

Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular
Universitat Autònoma de Barcelona. 08193-Bellaterra, Barcelona. Spain.

Phone: +34-93-5812807

Fax: +34-93-5812011

Email: jcedano@servet.uab.es

ABSTRACT:

Is it possible to know what the best solution is of the output of a docking program? The apparent obvious answer to this question is the highest-score solution, but the interaction between proteins seems to be a dynamic mechanism where the interaction region has to be wide enough to allow protein-protein interactions coming from different orientations, and sometimes, as in a multimeric complex, several interacting regions are possible. Consequently, in all these cases, there is not a single solution. To extract the significant solutions from the docking-output data, we have developed an unsupervised and automatic cluster computational program supported on a mathematical algorithm. This algorithm is based on the DBscan clustering method, which searches for continuities among clusters generated by the output-docking data representation. The DBscan clustering method solves some of the inconsistency problems of the classical clustering methods like k-means: a) it allows the easy treatment of outliers (isolated points) taking into account only those clusters that are large enough to be an interaction region, b) it allows the finding of all cluster members independently of the cluster shape and, c) it is not dependent on the previously defined number of clusters. A pre-processing step finds the radius, which is necessary to run the DBscan method, without requiring any tuning parameter by the user. We used this approach to postulate complex interaction mechanisms that involve surface displacements among proteins that interact together to carry out a specific function. Another useful application for the mathematical model presented here could be the prediction of the putative structure of protein complexes. The algorithm and the application (implemented in the R package) are accessible on: http://www-eio.upc.es/~delicado/Rfiles/DockAnalyse.zip.

Keywords: Docking, Interaction, Algorithm, Clustering, Protein-complex.

INTRODUCTION:

Protein interaction is a key process by which most of the proteins accomplish their function and interactomics represents one of the main frontiers of biosciences [1, 2]. Moreover, protein interactions can help us to predict protein function and, therefore, many protein-function predictors have been developed using protein-protein interaction databases [3-9]. In a near future, it is expected that the number of protein complexes will surpass the number of proteins in a given proteome. A lot of protein interactions involve surface displacements among the members of the protein complex to fulfill the required biological function. As an example, we can mention the proteins that take part in the Iron-Sulfur Cluster (ISC) biogenesis mechanism in yeast [10]. It is also known that many proteins are single parts, labeled monomers, of a complex quaternary structure, a multimer. In any case, monomers alone do not have a specific function which is only achieved when the distinct parts interact together to carry out a certain function [11, 12].

Nuclear Magnetic Resonance (NMR) and X-ray crystallography are the main technologies for structure elucidation. They are frequently constrained by the methodological requirements: on the one hand, NMR works well only for small, highly soluble proteins so, it cannot be used for the structure determination of large protein complexes; on the other hand, X-ray crystallography demands such special conditions that, in most cases, they do not match those required for the protein complex structure formation and co-crystallization. As a whole idea, structural data for protein complexes is arduous to obtain because it is very difficult to fix the optimal conditions that are required for these experiments. It is assumed that these experimental limitations have reduced the amount of large protein complexes solved and, therefore, protein complexes have become less represented in the structural databases such as the Protein Data Bank (PDB; http://www.rscb.org/pdb/; [13]). With the aim of analyzing the dynamics of the interaction process among the proteins of a protein complex, which can involve large interaction zones and surface displacements, a NMR assay may not be feasible because, as said before, this technique is designed only for small proteins. In the same case, and assuming that it were possible to carry out an X-ray crystallography experiment despite the above-mentioned difficulties, the data obtained may not be useful either to represent the dynamic behavior because crystals show only a static image of the protein complex. In addition, in the case of the protein complexes where the constituents have several interaction sites, a static image of the interaction between them may not be informative enough to model the structure with which it will fulfill a specific function. In general, the tri-dimensional structure obtained might not be exhaustive enough to explore all the possible interaction sites between the proteins considered and, therefore, it might not have the required detail. Consequently, despite almost everybody tending to apply these two experimental technologies for protein structure determination, other complementary strategies may be useful to accurately model the interaction among the proteins of a protein-complex.

These alternative methods prompted us to propose a recognition interface within the proteins of a macro-molecular complex to help to elucidate its putative quaternary structure, from an exhaustive data analysis, without requiring any extra-experimental data. Furthermore, these methods can also lead us to model the movement among the components of a macromolecular structure which is necessary to fulfill a specific function, also minimizing the use of experimental procedures. In this context, some theoretical methods to study protein complexes at a structural level, such as docking, are

now emerging. Protein-protein docking is a computational method to predict the best way by which two proteins could interact [14, 15]. It is important to be careful using docking programs because, even those proteins that, experimentally, do not interact, can generate an output file including a list of putative interactions. Also, it is important to choose a suitable docking program which generates the most realistic docking model. In rigid-body docking approaches, conformational changes during the complex formation are not permitted in order to save computation time. This technique may be appropriate when non-substantial conformational changes are expected to take place in the interacting proteins. Usually, it is considered that the best solution given by a docking program is the one with the best interaction energy, but quite a lot of the real interactions tend to involve large surface displacements with non-optimal interaction energy to form the protein complex. They take place along the protein's surface generating multiple low-energy interaction complexes. In these cases, these low-interaction energy regions might not be, in reality, less important from a functional point of view and the interaction region has to be wide enough to allow protein-protein interactions coming from different orientations, like, for instance, proteins that require movements among them when they act as a protein-complex. Owing to all of these facts, interaction among proteins seems to be a dynamic mechanism where there is not only a single solution with the highest interaction energy, like most of the current docking programs consider, but rather there are several solutions with more or less interaction energy [16].

Our approach tries to deal with these particularities considering the global contribution of the different, calculated solutions and select those representatives that describe a general behavior of a subset of solutions, but do not improve an unrealistic docking output file. To extract those solutions that best describe the real dynamic mechanism of interaction from the output data of the current protein-protein docking programs, we have developed an algorithm based on an unsupervised and automatic cluster analysis. The aim of the algorithm is to choose the appropriate solutions, not only by taking into account the interaction energy, but also the dependence among the clusters generated by the docking-output data representation. Choosing the representative solutions is made by searching for continuities among these clusters [17]. The real challenge is the ability to identify the correct structures from among the huge amount of previously calculated solutions. That is why we developed the algorithm, called *DockAnalyse*, to do this easily without requiring any tuning parameter from the user. Normally, the decision on which of the docked structures is the most important entails the implication of an expert researcher in the field, however *DockAnalyse* guides the search of a good docking candidate, without requiring any previous expertise, thus reducing the amount of putative solutions to check.

The exhaustive analysis of all of the protein interaction regions considered by *DockAnalyse*, may help us to theoretically postulate the structure of a protein complex. Additionally, it could be useful in proposing the way in which certain proteins interact together to execute their biological function. As an example of one of the applications of the developed program, we modeled the dynamic interaction mechanism between the yeast proteins Isu1 and Isu2, which have been demonstrated experimentally to interact. These two proteins generate the central platform for ISC biogenesis inside the yeast mitochondria [10] and, moreover, are important targets in the yeast model of the human neuro-degenerative disease called Friedreich ataxia (FRDA) [18-21]. Above all, we could see that they interact in a mobile fashion and, therefore, their interaction seems to imply large surface displacements.

MATERIALS AND METHODS:

THE ALGORITHM:

With the aim of elucidating which of the docked structures between two proteins are the most important from a functional point of view, an unsupervised mathematical algorithm, based on the DBscan clustering method, was designed and implemented with R package. The shape, size and movement, expressed in rotations and translations described by the proteins, were considered in the algorithm to finally obtain the cluster distribution with the best internal coherence among clusters generated by the docking output data representation. A pre-processing step finds the radius necessary to run the DBscan method without requiring any tuning parameter from the user. Regarding the DBscan, this clustering method was chosen because it is extremely robust and solves some inconsistence problems that may appear when applying other clustering methods. In general, classical clustering methods do not manage the outliers well. DBscan tends to treat these isolated points much better and it allows for the finding of all cluster members independently of the cluster shape. Finally, and one of the main problems of clustering, is that the classical clustering methods are dependent on the previously defined number of clusters while DBscan is not. The developed mathematical algorithm, named *DockAnalyse*, was applied to interpret the results obtained from different docking assays. *DockAnalyse* generates a lot of information extracted from the output data file of these docking assays. Among this information, a visual representation of the 1000 docking solutions was obtained from the docking experiment, which are represented as single points grouped in different clusters, is one of the most important results. The representative solutions of each of the calculated clusters are highlighted and they refer to the significant points among all of the 1000 docked structures tested (see Fig.1). These points represent the most relevant solutions obtained from the protein-protein docking calculation and they allow us to identify which solutions among those could be more directly involved in the interaction, because it is a central member of the cluster and it has a high interaction energy. As a whole, what represents a real challenge that can be achieved with *DockAnalyse* is the reduction of the number of solutions to analyze after a protein-protein docking experiment. With our program the docking output-data analysis is facilitated, because the number of solutions is reduced from a huge number (e.g., 1000) to approximately less than 10 in most of the cases. Therefore, it enables an accurated study of the most interesting protein-protein docking solutions.

APPLICATIONS OF DOCKANALYSE:

Considering all of the facts stated above, we could go further proposing new functional interpretations that involve our proteins of interest. In terms of these new hypothesis, when the initially docked proteins are monomers, a proposal on the putative structure of a multimeric protein complex might be postulated [11, 12]. Another procedure to visualize the expected surface displacements between two interacting regions. This approach colud be applied to pairs of proteins that require movements between them to fulfill a specific function [10]. In this second case, the representative solutions obtained from the *DockAnalyse* analysis, which represent different protein-structure configurations, were captured and subsequently viewed in a protein modeling or visualization program. This procedure allowed us to build a point-to-point pseudo-

trajectory to postulate a model to explain the surface displacements between the given proteins. This pseudo-trajectory could be reconstructed by means of the selection of other solutions along the cluster or joining different cluster representatives. These solutions could be considered as static frames that describes the motion between the two given proteins.

An example of proteins taking part in a protein-complex that involve surface displacements in the interaction between them to accomplish a specific function, are the yeast proteins Isu1 and Isu2, which are required for the ISC biogenesis inside the mitochondria [10]. With the purpose of making a dynamic model by which Isu1 and Isu2 interact together for the generation of the ISCs, these two yeast proteins were studied in detail from a sequential, structural and functional point of view. Firstly, the sequences for both Isu1 and Isu2 from evolutionarily distinct organisms were retrieved. With these sets of sequences, we could perform a multi-alignment sequence analysis with which we could see that these two proteins are extremely evolutionary conserved and, therefore, have a high sequential homology. Moreover, for each of the two yeast protein sequences, several classical bioinformatics analyses were made to establish some important characteristics for these proteins. The cellular localization, the putative DNA interaction or trans-membrane regions and the signal peptide length for the two proteins were studied. Despite the tri-dimensional structure of these two proteins not being available in the PDB [13], we could model its structure due to the sequential and structural homology to some already-solved protein family members. Besides, the residues considered to model the structure of Isu1 and Isu2 were part of the functional protein inside the mitochondria because, as studied before, the signal peptide length was not overlapped them. The 2D and 3D structure modeling was done applying three widely used applications designed for this purpose. On the one hand, the secondary structure was predicted using PsiPred [22], which incorporates neuronal networks to the outputs of PSI-BLAST. On the other hand, the tridimensional structure was predicted using EsyPred3D [23], which is an homology-based application that uses the MODELLER package, and 3D-PSSM [24], which is a threading-based application that uses both the 1D-2D-3D structural information and the solvation potentials information of the protein. Secondly, the functions for Isu1 and Isu2 were analysed with several bioinformatics tools and the results were contrasted with the already published material which referred to these proteins [25]. Furthermore, with the aim of identifying the protein-protein interaction regions in Isu1 and Isu2, two analysis were performed with programs designed for this purpose. The first program used was ProMate [26], which is a protein-structure-based program, and then we used PPI-Pred [27], which is a support-vector-machine-based program that we applied to corroborate the results. Lastly, searches in interactomics databases were performed to elucidate the protein interaction partners of Isu1 and Isu2 [28-34]. Obviously, the above mentioned studies were complemented with the appropriate literature information to contrast all of the data obtained. Besides, most of the previously described studies needed molecular visualization or modeling tools widely used in bioinformatics like, for instance, RasMol or UCSF Chimera [35, 36].

PROTEIN-PROTEIN DOCKINGS:

The Escher NG protein-protein automatic docking system of the VEGA ZZ project was the program used to execute the dockings between Isu1 and Isu2 [37]. In more detail, two parameters were modified of the docking procedure: on the one hand, we set the rotation step to 3 degrees because small rotations result in more docking

details and, on the other hand, we established the maximum number of collisions to 100 with the aim of avoiding errors during the bump-check process. Despite the existence of the possibility of reducing the number of given solutions for the docking, we conserved it with the default parameter, which is 1000, because the more solutions obtained in the docking assay the, more robust, the *DockAnalyse* results would be. Taking these premises into account, the docking output datafile that was obtained contained information for the one thousand solutions tested coming from the different, docked protein-structure configurations (see Fig. 2). Tri-dimensional structure information for each of the proteins studied was obtained from the PDB as pdb files [13]. To evaluate the reliability of *DockAnalyse* and see if it was helpful to predict the movement between two given proteins, we chose the most representative solutions given by the newly developed mathematical algorithm and we extracted the PDB files that represent the protein-structure configuration between the docked monomers for each of the solutions. The PDB file generation from the solutions given by *DockAnalyse* was done using a tool of the VEGA ZZ package, where Escher NG is linked [37, 38]. Then, the generated PDB files for each of the representative solutions of *DockAnalyse* were loaded in a protein modeling or visualization tool with which we could analyze the putative surface displacements between the initially docked proteins. We have to take into account that some expert knowledge about the living-protein context would be necessary to propose a coherent trajectory involving several surface displacements. With this experiment we could evaluate the utility of our algorithm in facilitating the search for the correct docking solution among all of those that result from the protein-protein experiment. Furthermore, this was helpful to see if *DockAnalyse* could accurately predict the interaction movements between two given proteins.

RESULTS AND DISCUSSION:

DETAILS OF DOCKANALYSE:

The clustering algorithm DBscan [39] relies on a density-based notion of clusters and is designed to discover clusters of arbitrary shape as well as to distinguish noise. The algorithm is based on the definition of density-connection: two points in a dataset are density-connected if there is a chain of points in the dataset that allows for moving from one to the other. The connecting chain must verify two conditions: first, each point in the chain (except maybe the first and the last ones) has at least k observed data at a distance less than the radius epsilon (that is, they are in places where the density of data is not too low), and second, the distance between two consecutive points in the chain is less than epsilon. This definition induces a partition in the set of observed points, each part defined as a subset of points which are density-connected between them. The clusters provided by DBscan are those components in the partition with two or more elements. DBscan marks those not density-connected with any other as isolated points (that is, parts with only one member).

The algorithm DBscan depends on two tuning parameters: k and epsilon, defining density-connection. Ester et al. (1996) [39] indicate that choosing parameter epsilon is much more important than choosing k (they argue that the results in their databases are quite similar for any $k \geq 4$) and they propose to fix k as being equal to 4. In our experiments we have verified that values of k greater than 4 provide better results. So we use k=15 in all computations.

The better epsilon parameter is chosen according to a battery of cluster-quality measures. To be specific, we have considered the following clustering indexes (see Walesiak and Dudek (2007) [40], for more details): Davies-Bouldin (multiplied by -1), Calinski-Harabasz, Hubert-Levine (multiplied by -1) and Silhouette, all of them taking a high value for a high quality clustering. Algorithm DBscan was applied for several epsilon-candidate values and the resulting clustering structures were evaluated by these criteria. Then the candidate values for epsilon were ranked according to every index. The score of a candidate value for epsilon was the mean of its ranks. Then the value with highest score is taken as the final epsilon and the corresponding cluster structure is considered the right one.

THE DESIGNING PROCEDURE:

DockAnalyse can be used after the docking assays of a more complex procedure where it might be useful to model the behavior between the proteins that take part in a biologically functional protein-complex. A descriptive scheme on how to use *DockAnalyse* in this whole bioinformatics procedure is shown in Fig. 3. First of all, an extensive literature-mining analysis coupled with a profound study of the sequence, the structure, the function and the interactome involving the proteins of interest is required. Secondly, the necessary protein-protein docking experiments have to be executed, taking into account that the more solutions tested during the docking assays, the more robust the results from *DockAnalyse* would be. After that, the newly developed mathematical algorithm, *DockAnalyse*, has to be applied to each of the docking output-data files allowing us to obtain the best docking solutions among those thousands calculated. Lastly, manual curation of the docking structures obtained might be necessary to fit the solutions given with the appropriate biological function eliminating the putative aberrant results.

MODELLING THE ISU1-ISU2 PROTEIN COMPLEX:

The yeast proteins Isu1 and Isu2 are conserved proteins of the mitochondrial matrix which perform a scaffolding function during the assembly/maturation of Iron-Sulfur clusters (ISC). Therefore, they physically and functionally interact, leading to the formation of the protein complex required for ISC biogenesis [10, 41-44]. Moreover, these two proteins have been shown to be involved in the lack of ISC generation of the human disease called Friedriech Ataxia (FRDA) [21, 45]. FRDA is a neurological, progressive and hereditary disease which, basically, through the nervous system, the spinal cord, the neurons, and cortico-spinocerebellar routes concerns the balance and the coordination of movements. Furthermore, it is the most common autosomal recessive ataxia and it is associated with a pronounced lack of a conserved mitochondrial protein of a not fully-understood function called Frataxin. An important expansion of a triplet GAA in the first intron of the gene FRDA involves an aberrant structure of the DNA helix complicating the passing of the RNA polymerase enzyme during the transcription of the gene. This event causes a reduced expression of Frataxin. This protein has been associated with iron accumulation in the mitochondria, increased sensitivity to oxidative stress and, more recently, with the assembly/maturation of the mitochondrial ISCs ([Fe-S] clusters). Consequently, Frataxin and its partners Isu1 and Isu2 play an important role in ISC protein assembly, avoiding the depletion of proteins like Aconitase and respiratory chain complexes I-III inside the mitochondria. The study has been done using *Saccharomyces cerevisiae* proteins as a model, because a high degree of similarity

is supposed between the human and the yeast molecular mechanisms for the ISC biogenesis [18-20, 46].

In terms of the key characteristics found in the sequence, the structure and the function of Isu1 and Isu2, we could see that both proteins have a clearly visible tail, which has been predicted as a protein-protein interaction region. Just on the opposite side of each of the proteins, another interaction zone has been identified coinciding with an iron-binding pocket, which is composed of 3 Cysteines placed to be spatially suitable for forming a typical iron-binding pocket [see images (a) and (b) of Fig.4]. A rigid-body protein-protein docking procedure was made between the modeled tri-dimensional structures of Isu1 and Isu2 setting a small rotation step to exhaustively explore a great number of solutions in a reasonable computing time. Then, we applied *DockAnalys* algorithm with which we could reduce the huge amount of output docking data obtained to five representative solutions that are the 1st, 4th, 28th, 89th and 296th of the initial ranked solutions of the Escher NG output docking datafile. Here, one utility of *DockAnalyse*, in reducing the number of solutions to analyze after a protein-protein docking calculation, is shown. However, specific knowledge about the ISC biogenesis process involving Isu1 and Isu2 was also necessary to extract biological sense from the representative protein-protein docking solutions which were transformed to structure configuration images. Consequently, two out of the five images were finally considered, according to the expected biological function. These two images are the initial, ranked Escher docking solutions 1st and 28th respectively, which are, respectively, the 1st [see image (d) of Fig. 4] and the 3rd [see image (c) of Fig. 4] *DockAnalyse* representatives. Therefore, the 2nd, the 4th and the 5th *DockAnalyse* representatives were discarded due to the inexistent biological concordance with the other representatives.

In summary, the postulated dynamic interaction mechanism between the yeast proteins Isu1 and Isu2 is as follows: According to the subsequent images analyzed, it seems that Isu1 continues being positioned, with respect to Isu2, until it achieves the appropriate orientation [see image (d) of Fig. 4]. When Isu1 is positioned in front of Isu2, the two iron-binding pockets seem to be correctly placed spatially and, moreover, the interaction tails seem to play an important role acting as a hinge between the two proteins. This hinge allows for the movement of Isu1, with respect to Isu2, and permits the iron donation and the sulfur donation, thanks to other donor proteins such as Frataxin, which was mentioned before, to the ISC biogenesis when they are needed [see the reversible pass between (c) and (d)]. We have to consider that during all of the interaction process, the tails of the two yeast proteins clearly facilitate their interaction.

ACKNOWLEDGMENTS:

This research was supported by Grants (BIO2007-67904-C02-01, MTM2006-09920) from the MCYT (Ministerio de Ciencia y Tecnología, Spain) and from the Centre de Referència de R+D de Biotecnologia de la Generalitat de Catalunya. I. Amela is a predoctoral fellowship recipient from the UAB (Universitat Autònoma de Barcelona). The English of this manuscript has been corrected by Mr. Chuck Simmons, a native English-speaking Instructor of English of this University.

REFERENCES:

1. Pache, R.A., et al., *Towards a molecular characterisation of pathological pathways.* FEBS Lett, 2008. **582**(8): p. 1259-65.

2. Gavin, A.C. and G. Superti-Furga, *Protein complexes and proteome organization from yeast to man.* Curr Opin Chem Biol, 2003. **7**(1): p. 21-7.
3. Chen, X.W., M. Liu, and R. Ward, *Protein function assignment through mining cross-species protein-protein interactions.* PLoS ONE, 2008. **3**(2): p. e1562.
4. Chua, H.N., W.K. Sung, and L. Wong, *Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions.* Bioinformatics, 2006. **22**(13): p. 1623-30.
5. Espadaler, J., et al., *Prediction of enzyme function by combining sequence similarity and protein interactions.* BMC Bioinformatics, 2008. **9**: p. 249.
6. Gabow, A.P., et al., *Improving protein function prediction methods with integrated literature data.* BMC Bioinformatics, 2008. **9**: p. 198.
7. Jaeger, S., et al., *Integrating protein-protein interactions and text mining for protein function prediction.* BMC Bioinformatics, 2008. **9 Suppl 8**: p. S2.
8. Sun, S., et al., *Faster and more accurate global protein function assignment from protein interaction networks using the MFGO algorithm.* FEBS Lett, 2006. **580**(7): p. 1891-6.
9. Vazquez, A., et al., *Global protein function prediction from protein-protein interaction networks.* Nat Biotechnol, 2003. **21**(6): p. 697-700.
10. Lill, R. and U. Muhlenhoff, *Iron-sulfur protein biogenesis in eukaryotes: components and mechanisms.* Annu Rev Cell Dev Biol, 2006. **22**: p. 457-86.
11. Cohen, G.H., et al., *Water molecules in the antibody-antigen interface of the structure of the Fab HyHEL-5-lysozyme complex at 1.7 A resolution: comparison with results from isothermal titration calorimetry.* Acta Crystallogr D Biol Crystallogr, 2005. **61**(Pt 5): p. 628-33.
12. Jackson, S.E., et al., *Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhibitor 2.* Biochemistry, 1993. **32**(42): p. 11259-69.
13. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
14. Ritchie, D.W., *Recent progress and future directions in protein-protein docking.* Curr Protein Pept Sci, 2008. **9**(1): p. 1-15.
15. Vakser, I.A. and P. Kundrotas, *Predicting 3D structures of protein-protein complexes.* Curr Pharm Biotechnol, 2008. **9**(2): p. 57-66.
16. Halperin, I., et al., *Principles of docking: An overview of search algorithms and a guide to scoring functions.* Proteins, 2002. **47**(4): p. 409-43.
17. Ester, M., et al. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. in *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining.* 1996. Portland.
18. Pandolfo, M., *Molecular genetics and pathogenesis of Friedreich ataxia.* Neuromuscul Disord, 1998. **8**(6): p. 409-15.
19. Pandolfo, M., *Molecular pathogenesis of Friedreich ataxia.* Arch Neurol, 1999. **56**(10): p. 1201-8.
20. Pandolfo, M., *Friedreich's ataxia: clinical aspects and pathogenesis.* Semin Neurol, 1999. **19**(3): p. 311-21.
21. Pandolfo, M., *Friedreich ataxia.* Arch Neurol, 2008. **65**(10): p. 1296-303.
22. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.
23. Lambert, C., et al., *ESyPred3D: Prediction of proteins 3D structures.* Bioinformatics, 2002. **18**(9): p. 1250-6.
24. Kelley, L.A., R.M. MacCallum, and M.J. Sternberg, *Enhanced genome annotation using structural profiles in the program 3D-PSSM.* J Mol Biol, 2000. **299**(2): p. 499-520.
25. Garland, S.A., et al., *Saccharomyces cerevisiae ISU1 and ISU2: members of a well-conserved gene family for iron-sulfur cluster assembly.* J Mol Biol, 1999. **294**(4): p. 897-907.
26. Neuvirth, H., R. Raz, and G. Schreiber, *ProMate: a structure based prediction program to identify the location of protein-protein binding sites.* J Mol Biol, 2004. **338**(1): p. 181-99.
27. Bradford, J.R. and D.R. Westhead, *Improved prediction of protein-protein binding sites using a support vector machines approach.* Bioinformatics, 2005. **21**(8): p. 1487-94.
28. Breitkreutz, B.J., C. Stark, and M. Tyers, *The GRID: the General Repository for Interaction Datasets.* Genome Biol, 2003. **4**(3): p. R23.
29. Gilbert, D., *Biomolecular interaction network database.* Brief Bioinform, 2005. **6**(2): p. 194-8.
30. Guldener, U., et al., *MPact: the MIPS protein interaction resource on yeast.* Nucleic Acids Res, 2006. **34**(Database issue): p. D436-41.
31. Hermjakob, H., et al., *IntAct: an open source molecular interaction database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D452-5.

32. Mishra, G.R., et al., *Human protein reference database--2006 update.* Nucleic Acids Res, 2006. **34**(Database issue): p. D411-4.
33. Prieto, C. and J. De Las Rivas, *APID: Agile Protein Interaction DataAnalyzer.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W298-302.
34. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
35. Sayle, R.A. and E.J. Milner-White, *RASMOL: biomolecular graphics for all.* Trends Biochem Sci, 1995. **20**(9): p. 374.
36. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.
37. Ausiello, G., G. Cesareni, and M. Helmer-Citterich, *ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure.* Proteins, 1997. **28**(4): p. 556-67.
38. Pedretti, A., L. Villa, and G. Vistoli, *VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs.* J Mol Graph Model, 2002. **21**(1): p. 47-9.
39. Ester, M., et al. *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. in *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*. 1996. Portland.
40. Walesiak, M. and A. Dudek, *clusterSim: Searching for Optimal Procedure for a Data Set*. 2007.
41. Barras, F., L. Loiseau, and B. Py, *How Escherichia coli and Saccharomyces cerevisiae build Fe/S proteins.* Adv Microb Physiol, 2005. **50**: p. 41-101.
42. Gerber, J. and R. Lill, *Biogenesis of iron-sulfur proteins in eukaryotes: components, mechanism and pathology.* Mitochondrion, 2002. **2**(1-2): p. 71-86.
43. Gerber, J., et al., *The yeast scaffold proteins Isu1p and Isu2p are required inside mitochondria for maturation of cytosolic Fe/S proteins.* Mol Cell Biol, 2004. **24**(11): p. 4848-57.
44. Muhlenhoff, U., et al., *Components involved in assembly and dislocation of iron-sulfur clusters on the scaffold protein Isu1p.* Embo J, 2003. **22**(18): p. 4815-25.
45. Pandolfo, M., *Friedreich ataxia.* Semin Pediatr Neurol, 2003. **10**(3): p. 163-72.
46. Delatycki, M.B., et al., *Direct evidence that mitochondrial iron accumulation occurs in Friedreich ataxia.* Ann Neurol, 1999. **45**(5): p. 673-5.

```
#ESCHERNG_VER 1

Sat Jul 12 17:57:50 2008


#DOCKING_INFO
Target file name..: "1BQL.pdb"
Probe file name...: "193L.pdb"
Solutions.........: 1000
Center (x, y, z)..: -0.294500 22.750500 19.122499
#END

#SOLUTIONS
```

Sol.	Score	Rms	Bumps	Chg.	Pos.	Neg.	Apo.	Pol.	RotX	RotY	RotZ	TransX	TransY	TransZ
1	553	20.9	56	-66	3	18	3	54	17	-43	69	-4.9	-2.3	-17.3
2	550	21.9	71	-87	0	12	0	75	26	-43	72	-4.0	-2.5	-18.6
3	544	21.6	77	-83	0	6	0	77	23	-37	72	-4.2	-2.8	-18.3
4	542	23.7	187	-188	0	26	0	162	35	-40	81	-1.7	-2.8	-20.4
5	540	21.7	75	-86	0	7	0	79	23	-37	70	-4.2	-2.7	-18.5
6	540	22.9	145	-152	2	25	1	130	32	-43	81	-1.8	-2.3	-19.4
7	537	21.4	63	-72	3	15	0	60	23	-43	70	-4.3	-2.6	-18.0
8	536	21.3	64	-83	0	15	0	68	20	-40	71	-4.3	-2.7	-17.8
9	535	21.0	60	-70	0	15	0	55	20	-40	69	-4.1	-2.3	-17.7
10	532	16.9	24	-29	0	6	0	23	11	-22	60	-7.9	-2.6	-12.2
11	532	22.0	83	-88	3	8	0	83	26	-40	72	-3.9	-2.7	-18.8
12	530	21.7	88	-96	2	14	0	84	26	-43	70	-2.5	-2.1	-18.6
13	528	21.5	70	-78	0	7	0	71	20	-34	72	-4.6	-2.9	-18.0
14	527	23.3	164	-177	0	28	0	149	35	-40	74	-1.3	-2.3	-20.4
15	527	24.2	180	-181	2	26	0	157	38	-40	81	-1.8	-2.4	-21.1
16	527	23.7	207	-219	0	28	0	191	44	-46	82	1.0	-2.1	-20.2
17	526	19.3	23	-24	0	0	0	24	20	-28	68	-4.9	-2.1	-15.7
18	526	19.4	27	-27	0	2	2	27	17	-28	69	-5.4	-2.6	-15.6
19	525	18.7	0	-3	2	0	2	7	11	-28	59	-7.7	-2.5	-14.8
20	525	17.5	7	-14	0	0	0	14	8	-28	60	-7.9	-2.4	-12.9

Fig1. One of the most important output windows of *DockAnalyse* which shows the clustering graph of the 1000 docking solutions tested where the axes are the two extracted components of the computed Principal Components Analysis (PCA). The clusters found by the program are depicted in different colors and the representative points of each cluster are highlighted.

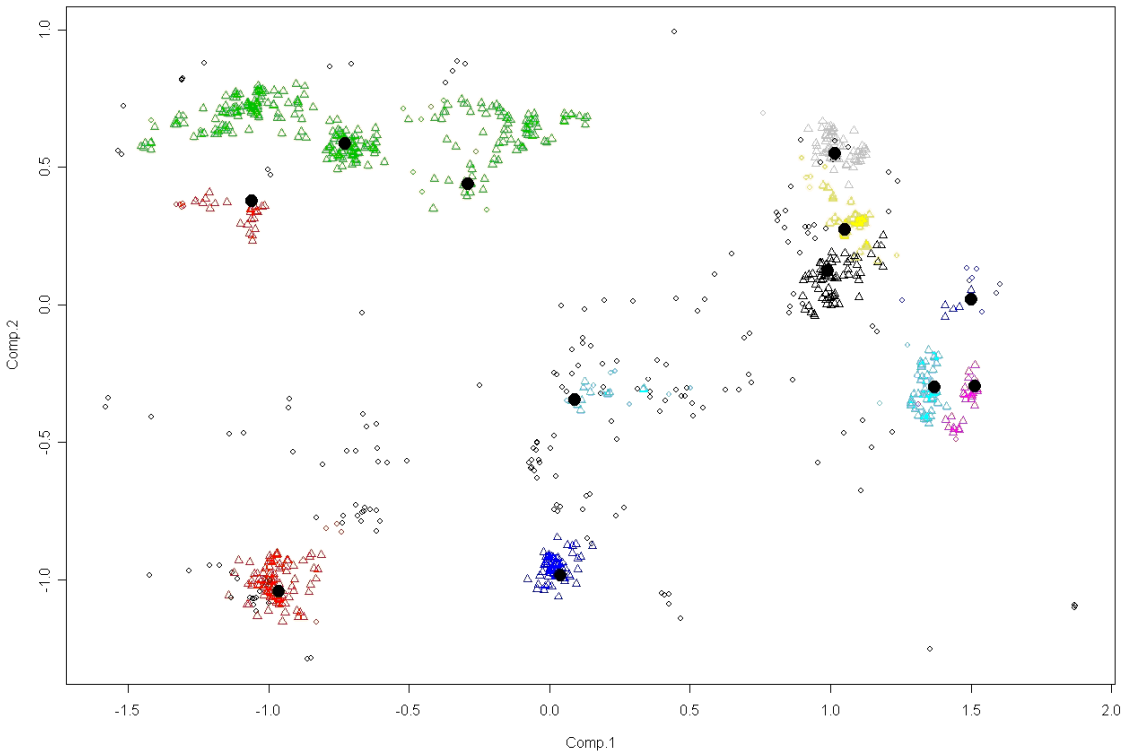

Fig2. Section of the Escher NG output datafile that results from the execution of the docking experiment between the proteins PDB ID: 1BQL and PDB ID: 193L. The file contains information for 1000 solutions (rows) where each of them has data divided into 14 sections (columns) giving information for each of the docking structures tested during the docking assay. This information is divided into: solution, score, root mean square, collisions, total charge score, positive<->negative charge score, positive<->positive and negative<->negative charge score, apolar score, polar score, X-Y-Z rotation and X-Y-Z translation (see Escher NG manual).

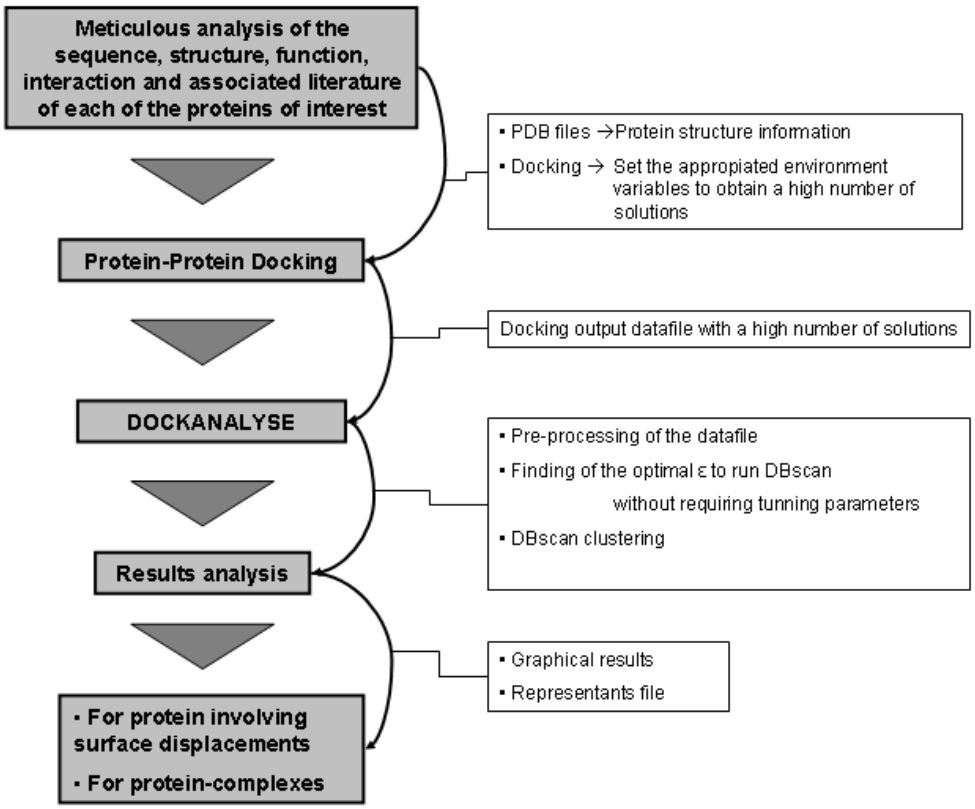

Fig3. Schematic chart flow where the sequential steps of the complete bioinformatics study in which *DockAnalyse* can be used are described.

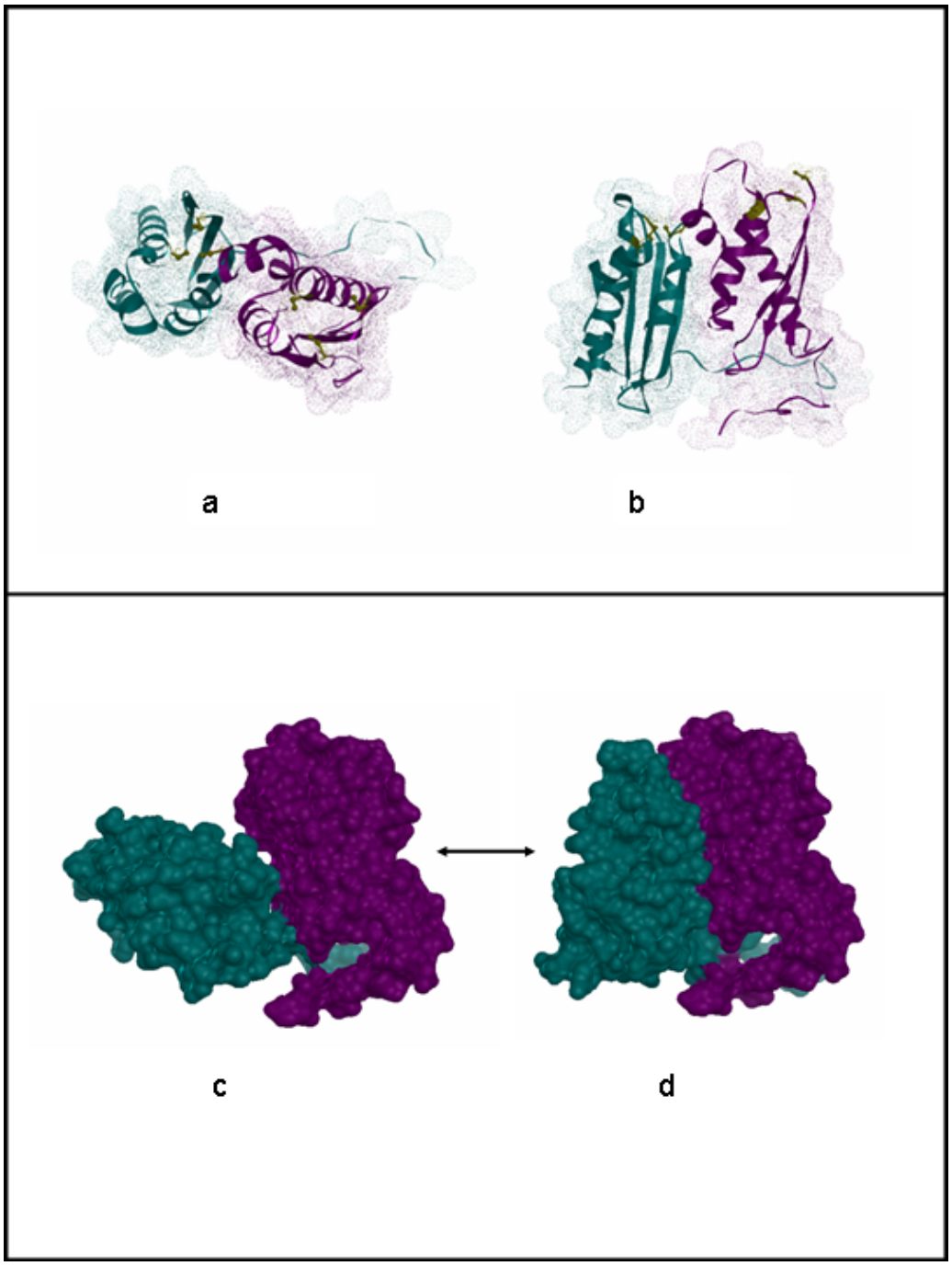

Fig4. The images (a) and (b) represent the tri-dimensional structure of the first ranked docking solution between the yeast proteins Isu1 and Isu2. The structures are displayed in "ribbons" and colored in magenta and cyan, respectively. Moreover, the calculated surfaces are depicted as dotted in both cases. In (a), a top view of the protein-complex structure is shown, where the atoms and bonds of the two iron-binding pockets (6 cys in total and 3 residues per protein) of each protein are colored in yellow in the "ball and stick" format. In (b), a side view of the same protein-complex is shown, where the tails, which have been predicted as interaction sites, are clearly visible at the bottom of the protein-complex structure. The images (c) and (d) represent the postulated dynamic interaction mechanism between the yeast proteins Isu1 (magenta in solid surface) and Isu2 (cyan in solid surface). These two images are the real tri-dimensional structures of the solutions given by *DockAnalyse* for the interaction between Isu1 and Isu2. According to the biologically expected meaning, frame (c) corresponds to the 3rd *DockAnalyse*-ranked solution and the frame (d) to 1rst. The edges intend to show the sequence of frames that may occur when these proteins interact, taking into account that the passing between frames (c) and (d) is putatively reversible, also according to the expected biological function.