Bioinformatics Approaches to Protein Interaction and Complexes: Application to Pathogen-Host Epitope Mimicry and to Fe-S Cluster Biogenesis Model

Isaac Amela Abellan

PhD Thesis

Institut de Biotecnologia i de Biomedicina
Universitat Autònoma de Barcelona

June 2013

Universitat Autònoma de Barcelona

Bioinformatics Approaches to Protein Interaction and Complexes: Application to Pathogen-Host Epitope Mimicry and to Fe-S Cluster Biogenesis Model

Submitted for the Degree of Doctor of Philosophy in Biotechnology by
Isaac Amela Abellan

The work was supervised by Dr. Enrique Querol Murillo and
Dr. Juan Cedano Rodríguez of the Institut de Biotecnologia i de
Biomedicina of the Universitat Autònoma de Barcelona

Isaac Amela Abellan

Enrique Querol Murillo

Juan Cedano Rodríguez

Bellaterra (Cerdanyola del Vallès), June 2013

AGRAÏMENTS:

Als meus companys de laboratori per compartir les estones del dia a dia, a en Juan Cedano pels seus consells transcendentals i a l'Enrique Querol per acceptar-me al laboratori amb els handicaps que això comporta.

A tots els membres de l'IBB, però en especial als del grup d'Immunologia Cel·lular amb els que he compartit la majoria de moments fora d'hores de feina. Aquí faig menció especial a la Cris per ser allà en molts moments.

Sobretot agraeixo als meus pares per l'educació rebuda i per l'enfocament de la vida que he rebut tot i les dificultats, a la meva germana pel suport rebut en tots els aspectes i a la meva dona per estar sempre al meu costat en el dia a dia.

TABLE OF CONTENTS:

LIST OF ABBREVIATIONS:

BCR: B-cell Receptor.

BioGRID: Biological General Repository for Interaction Datasets.

BLAST: Basic Local Alignment Search Tool.

BLASTP: Protein BLAST.

BOND: Biomolecular Object Network Databank.

CAPRI: Critical Assessment of PRedicted Interactions.

DBscan: Density-based spatial clustering of applications with noise.

DIP: Database of Interacting Proteins.

EBI: European Bioinformatics Institute.

ED: Extensive sequence Databasse.

EM: Electron Microscopy.

EMBL: European Molecular BiologyLaboratory.

ExPASy: Expert Protein Analysis System.

FFT: Fast Fourier Transform.

FRDA: Friedreich's Ataxia.

HAMAP: High-quality Automated and Manual Annotation of Proteins.

HPRD: Human Protein Reference Database.

HRTEM: High-Resolution Transmission Electron Microscopy.

IBB: Institut de Biotecnologia i Biomedicina.

ID: Initial sequence Database.

IEDB: Immune Epitope DataBase.

IntAct: EMBL-EBI Database of Protein InterActions.

ISC: Iron-Sulfur Cluster.

LF: Last File.

MIPS: Mammalian Protein-Protein Interaction Database.

MINT: Molecular Interaction database.

ORF: Open Reading Frame.

PDB: Protein Data Bank.

PPD: Protein-Protein Docking.

PPI: Protein-Protein Interaction.

NCBI: National Center for Biotechnology Information.

NMR: Nuclear Magnetic Resonance.

RMSD: Root Mean Square Deviation.

TAP: Tandem Affinity Purification.

TCR: T-cell Receptor.

UAB: Universitat Autònoma de Barcelona.

XRC: X-Ray Crystallography.

Yfh1: Yeast Frataxin Homolog 1.

Y2H: Yeast Two-Hybrid.

PRELIMINARY CONSIDERATIONS:

Since approximately 1985, the Group of Molecular Biology has been using bioinformatics for the analysis of protein structure/function relationship and also for the study of gene expression. Initially, bioinformatic tools were used to help protein engineering and design, but later bioinformatics became a principal objective and, thus, several programs and databases have been designed by the team. At present, the main objectives of our bioinformatics team are:

a) The help in Proteogenomics experimental investigations in order to re-annotate *Mycoplasma genitalium* genome and proteome. This organism is a very good model of a minimal cell/genome.
b) The identification of targets for the design of vaccines by *reverse vaccinology*.
c) The bioinformatics identification of moonlighting (multitasking, multifunctional) proteins.
d) The development of new statistical methods and programs for analysis protein characteristics and gene expression as a tool for target and drug discovery.
e) The work on protein sequence, structure, function and interaction in general. Specially applied to rare diseases and their putative therapy.

When I came into the group, one of the main issues in which the laboratory was interested was the design of vaccines using recombinant DNA and *reverse vaccinology*. A key problem of these strategies is focused on designing a good strategy to identify which of the many pathogenic proteins are important for vaccine design. The general issues commonly questioned are: How the host chooses pathogen targets to

elicit a protective immune response? There are two sub-questions to answer: First, why the host immune system rejects some pathogen targets? And second, which pathogen protein characteristics make them eligible for eliciting the host immune response? About this point was my first introduction to bioinformatics where we established approaches for a better identification of pathogenic antigens that avoid the host autoimmune response. This corresponds to Chapter I of the thesis, which is based on the article that was published as: Amela I. et al. *Pathogen proteins eliciting antibodies do not share epitopes with host proteins: A Bioinformatics approach.* PLoS ONE, 2007.2: e512. A method to identify potential protective antigens from the pathogen proteome would be very helpful and, therefore, some work is in progress to try to answer this question.

This first experience helped me to better understand how the transition from the sequence of a protein to its structure occurred, enabling me to work with leading bioinformatic applications to address a problem that worried me: My own disease, Friedreich's Ataxia. To go further in this problem, we realized the need of understanding how the different elements that constitute the key protein complex involved in this disease, which is the ISC biogenesis protein complex, interact with each other. This entailed the modeling of the 3D structures of the individual components of protein complex as well as its dynamic interactions by means of docking techniques. When working with these programs we figured out that manipulating such high number of solutions was not an easy task and, on the attempt to systematize the problem, an application that can be useful in analyzing docking solutions was developed (Chapter II of the thesis, which is based on the article that was published as: Amela I. et al. *DockAnalyse: an application for the analysis of protein-protein interactions*. BMC Structural Biology, 2010,10:37). A modification of *DockAnalyse* can be done in order to directly work with the resulting PDB files of the docking assays in spite of using the docking

output text file. This may be easier and portable for he users.

Putting together the observations found working with docking, *DockAnalyse* and other protein modeling tools, in terms of the ISC biogenesis protein complex, with the data provided by many references we were able to merge the protein complex model and the expected biological function. Summarizing we could solve the dynamic process by which several proteins, among which is the Friedreich's Ataxia causing protein (Frataxin), interact together to form the protein complex that assemble ISCs in the cell. Not only these studies can contribute to the current knowledge about Friedreich's Ataxia pathophysiolgy, but also give insight into one of the most important ways by which essential components in many red-ox reactions in the cell are generated (Chapter III of the thesis, which is based on the article that was published as: Amela I. et al., *A Dynamic Model of the Proteins that Form the Initial Iron-Sulfur Cluster Biogenesis Machinery in Yeast Mitochondria.* The Protein Journal, 2013, 2(3):183-96. This work can be amplified by modeling the different protein complexes that are formed in the different stages of the ISC biogenesis process.

Another work in progress regarding Friedreich's Ataxia, involves a power statistical analysis tool that was previously developed in our group, PCOPs (Principal Curves of Oriented Points), which I am using to find Frataxin co-expressed genes and also for the search of putative existing drugs that modulate these genes and might be used in a repositioning strategy.

Frataxin function has been associated to many processes related with iron binding, but other functions such as mitochondrial organization, mitochondrial dysfunction or lipid metabolism are now emerging. Our group has been working with moonlighting proteins during many years and we think whether Frataxin could be one of these proteins.

ABSTRACT

Bioinformatics is the use of computers in the field of biology to analyze amounts of data and generate new hypothesis. Along with many other things, protein sequence, structure, function and interactions can be studied with some current bioinformatic programs. Bioinformatics includes the design of new algorithms to analyze all of the previously mentioned protein properties and many other biological processes and data.

This thesis is structured in three main chapters, based one the use of different bioinformatic approaches to some biological problems related with human diseases. The present work has been done at the Institut de Biotecnologia i Biomedicina (IBB) of Universitat Autònoma de Barcelona (UAB).

CHAPTER I: Pathogen-Host Epitope Mimicry.

This section describes a work where sequence-based bioinformatic techniques were used for vaccine design.

The identification of epitopes eliciting antibodies in proteins is a fundamental, preliminary step in designing effective vaccines for human infectious diseases. Thus, the best way to prevent this diseases caused by human pathogens is by the use of vaccines. Autoimmune diseases caused by epitopes that generate auto-antibodies are highly important in immunology. In 2000, Rappuoli et al. described for the first time ever the term *reverse vaccinology* as the use of computers to rationally design vaccines starting with information present in the genome. The most common strategy to apply *reverse vaccinology* is by designing subunit recombinant vaccines, which usually generate humoral immune response due to B-cell epitopes in proteins. A major problem for this strategy is the identification of the few protective immunogenic proteins from the surfome of the pathogen. Epitope mimicry may lead to auto-immune phenomena related to several human diseases. A sequence-based bioinformatic analysis was carried out and two huge databases

were created, one with the most complete and current linear B-cell epitopes and the other one with the surface-protein sequences of the main human respiratory bacterial pathogens. We found that none of the 7353 linear B-cell epitopes analysed were found to share any sequence identity region with human proteins capable of generating antibodies, and only 1% of the 2175 exposed proteins analysed contained a stretch of shared sequence with the human proteome. These findings suggested the existence of a mechanism to avoid autoimmunity. Furthermore, a strategy for corroborating or warning about the viability of a protein linear B-cell epitope as a putative vaccine candidate in a *reverse vaccinology* study was also proposed. In this strategy, epitopes without any sequence identity with human proteins should be very good vaccine candidates for human diseases, and the other way round.

CHAPTER II: Analysis of Protein Interactions.

This section presents the process followed to design a new program for protein docking analysis.

Continuing with human diseases, many of them are related with protein function defects and, more precisely, with the formation of protein complexes. These protein complexes are assembled by protein-protein interactions (PPIs), which are the way by which most of the proteins fulfill their function. Protein monomers alone, in many cases, do not have a specific function which is only achieved when the distinct parts interact together to accomplish a certain function. Due to PPIs, it is expected that in the near future the number of protein complexes will surpass the number of proteins in some organisms. Therefore, interactomics represents one of the current frontiers of biosciences. From a bioinfomatics point of view, the method to predict the best way by which proteins interact is called protein-protein docking (PPD). But, is it possible to identify what the best solution of a docking program is? The usual answer to this question is the highest score solution, but

interactions between proteins are dynamic processes, and many times the interaction regions are wide enough to permit PPIs with different orientations and/or interaction energies. In some cases, as in a multimeric protein complex, several interaction regions are possible among the monomers. These dynamic processes involve interactions with surface displacements between the proteins to finally achieve the functional configuration of the protein complex. Consequently, there is not a static and single solution for the interaction between proteins, but there are several important configurations that also have to be analyzed. To extract those representative solutions from the docking output datafile, an unsupervised and automatic clustering application, called *DockAnalyse*, was created. This application is based on the already existing DBscan clustering method, which searches for continuities among the clusters generated by the docking output data representation. The DBscan clustering method is very robust and, moreover, solves some of the inconsistency problems of the classical clustering methods like, for example, the treatment of outliers and the dependence of the previously defined number of clusters. *DockAnalyse* makes the interpretation of the docking solutions, through graphical and visual representations, easier and guides the user to find the representative solutions. This new approach was applied to analyze several protein interactions and, therefore, model the dynamic protein interaction behavior of the protein complex of Chapter III of this thesis. *DockAnalyse* might also be used to describe interaction regions between proteins and, therefore, guide future flexible dockings. The application (implemented in the R package) is accessible.

CHAPTER III: Iron-Sulfur Cluster Biogenesis and Friedreich's Ataxia.

This section details the application of different bioinformatic tools to study the Iron-Sulfur Cluster biogenesis protein complex, which is clue in Friedreich's Ataxia.

One of those human diseases caused by a deficit in a protein function and that this protein seems to participate in the formation of a protein complex is Friedreich's Ataxia (FRDA). This syndrome is a human neurodegenerative and hereditary disease which mainly affects the equilibrium, coordination, muscles and heart. It is the most common autosomal recessive ataxia, and it is associated with a pronounced lack of a protein named Frataxin. This protein has been associated with iron inside the mitochondria, and it also seems to play an important role in the assembly of the mitochondrial Iron-Sulfur clusters (ISCs). High similarities have been suggested between the human and yeast molecular mechanisms that involve Frataxin. Moreover, in yeast, it has been demonstrated experimentally that Yeast Frataxin Homolog 1 (Yfh1) interacts with the protein Isu, which also interacts with the protein complex Nfs1-Isd11. Together, this set of proteins might generate the central platform for ISC biogenesis. Protein function involves interaction with other protein partners, however, not enough is known about the structure of the complex in which Frataxin works. The objective of this work was to model that complex in order to gain insight into its biological function. This objective was accomplished by the application of some bioinformatic tools, different protein docking programs and exhaustive clustering analyses of the docking results like that designed in Chapter II of this thesis. The structure of the protein complex and the dynamic behavior of its components, along with that of the iron and sulfur atoms required for the ICS biogenesis, were suggested. That hypothesis might be a seed to better understand the function and molecular properties of Frataxin and its protein partners.

Therefore, it may contribute to finally solve the exact ISC generation procedure and it could also be helpful for future treatment of FRDA.

The three main sections presented in this thesis have produced the following manuscripts:

Chapter I:

- **Amela I.**, Cedano J. & Querol E. *Pathogen proteins eliciting antibodies do not share epitopes with host proteins: A bioinformatics approach.* **PLoS ONE, 2007**. 2: e512.

Chapter II:

- **Amela, I.**, Delicado P., Gómez A., Bonàs S., Querol E. & Cedano J. *DockAnalyse: an application for the analysis of protein-protein interactions.* **BMC Structural Biology, 2010**. 10: p. 37.

Chapter III:

- **Amela I.**, Delicado P., Gómez A., Querol E. & Cedano J. *A Dynamic Model of the Proteins that Form the Initial Iron-Sulfur Cluster Biogenesis Machinery in Yeast Mitochondria.* **The Protein Journal, 2013,** 2(3):183-96.

These three papers are those where I contributed substantially, although during these years as a PhD student other works have been also published.

It must be taken into account that in Chapter II both Amela I. and Delicado P. contributed equally to the work. Dr. Delicado is a mathematician of the Universitat Politècnica de Catalunya (UPC) with which we usually collaborate. He basically programmed the application and Isaac Amela designed and tested *DockAnalyse*.

BIOINFORMATICS

Definition

Bioinformatics is the field where biology, computer science and information technology merge to create a discipline with solutions to biological problems. This scientific area involves, among many other things, databases, algorithms, modeling and simulations. Taking into account that bioinformatics is a theoretical science, the resulting hypothesis always have to be experimentally corroborated.

Approaches used in Chapter I

One of the branches of bioinformatics has to do with the creation, growth and maintenance of databases with biological information. These databases contain many types of scientific information that come from different sources, such as laboratory experiments or computational analyses mainly based on the called *omics* areas. Normally, each database entry is described by a unique accession number and, moreover, the data is usually structured in tables or easy-treatment frameworks, like for instance "comma separated values", that facilitate the search and manipulation of the information contained in that database. Some of the most commonly used biological databases, which are specially designed for researchers, are for example:

- PubMed → an open access database comprising citations for life science journals, online books and biomedical literature.
- GenBank → an open access collection of nucleotide sequences.
- UniProt → an open access database of protein sequences.
- Protein Data Bank (PDB) → an open access Information Portal to Biological Macromolecular Structures of proteins and nucleic acids.

The journal *Nucleic Acids Research* (NAR) edits every year a special issue that is only focused on open access biological databases. Below, two plots that demonstrate the importance and constant growth of these repositories of scientific information are shown as an example.

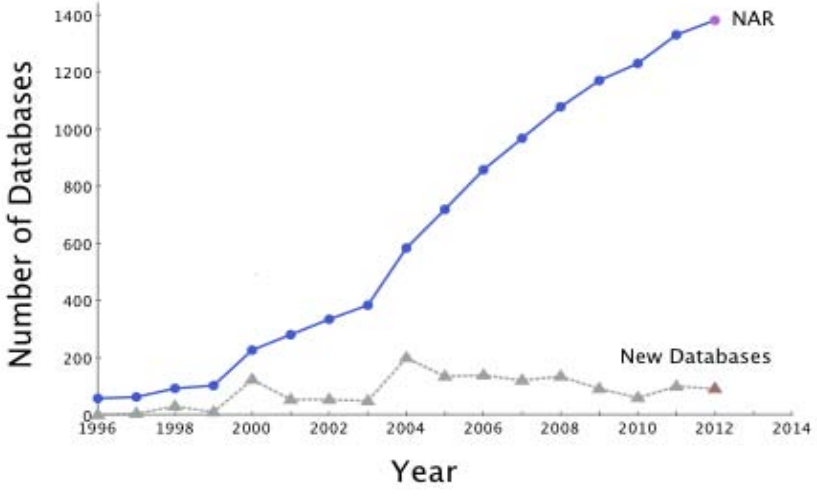


Figure 0-1. A general view of biological database growth. The number of existing databases (blue) and that of new databases (grey) over the years is shown here (Geospiza 2012).

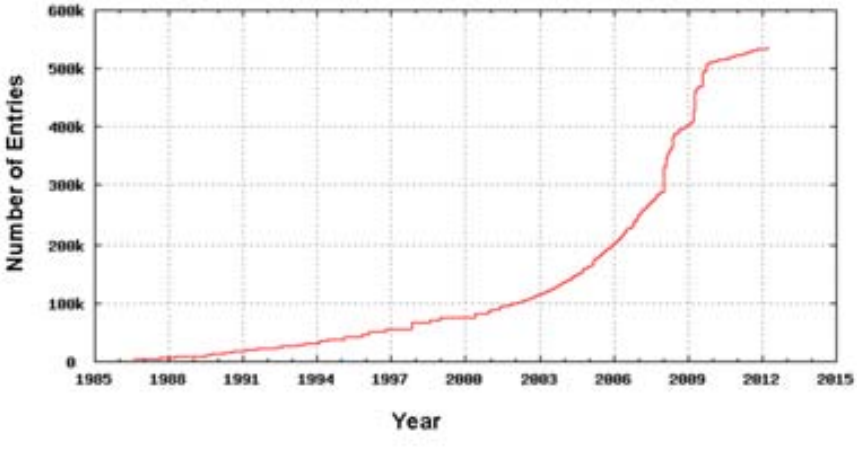


Figure 0-2. An example of a particular database growth. The number of entries in UniProtKB/Swiss-Prot over the years is shown here (ExPASy release 2012_03).

These and many other existing databases allow researchers to access the existing information and assist them in their investigations. Furthermore, new entries to a certain database might be submitted and, in fact, when researchers publish a paper in a journal, this article and its contents are automatically deposited in one of the current bibliography databases. Nowadays, many efforts are focused in developing tools to capture and analyze the data of these biological databases, which might be important for our investigations to generate new hypothesis. Some of this biological data might be DNA or protein sequences. If you want to compare sequence information from a bioinformatics point of view, the Basic Local Alignment Search Tool (BLAST) is the algorithm desigjned for this purpose [Altschul et al., 1990]. Roughly, BLAST relates DNA or protein sequences and compares a query sequence with a sequence database. This program searches for sequences in the database that are similar to the query sequence. There are different types of BLAST depending on the type of query sequence and BLASTP is the one most commonly used when working with protein sequences. In Chapter I of this thesis, these bioinformatic approaches are used in an immunology context to ease the decision making process in vaccine design studies.

Approaches used in Chapter II

Different information could be generated when applying whatever bioinformatic program and, thus, the development of new algorithms and programs to manage that data is another area in which bioinformatics can be applied. That is a similar approach to the previously mentioned database management, but in this case the data is generated by a bioinformatics program running locally in our computer rather than obtaining the data retrieved from a biological information server. In Chapter II of this thesis, the creation of a new program to treat

the data created locally by a bioinformatic program is addressed. In this case, the huge amount of data is generated by a docking program and we want to facilitate the selection of the best docking solutions as the base for the work in Chapter III.

Programming languages

Both in Chapter I and Chapter II, extensively used programming languages in bioinformatics are required. In particular, Perl, R and Shell programming languages have been used along the thesis. A figure showing the programming language usage in the field of bioinformatics is as follows:

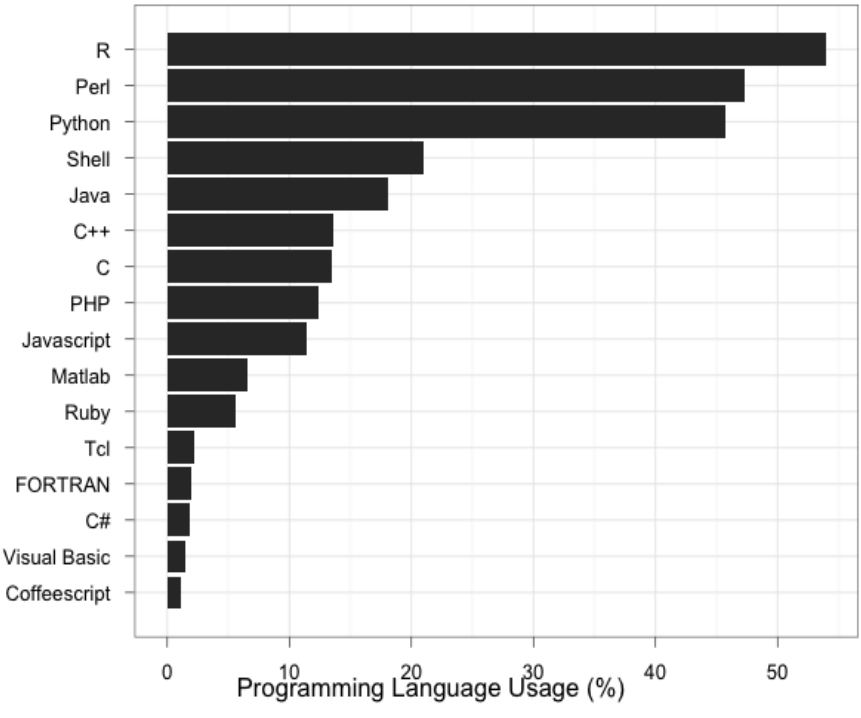


Figure 0-3. Programming language usage in bioinformatics. This figure shows the relative percent of programming language usage in the bioinformatics community according to the 2012 survey (http://bioinfsurvey.org).

Perl is one of the most popular programming languages in boinformatics because it allows an easy treatment of strings trough its regular expression and dynamic scripting potential. DNA or protein sequences are, in fact, concatenations of letters (strings). Many modules to easily work with bioinformatic resources that are based on Perl have been created, for example BioPerl. Regarding R, this programming language is now becoming more and more used due to its statistics capacities. Taking into account that Linux distributions based on the Unix operating system are widely used for bioinformaticians, Shell scripting represent a key tool considering the high command line usage in this Linux environments. Summarizing, Perl was used basically in Chapter I, R in Chapter II and Shell whenever it was required.

Approaches used in Chapter III

Plenty of bioinformatic web services designed to study different protein properties are currently available online. These tools can be used to bioinformatically analyze different sequence characteristics, structure traits, function attributes, interaction features and current bibliography of a certain protein. Some of these services are grouped in reference web pages especially dedicated to these purposes like for instance the National Center for Biotechnology Information (NCBI), Expert Protein Analysis System (ExPASy) and European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), which is shown below:

Figure 0-4. EMBL-EBI web page. A portion of the EBI home page is shown (http://www.ebi.ac.uk).

Another area of bioinformatics concerns protein structure prediction and molecular interaction prediction and is termed Structutal Bioinformatics. These approaches contain protein modeling and protein docking techniques. The modeling techniques predict the structure of a protein from its sequence while docking tools try to guess the way by which proteins interact. In Chapter III of this thesis, the use of the previously stated web services and these modeling/docking tools are described. It must be taken into account that the program developed in Chapter II was used as the key tool in this part of the thesis.

CHAPTER I:

Pathogen-Host Epitope Mimicry

INTRODUCTION:

This first chapter is related with the previously mentioned bioinformatics branch, which deals with the handling of biological databases to postulate new hypothesis. Particularly, the bioinformatic approaches used in this section comprise from literature analysis, Perl scripting, database management and data retrieval, to sequence analysis using the Protein BLAST (BLASTP).

The immune system

The initial part of this thesis deals with vaccinology and especially with the new branch of *reverse vaccinology*. Therefore, to focus this part some clue aspects about this field and also about the immune system must be introduced. Generally speaking, the immune system is a group of biological processes that protect the organism against disease. There are two types of immune responses depending on which biological processes are involved:

- The innate immune response/system.
- The adaptative immune response/system.

The first type of response is present in most of the organisms, it deals with non-specific immediate responses and do not generate immunological memory. In contrast, the second type of response is only found in jawed vertebrates, deals with antigen-specific late responses and produce immunological memory. Both subsystems cooperate to fight against infections in humans. Regarding the adaptative immune

system, it is mediated by a type of white blood cells called lymphocites and is based on the specific recognition of certain pathogen structures called antigens. An antigen is a part of a molecule, typically of a pathogen protein, that is recognized by an antibody and stimulates its production. Antibodies or immunoglobulins are molecules produced by a type of lymphocites called B-cells that label pathogens to facilitate their attack by the immune system. This antigen/antibody-mediated response produced by B-cells is known as the humoral immune response.

Antigens are recognized by a B-cell receptor molecule termed BCR that is a membrane-bound antibody allowing for a direct recognition of the pathogen antigen. When an antigen is recognized by a BCR, that B-cell becomes activated and begins a differentiation to plasma B-cells, which produce soluble antibodies specific for that antigen, and memory B-cells, which remain in the body to faster recognize this same antigen and better respond to future infections. It is interesting that aged people who survived the deadly 1918 "Spanish influenza" yet present immunological memory to that virus, which was reconstructed by the team of Taubenberg recently.

From a general point of view, the most important lymphocites in the adaptative immune response are T-cells and B-cells. The first type of lymphocites is involved in the cell-mediated or cellular immune response, but they also contribute to the B-cell or humoral immune response trough direct cellular interactions or by the secretion of signaling molecules called cytokines. Antigens could be recognized by T-cell receptor molecules (TCR) or by B-cell receptor molecules (BCR, membrane-bound antibody) dealing with this two types of responses, which, as said before, are not independent but they share many

processes and they usually cooperate (See Figure I-1).

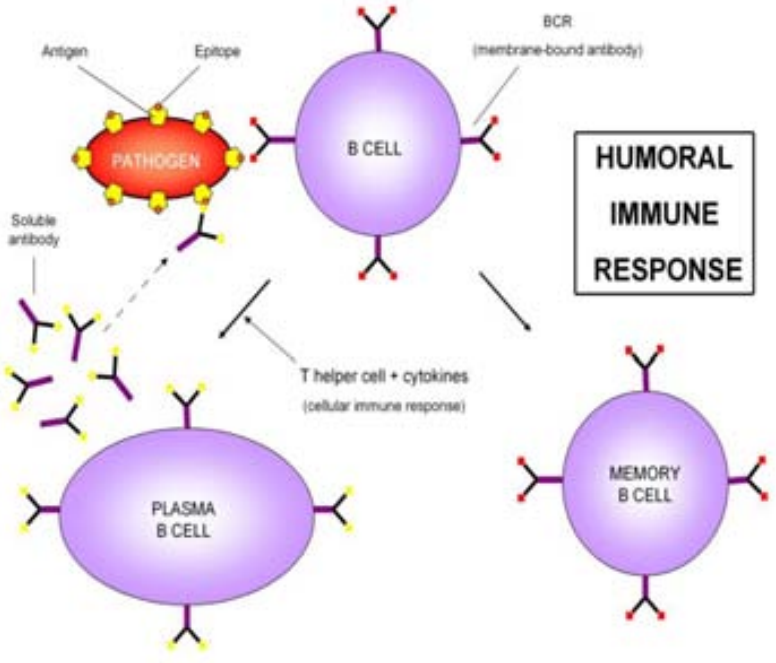


Figure I-1. Humoral immune response. Schematic representation of the B-cell mediated adaptative immune response.

More precisely, the portion of an antigen that is recognized by the antibody is called epitope. An epitope or determinant is a combination of amino acids of a protein and can be divided in these two categories (See Figure I-2):

- Linear or continuous epitopes (most common) → Amino acids that are sequential both in the tertiary structure and primary sequence of the protein.

- Conformational or discontinuous epitopes → Amino acids that are grouped in the tertiary structure of the protein but are not sequential in its primary sequence.

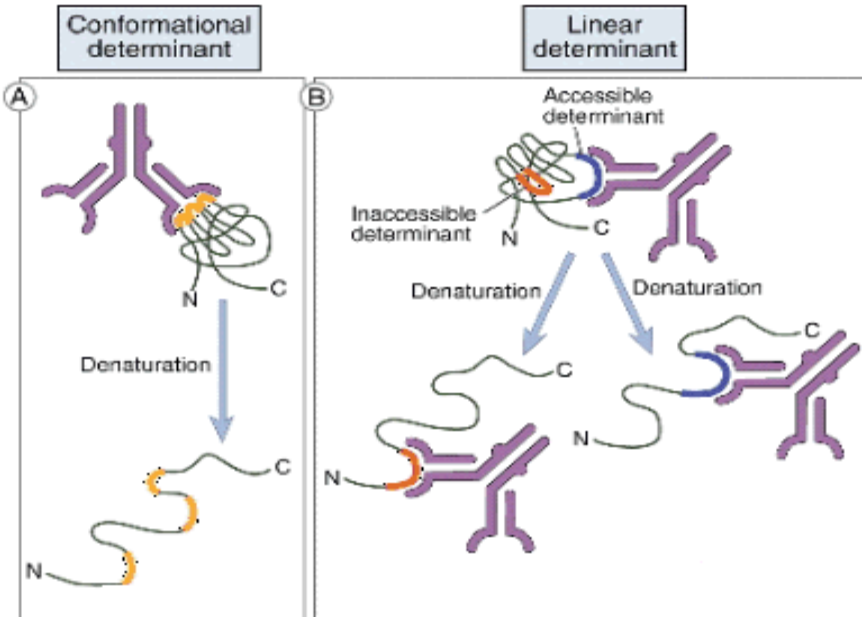


Figure I-2. The two types of existing epitopes/determinants. In this picture the essential differences between conformational and linear epitopes are shown (Elsevier Science 2003).

In addition to the differentiation between conformational and linear epitopes, immunologists refer to T-cell or B-cell epitopes depending on the type of cell that recognizes this epitope, a T-cell or a B-cell respectively. Owing to the direct correlation between antigenic recognition via B-cell epitopes, B-cell activation, antibody production and immune reponse, current vaccine development is mainly focused on finding pathogen protective antigens that contain one or more

accessible B-cell epitopes, which should evoke B-cell activation and produce protective antibodies. Most of the pathogen proteins that conform to these rules are those located in the surface of the pathogen, the *surfome*. It has been reported that some of the proteins of the *surfome* of a certain pathogen effectively deal with this type of response [Rodriguez-Ortega et al., 2006].

Vaccine design

Vaccination is the preventive method of choice to fight against microbial pathogens and presents the best cost/benefit ratio among current clinical and pharmaceutical practices. There are many reasons and serious threats which make the development of new advanced vaccines necessary, for example, avian flu and the spread of antibiotic-resistant strains of pathogens [Borchardt, 2004]. Vaccines can be divided in different groups depending on their development:

- Conventional vaccines:
    - Killed → with dead pathogens.
    - Attenuated → with live pathogens but without virulent properties.
    - Toxoid → with inactivated toxic proteins of the pathogen.
    - Subunit → with surface or exposed proteins of the pathogen.
    - Conjugate → with coverings and immunological proteins of the pathogen.

- Synthetic vaccines:
    - Recombinant → with the cell of one pathogen and the DNA of another one.
    - DNA → with the pathogen DNA.

The advent of genomics and high-throughput cloning/expression of large sets of genomic open reading frames (ORFs) from pathogens make genome-wide searches of new vaccine candidates possible. This systematic identification of potential antigens and virulence factors of a pathogen, without the need for its cultivation and using bioinformatic approaches has been termed *reverse vaccinology* [Rappuoli, 2000; Rappuoli and Covacci, 2003]. This method is not as time-consuming as the conventional one and it generally reduces the period of vaccine development more than five times. In Figure I-3 the comparison of these two methods can be seen:

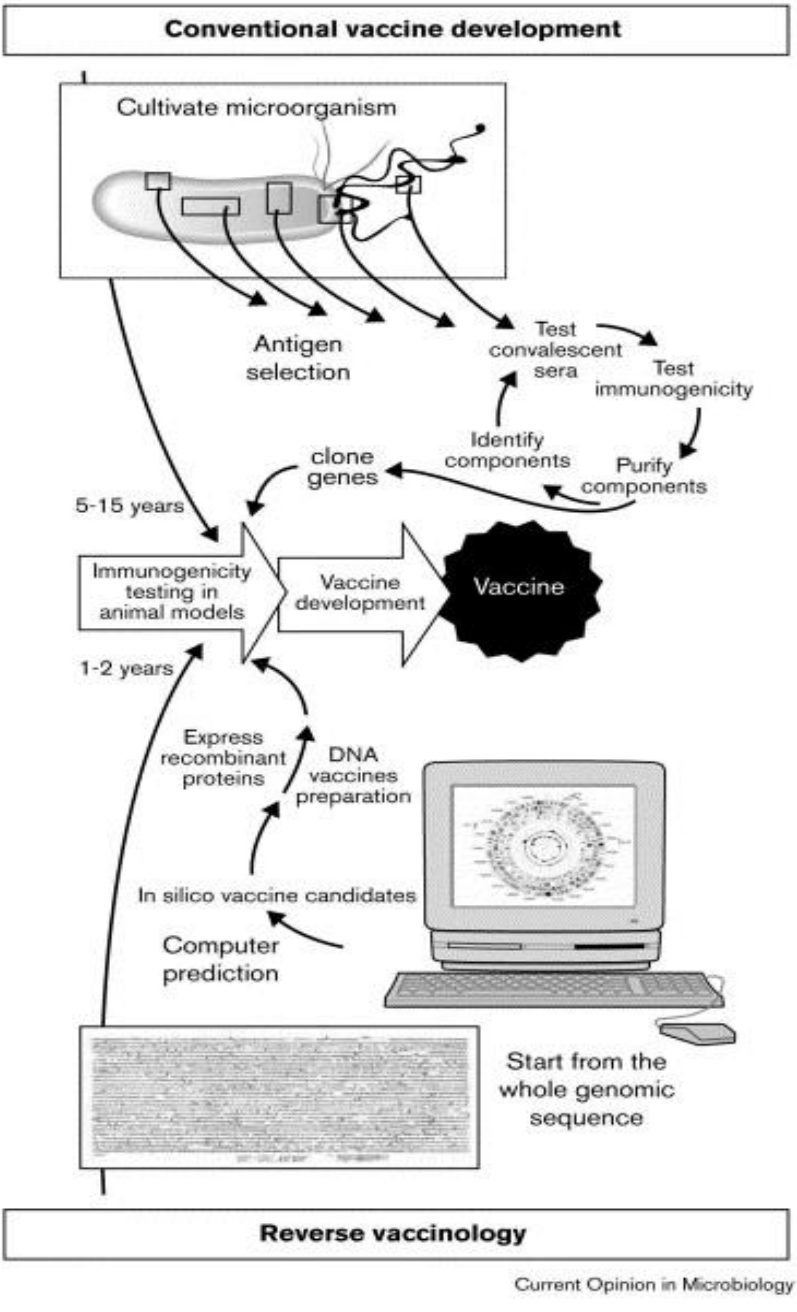


Figure I-3. Conventional vaccine development vs *reverse vaccinology*. Schematic representation of the essential steps of vaccine development by the conventional approach and by *reverse vaccinology* (Rappuoli 2000).

The objective of vaccine development is to find proteins eliciting antibodies capable of binding to the bacterial surface, and through interaction with the complement system, kill certain pathogen microorganisms. The complement system is a group of molecules that belong to the innate immune system and destroy the pathogen when this is labeled with specific antibodies. However, current large-scale antigen-screening studies show that only a small fraction of the pathogen proteins, most surface-exposed or secreted, appears to elicit antibodies with bactericidal activity [Poolman, 1995; van den Elsen et al., 1999; Rappuoli and Covacci, 2003]. It is generally considered that, in a bactericidal assay, an antigen that elicits *murine* antibodies capable of triggering bacterial cell death *in vitro* in a complement-dependent manner, is a good candidate for human vaccine development [Goldschneider et al., 1969; Pizza et al., 2000; Welsch et al., 2004; Rodriguez-Ortega et al., 2006]. A major obstacle to *reverse vaccinology*, besides sequence and antigenic variability, is the difficulty to identify, from the pathogen proteome, those proteins that will generate the wanted protective response.

As previously described, a linear or continuous B-cell epitope is a specific region of an antigen to which an antibody binds and its constitutive residues are sequential in the primary sequence of the protein. Moreover, it is generally assumed that a linear B-cell epitope is composed by a minimum of five sequential amino acids [Lucchese et al., 2007]. On the contrary, conformational or discontinuous B-cell epitopes are highly conformational-dependant and its constitutive residues are not sequential in the primary sequence of the protein. This makes the work with conformational epitopes almost unaffordable in *reverse vaccinology* due to the necessary use of protein sequences in this kind of studies. The lack of databases for conformational B-cell epitopes and the poorly developed method for predicting it from structure enforce this.

The low-simlarity hypothesis

This hypothesis postulates that pathogen protein sequences with zero or low similarity to the host proteome modulate the B-cell epitope pool in the humoral immune response. Consequently, those epitopes have the highest specificity and lowest cross-reactivity and that is why they should be taken into consideration when designing effective and safe vaccines. In summary, it can be said that immunogenicity is preferentially associated to low-similarity sequences, that is, B-cell epitopes [Kanduc, 2009].

Autoimmunity

The capacity to discriminate between self and non-self molecules is a key aspect that concerns the immune system promoting the attack only of foreign components and not of body structures. Failures in the ability to properly carry out this function result in autoimmune phenomena, which produce immune response against own constituents. In some occasions, autoimmunity takes place when a stretch of shared sequence, that could act as an epitope and is called *mimetope* (or *mimotope*), exists between a protein of a certain pathogen and a protein of its host. This event is known as *epitope mimicry*. Many of the autoimmune diseases are caused by pathogens that present this epitope mimicry [Benoist and Mathis, 2001]:

- *Borrelia burgdorferi* → Lyme Disease or Neuroborreliosis.
- *Streptococci* (several) → Rheumatic Fever.
- *Tripanosoma cruzi* → Chaga's Disease.
- *Campylobacter jejuni* → Guillain-Barré Syndrome.
- C*hlamydia pneumoniae* and a group of viruses → Multiple Sclerosis.

- B3 coxsakieviruses → Myocarditis.
- B4 coxsakieviruses or cytomegaloviruses → Type I Diabetes.
- Herpes virus 1 → Herpetic Stromal Kerartitis
- *Chlamydia pneumoniae* outer-membrane proteins mimicry to myosin.

For several of these kinds of autoimmune diseases, we do not yet know what the possible causing agent is like, for example, Primary Biliary Cirrhosis, Psoriasis, Scleroderma, Sjögren's Syndrome or Lupus.

OBJECTIVES:

- Check whether known epitopes from human respiratory pathogen-proteins that elicit the host immune response share, to some degree, amino acid sequence with host proteins.

- Propose some rules that should be taken into consideration in *reverse vaccinology* approaches and define strategies that should be followed in vaccine design studies.

MATERIALS AND METHODS:

General ideas and main concepts of the created databases

With the aim of proposing a method applicable in the previously described *reverse vaccinology* studies, databases were made of the exposed proteins from the up-to-date (at this moment, year 2005), sequenced main human bacterial respiratory pathogens, which are: *Neisseria meningitidis serogroup B*, *Legionella pneumophila* (Lens strain), *Streptococcus pneumoniae, Haemophilus influenzae, Pseudomonas aeruginosa, Streptococcus pyogenes* (serotype M1), *Yersinia pestis, Bordetella bronchiseptica*, *Staphylococcus aureus* (COL strain), *Pasteurella multocida, Bordetella parapertussis*, *Bordetella pertussis, Chlamydia pneumoniae* (or *Chlamydophila pneumoniae*) and *Mycoplasma pneumoniae*.

In addition, to see if our previously proposed theory postulating that some selectivity exists to try to avoid the auto-immune response, we obtained all of the available linear B-cell epitopes from human bacterial-pathogens of the three most complete and current epitope databases at this moment.

On one hand, an exposed-protein sequence database for each of the 14 pathogens analyzed was created downloading the protein sequences from the High-quality Automated and Manual Annotation of Proteins (HAMAP) system under the ExPASy web server [Gattiker et al., 2003; Gasteiger et al., 2003]. As well as that, exhaustive searches in NCBI resource (www.ncbi.nlm.nih.gov/entrez), in "SCIRUS for scientific information only" research tool (www.scirus.com) .and in specialized journals served us to include in these pathogen databases protein sequences with certain interest for our study because of their reported capacity to generate antibody immune response via B-cell epitope

activation. Summarizing, a total of 2175 proteins sequences from the surfome [Rodriguez-Ortega et al., 2006] of the studied pathogens were recruited.

On the other hand, we developed different computational scripts (See the "Programming languages" paragraph of the Bioinformatics section) to obtain the human bacterial-pathogen section of the Bcipep database [Saha et al., 2005], the most complete and specific B-cell epitope database available at this moment, from which 2275 linear B-cell epitopes were obtained. Moreover, we performed a similar procedure to collect the B-cell epitope portion of the IEDB database [Peters et al., 2005], an extensive immune epitope database appeared in early-2006 with a total of 2154 human bacterial-pathogen linear B-cell epitopes. Finally, we could computationally retrieve 2924 linear B-cell epitopes from human bacterial-pathogens of the AntiJen database [Toseland et al., 2005], an important immunological database that contains quantitative binding data for epitope peptides. Thus, a total of 7353 linear B-cell epitopes from human bacterial-pathogens were collected for the study (See Figure I-4). It has taken into account that most of that epitopes are of at least six or more amino acids length, which, as pointed out in the introduction, is considered to be the minimum length for an epitope.

Once the 14 databases containing the exposed protein sequences of the pathogens under study have been created, possible identity regions compared with the human proteome could be further analyzed using BLASTP. Regarding the 3 databases containing linear B-cell epitopes from human bacterial-pathogens, the same as before could be done to see if some of these B-cell epitopes have significant sequence identity to similar regions of the human proteome.

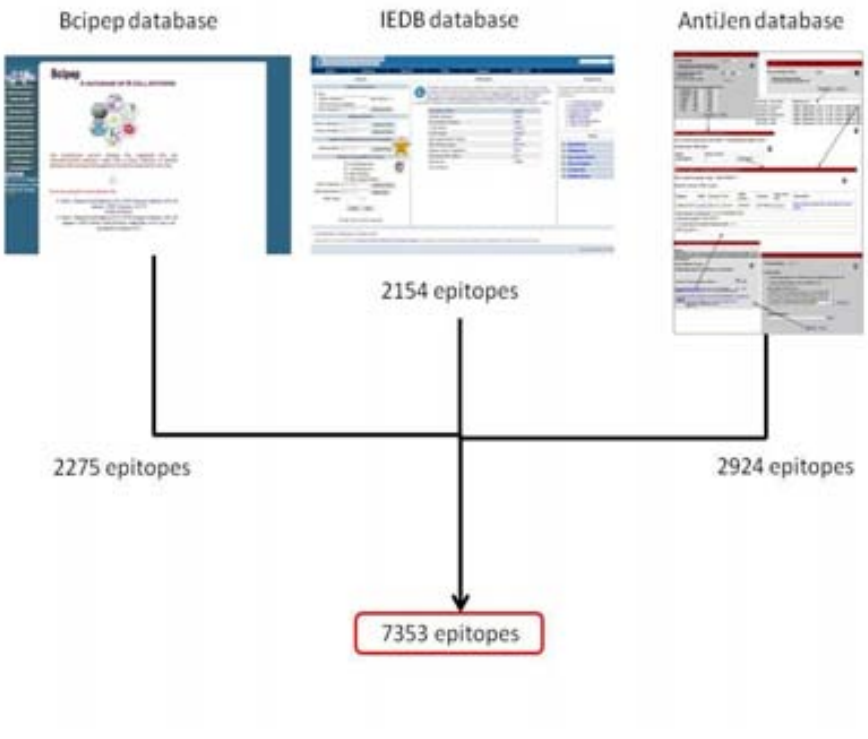


Figure I-4. B-cell epitopes sources. Images of the three databases used to capture all the linear B–cell epitopes.

Details of the exposed pathogen protein databases

For each of the pathogens analyzed, downloading was started from the HAMAP tool under the ExPASy web server of all protein sequences whose annotation included the following keywords: *Outer, membrane, lipoprotein, adhesin, surface, secreted or exposed*. Then, as said above, extensive searches were carried out through the NCBI web server, the SCIRUS web site and several specialized journals, with the aim of finding scientific articles talking about proteins that generate antibody immune response via B-cell epitope activation and could probably act as putative

vaccine candidates. An initial sequence database (ID) for every pathogen was made as explained before and was used to execute a BLASTP analysis [Altschul et al., 1990] against the protein non-redundant database at the NCBI ftp site (www.ncbi.nlm.nih.gov/Ftp). If sequence similarities were found at first glance between pathogen proteins and human proteins in the BLASTP output, a more extensive sequence database (ED) for each pathogen was generated. Again, the HAMAP tool under the ExPASy web server was used, but this time including all protein sequences that contained the words *hypothetical, probable, conserved* or *putative* in their annotation. These protein sequences ED files were employed to make another BLASTP analysis as explained before where each of the pathogens analyzed was carefully scrutinized to obtain a last file (LF) including the alignments that comprise local and significant sequence similarity between a pathogen protein and a human protein.

A stretch of shared sequence was considered as a putative *mimetope* (or *mimotope*) if there were at least the same five sequential residues or at least five sequential residues with relatively similar physico-chemical properties, which is the minimum length generally accepted for an epitope as pointed out in the introduction section.

For all of the fourteen pathogens analyzed, LF files of exposed protein sequences were created and, moreover, for each of these files we checked to see if these similar stretches correspond to transmembrane regions were checked because a B-cell epitope cannot exist in a transmembrane region. This was done by applying TransMem, a program for predicting transmembrane domains in proteins [Aloy et al., 1997]. We also checked to see if these stretches of shared sequence coincide with the signal peptide section of the protein, when these are close to the N-terminal extreme, using the SignalP 3.0 Server [Bendtsen et al., 2004]. Lastly, it was checked to determine if these stretches are

predicted as putative linear B-cell epitopes using prediction servers like ANTIGENIC [Kolaskar and Tongaonkar, 1990], ABCpred [Saha and Raghava, 2006], and BcePred [Saha and Raghava, 2004] (See Figure I-5).

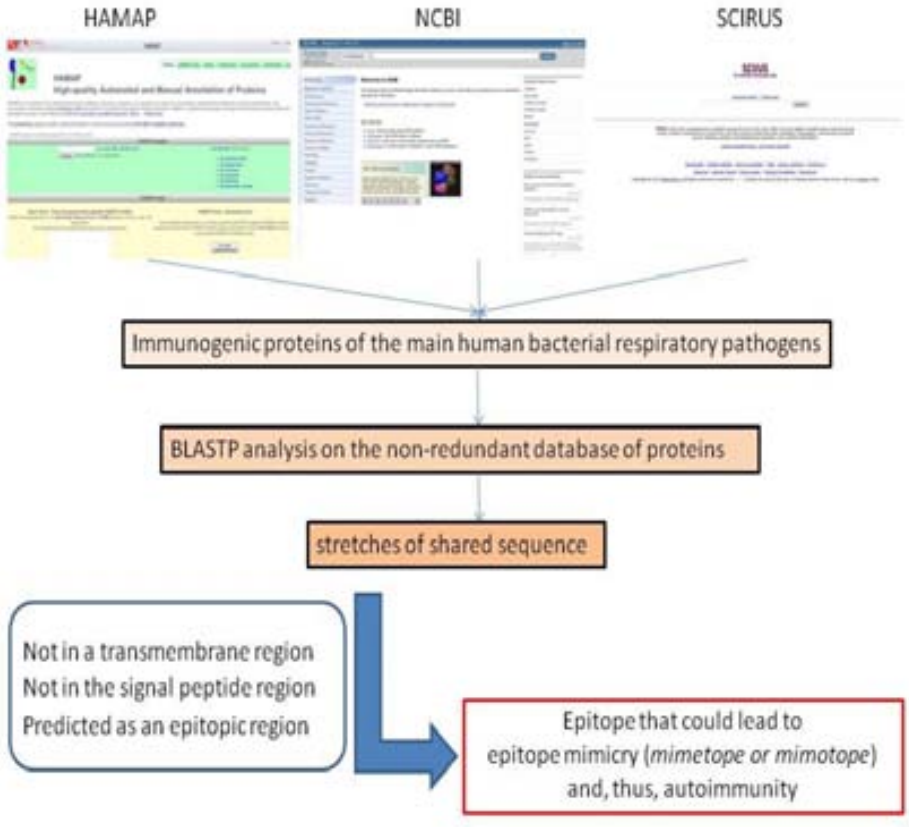


Figure I-5. Exposed protein study. Schematic flowchart of the procedure followed in the analysis of the exposed proteins.

Regarding *Neisseria meningitidis serogroup B*, the ID contained 62 protein sequences [Ala'Aldeen et al., 1994; Ala'Aldeen and Borriello, 1996; van den Elsen et al., 1999; Pizza et al., 2000; Turner et al., 2001; Vermont and van den Dobbelsteen, 2002; Fitzpatrick et al., 2005] and the ED comprised 263 protein sequences. In the case of *Legionella pneumophila* (Lens strain), the ID had 17

protein sequences [Weeratma et al., 1994; Cazalet et al., 2001; Chan et al., 2005] and the ED consisted of 352 protein sequences. The *Streptococcus pneumoniae* study allowed us to obtain an ID that comprised 27 protein sequences [Adamou et al., 2001; Wizemann et al., 2001; Chiavolini et al., 2003; Maione et al., 2005; Rodriguez-Ortega et al., 2006; Beghetto et al., 2006] and an ED of 338 protein sequences. Regarding *Haemophilus influenzae,* we were able to generate an ID of 25 protein sequences [Janson et al., 1993; El-Adhami et al., 1999; Zagursky et al., 2000; Liu et al., 2004; Murphy et al., 2006]. In this case, we did not search for more protein sequences because we did not find any previous sequence-identity region between the proteins from the ID and human ones, so the database remained consisting on a total of 25 protein sequences. Concerning *Pseudomonas aeruginosa,* we were able to assemble an ID of 34 protein sequences [Duchene et al., 1989; Moss et al., 2001; Price et al., 2001; Umelo-Njaka et al., 2001; Thomas et al., 2003; Corech et al., 2005; Goure et al., 2005; Worgall et al., 2005] and a huge ED that contained 569 protein sequences. The *Streptococcus pyogenes serotype M1* study allowed us to obtain an ID of 30 protein sequences [Guzman et al., 1999; McMillan et al., 2004; Banks et al., 2004; Okamoto et al., 2005; Vohra et al., 2005; Seepersaud et al., 2005; Rodriguez-Ortega et al., 2006; Abdissa et al., 2006; McArthur and Walker, 2006; Moyle et al., 2006] and an ED of 278 protein sequences. Regarding *Yersinia pestis*, we only created an ID of 32 protein sequences [Benner et al., 1999; Li et al., 2005; Philipovskiy et al., 2005]. We did not increase this ID because we did not find any region with a sequence identity between the proteins from this database and human proteins in the preliminary BLASTP analysis. Concerning *Bordetella bronchiseptica*, we could group together 46 protein sequences in its ID [Keil et al., 1999; Mattoo et al., 2000; Hausman and Burns, 2000]. Nothing else was added to this ID because we did not find stretches of shared sequence between the proteins from this database and human ones. In the case of *Staphylococcus aureus (strain COL)*, we were able to generate an ID of 43 protein sequences [Vytvytska et al., 2002; Weichhart et al., 2003; Viau and Zouali, 2005] and we could extend it

to 490 protein sequences, producing a large ED. Regarding *Pasteurella multocida* assembled an ID of 20 protein sequences [Lugtenberg et al., 1986]. Nothing else was added to this ID because we did not find any local sequence-identity regions between the sequences from this database and human proteins. We made the same analysis for the rest of the pathogens in our study, which are *Bordetella parapertussis*, *Bordetella pertussis, Chlamydia pneumoniae (Chlamydophila pneumoniae)* and *Mycoplasma pneumoniae*. In the cases of *Bordetella parapertussis* and *Bordetella pertussis,* we were able to generate an ID of 28 and 37 sequences respectively [Askelof et al., 1990; Novotny et al., 1991; Novotny et al., 1991; Dias et al., 1994; Mascarell et al., 2005]. For *Chlamydia pneumoniae (Chlamydophila pneumoniae),* the ID was made up of 68 protein sequences [Montigiani et al., 2002; Erkkila et al., 2004; Sambri et al., 2004; Finco et al., 2005]. Regarding *Mycoplasma pneumoniae*, the ID contained 198 protein sequences. These last four pathogens have not any sequence-identity region between the proteins from the IDs and human proteins, so the databases were not enlarged. Taking into account all of the pathogens analysed, we were able to study the sequence of a total of 2175 proteins.

RESULTS:

The captured B-cell epitopes were analyzed to see if these sequence-stretches share or not common regions with human proteins. Our pre-assumed hypothesis was that one of the most important tasks of the immune system is the differentiation between self and non-self proteins. Therefore, this system will avoid the elicitation of antibodies against pathogen proteins sharing epitopes with host proteins. The presence of epitope mimicry may cause autoimmune diseases as mentioned in the previous sections [Benoist and Mathis, 2001].

Nowadays it is thought that humoral response against foreign proteins can distinguish between dangerous proteins and nondangerous ones [Matzinger, 2002]. The paradigm analyzed here could be used to describe how the bacterial proteins sharing linear B-cell epitopes with human proteins avoid immuno-reactivity against the host in spite of the possibility to produce auto-antibodies. Many questions like the following ones were initially formulated: Are these produced antibodies non-protective ones? Or, are these proteins not producing antibody responses even though they were able to do it?

New autoimmune diseases might appear if the exposed proteins of the bacterial pathogens share linear B-cell epitopes with any human protein, so this study may be useful to identify them [Rappuoli, 2000; Rappuoli and Covacci, 2003]. The findings may also help us to identify proteins which should not be used as putative vaccine candidates in a *reverse vaccinology* study. It is well-known that the antibodies recognize a small part of a big molecule and, therefore, two different proteins could be identified by the same antibody if both shared a small epitope of five to six amino acids, more o less, which is the usually accepted minimal length for a linear B-cell epitope [Lucchese et al., 2007].

B-cell epitope analysis

After analysing the three database files of linear B-cell epitopes obtained via BLASTP [Altschul et al., 1990], we found that none of the well-known protective antigens analyzed presented common linear B-cell epitopes with human proteins suggesting the existence of a system that tries to avoid autoimmunity. This mechanism might select those linear B-cell epitopes not having sequence similarity with human proteins. In summary, we could see that none of those 7353 linear B-cell epitopes shared any sequence identity region with human proteins capable of generating antibodies (See Figure I-4 of the Materials and Methods section). This was found apart of the already known epitopes that generate auto-antibodies and cause autoimmune diseases, the allergies caused by some epitopes, or certain sequence identities found between some artificial peptides and human proteins after their administration.

Exposed protein anaysis

As detailed in the Materials and Methods section of this chapter, for each of the pathogens analyzed a LF was made from the output file of the BLASTP analysis [Altschul et al., 1990]. This LF file contains the alignments in which local sequence similarity between the pathogen protein and a human protein has been found after an exhaustive inspection. Despite the huge amount of exposed proteins and pathogens analyzed, more than 2000, we found around only 20 protein alignments having he previously mentioned characteristics. This only represents 1% of the total proteins analyzed (See Figure I-5 of the Materials and Methods section).

According to the hypothesis of the existence of a mechanism to avoid autoimmunity already mentioned, the finding of pathogen exposed proteins sharing sequence stretches with human proteins that are

considered to be linear B-cell epitopes is very difficult. Not only epitope mimicry between proteins may lead to autoimmune phenomena most of these related to several diseases, but also this effect might elicit antibodies with high difficulty demonstrating the immunotolerance effect [Chatenoud, 2000]. The way of trying to avoid this mimicry and the presence, in some cases, of this immunotolerance effect reinforces the difficulty in finding regions of sequence identity between pathogen proteins and host proteins.

Here it is shown, an example of a specific procedure for the development of new vaccines when corroborating or advising about the viability of a surfome protein of a pathogen [Rodríguez-Ortega et al., 2006] as a putative vaccine candidate in a *reverse vaccinology* study [Rappuoli, 2000; Rappuoli and Covacci, 2003]. In this particular case, this was applied to the main human bacterial respiratory pathogens as an example of a tool that might be used in *reverse vaccinology*. As stated before, only 20 proteins of a total of more than 2000 analyzed shared a significant sequence-identity region with human proteins, so we strongly advise against recommending these proteins as putative vaccine candidates.

DISCUSSION:

In order to exemplify the sequence similarity analysis that is proposed here to be done as a tool in a *reverse vaccinology* study, three examples of alignments between the exposed-pathogen proteins and human proteins are detailed and discussed below:

***Streptococcus pneumoniae* and *Chlamydia pneumoniae*:**

After the BLASTP output file analysis of these two pathogens, we could see that there exist several bacterial proteins sharing a significant sequence-identity region with human proteins. Here, a detailed study was made and a very interesting case is shown in Figure 1. In this two alignments it can be seen that a couple of proteins of these pathogens has a stretch of shared sequence with a human protein that is considered to have an uveal auto-antigen [Ring, 1927]. Furthermore, the region of each of the two proteins is not included in the predicted transmembrane sections and there are some references highlighting that these proteins may elicit antibody immune response [Wiezemann et al., 2001; Adamou et al., 2001; Montigiani et al., 2002; Chiavolini et al., 2003; Erkkila et al., 2004; Sambri et al., 2004; Finco et al., 2005; Maione et al., 2005; Beghetto et al., 2006]. These two proteins are the Zinc metalloprotease zmpB precursor, in the case of *Streptococcus pneumoniae*, and the Cpn0042 protein for *Chlamydia pneumoniae*. Moreover, the stretch of shared sequence for the *Streptococcus pneumoniae* protein does not correspond to a putative signal peptide because it is not close to the N-terminal region. The stretch of shared sequence for *Chlamydia pneumoniae* protein is not predicted to be a predicted signal peptide either. Although the epitope prediction servers could not corroborate that these sequence sections correspond to putative linear B-cell epitopes, we considered these cases

as important ones.

Although these two proteins seem to be good putative vaccine candidates, our preliminary analysis of *reverse vaccinology* enables us not to recommend the use of these proteins for the development of new vaccines. This is due to the fact that they could probably generate antibodies against the human protein with which they share a stretch of sequence. In this particular case, the presence of an auto-antigen in the human protein has to be also considered and, therefore, a previous infection with *Sterptococcus pneumoniae* or *Chlamydia pneumoniae* might promote an auto-immune reaction at the uveal tract apart from producing infection by their own.

I-6 a)

***Streptococcus pneumoniae*:**

```
*** BLASTP SUBJECT:

>gb|AAG49577.1| uveal autoantigen [Homo sapiens]
          Length = 1416

 Score = 46.6 bits (109), Expect = 0.008
 Identities = 78/373 (20%), Positives = 148/373 (39%), Gaps = 37/373 (9%)

Query: 461  KQPEVNSETNKLKTAIDALNVDKTELNNTIADAKTKVKEHYSDRSW--QNLQTEVTKAEK 518
            K  ++N E  K+K   + L  +     N I       + EH +  S   Q+++        +
Sbjct: 889  KFEDINQEFVKIKDKNEILKRNLENTQNQIKAEYISLAEHEAKMSSLSQSMRKVQDSNAE 948

Query: 519  VAANTDAKQSEVN------EAVEKLTATIEKLVELSEKPILTLTSTDKKILEREAVAKYT 572
            + AN    Q E+      +A +K   TI++ +++   PI++    ++K    E   K
Sbjct: 949  ILANYRKGQEEIVTLHAEIKAQKKELDTIQECIKVKYAPIVSFEECERKFKATEKELKDQ 1008

Query: 573  LENQNKTKIKSITAELKKGEEVINTVVLTDDKVTTETISAAFKNLEYYKEYTLSTTMIYD 632
            L  Q + K       E+KK ++          +DK+   E  +   K+L   K    +  +
Sbjct: 1009 LSEQTQ-KYSVSEEEVKKNKQ--------ENDKLKKEIFTLQ-KDLRD-KTVLIEKSNEME 1058

Query: 633  RGNGEETETLENQNIQLDLKKVELKNIKRTDLIKYENGKETNESLITTIPDDKSNYYLKI 692
            R    +T+ L  Q   L  K  E+KN+K  + +  EN K+T+E L       K +  L+
Sbjct: 1059 RALSRKTDELNKQLKDLSQKYTEVKNVK--EKLVEENAKQTSEILAVQNLLQKQHVPLEQ 1116

Query: 693  TSNNQKTTLLAVKNIEETTVNGTPVYKVTAIADNLVSRTADNK----------------F 736
                +K+    ++N++E   +    Y+         + +  +N+                F
Sbjct: 1117 VEALKKSLNGTIENLKEELKSMQRCYEKEQQTVTKLHQLLENQKNSSVPLAEKLQIKEAF 1176

Query: 737  EEEYVHYIEKPKVHEDNVYYNFKELVEAIQNDPSKEYRLGQSMSARNVVPNGKSYITKEF 796
            E+E V  I+      ++    N  E V  +Q++      +  + +  R VV   K   TK
Sbjct: 1177 EKE-VGIIKASLREKEEESQNKMEEVSKLQSEVQNTKQALKKLETREVVDLSKYKATKSD 1235

Query: 797  TGKLLSSEGKQFA 809
                +SS  ++ A
Sbjct: 1236 LETQISSLNEKLA 1248

*** BLASTP QUERY:

>UniProt/Swiss-Prot|Q9L7Q2|ZMPB_STRPN Zinc metalloprotease zmpB precursor
```

I-6 b)

***Chlamydia pneumoniae*:**

```
*** BLASTP SUBJECT:

>gb|AAG49577.1| uveal autoantigen [Homo sapiens]
          Length = 1416

 Score = 40.4 bits (93), Expect = 0.068
 Identities = 44/182 (24%), Positives = 81/182 (44%), Gaps = 23/182 (12%)

Query: 1    MEEVSEYLQQVENQLESCSKRLTKMETFALGVRLEAKEEIESII------LSDVVNRFEV 54
            MEEVS+    +V+N  ++  K+L   E   L      K ++E+ I      L+++  ++E
Sbjct: 1198 MEEVSKLQSEVQNTKQAL-KKLETREVVDLSKYKATKSDLETQISSLNEKLANLNRKYEE 1256

Query: 55   LCRDI----EDMLSRVEEIERMLRMAELPLLPIKEALTKAFVQ----HNSCKEKLTKVEP 106
            +C ++     +  +S  +E E +     E  +   KE   K+            +E   ++E
Sbjct: 1257 VCEEVLHAKKKEISAKDEKELLHFSIEQEIKDQKERCDKSLTTITELQRRIQESAKQIEA 1316

Query: 107  YFKESPAYLTSEERL-QSLNQTLQRAY-------KESQKVSGLESEVRACREQLKDQVRQ 158
              +    L   ERL Q+LN   Q  Y        ++SQ +  L+ +V++  +QL D  RQ
Sbjct: 1317 KDNKITELLNDVERLKQALNGLSQLTYTSGNPTKRQSQLIDTLQHQVKSLEQQLADADRQ 1376

Query: 159  FE 160
             +
Sbjct: 1377 HQ 1378

*** BLASTP QUERY:

>AAD18195.1 cpne1 CPn0042
```

Figure I-6. Example of mimetope identification - 1. (a) Partial sequence alignment coming from the output file of the BLASTP algorithm analysis between the exposed-protein database from *Streptococcus pneumoniae* and the non-redundant protein database. A putative linear B-cell epitope is highlighted. **(b)** Partial sequence alignment coming from the output file of the BLASTP algorithm analysis between the exposed-protein database from *Chlamydia pneumoniae* and the non-redundant protein database. A putative linear B-cell epitope is highlighted (Amela et al, 2007).

***Pseudomonas aeruginosa* and *Chlamydia pneumoniae*:**

In terms of these pathogenic microorganisms, the analysis of their BLASTP output files allowed us to identify two proteins containing a significant sequence-identity region with a human protein differently annotated from the pathogen protein (See Figure 2). Additionally, these sections do not correspond to transmembrane regions and there are also several references highlighting the antibody elicitation and,

therefore, the consequential immune response [Duchene, 1989; Umelo-Njaka, 2001; Price et al., 2001; Montigiani et al., 2002; Thomas et al., 2003; Erkkila et al., 2004; Sambri et al., 2004; Finco et al., 2005; Worgall, 2005]. Furthermore, the epitope-prediction servers mentioned in the Materials and Methods section corroborated that these two identity regions correspond to putative linear B-cell epitopes. These proteins are the Translocator outer membrane protein PopD, a constituent of the *Pseudomonas aeruginosa* Type III Apparatus, and the Inclusion membrane protein A in the case of *Chlamydia pneumoniae*. Obviously, the stretches of shared sequence do not correspond to the signal peptide because they are not close to the N-terminal region of the protein.

I-7 a)

***Pseudomonas aeruginosa*:**

```
*** BLASTP SUBJECT:

>ref|NP_078789.1| FYVE and coiled-coil domain containing 1 [Homo sapiens]
 emb|CAC33883.1| FYVE and coiled-coil domain containing 1 [Homo sapiens]
          Length = 1478

 Score = 36.6 bits (83), Expect = 1.2
 Identities = 35/122 (28%), Positives = 57/122 (46%), Gaps = 19/122 (15%)

Query: 152 EKTLQKNIDGRNELIDAKMQAL----GKTSDEDRKIVGKVWAADQVQDSVALRAAGRAFE 207
           EK LQ N+ GRN+L++ K+QAL          +  I G + + +  Q S+  R  G   E
Sbjct: 622 EKELQ-NVVGRNQLLEGKLQALQADYQALQQRESAIQGSLASLEAEQASI--RHLGDQME 678

Query: 208 SRNGALQVANTVIQSFVQMANASVQVRQGESQ------------ASAREGEVNATIGQSQ 255
           +   A++ A   +++ +    A +Q ++GE Q            A AR  E+ A   Q Q
Sbjct: 679 ASLLAVRKAKEAMKAQMAEKEAILQSKEGECQQLREEVEQCQQLAEARHRELRALESQCQ 738

Query: 256 KQ 257
           +Q
Sbjct: 739 QQ 740


*** BASTP QUERY:

>UniProt/TrEMBL|Q9I323|Q9I323_PSEAE Translocator outer membrane protein PopD
```

I-7 b)

***Chlamydia pneumoniae*:**

```
*** BLASTP SUBJECT:

>ref|NP_003557.2| early endosome (membrane-bound compartment inside cells)
antigen 1, 162kD [Homo sapiens]
 sp|Q15075|EEA1_HUMAN Early endosome antigen 1 (Endosome-associated protein
p162) (Zinc
           finger FYVE ( = P.aeruginosa) domain-containing protein 2)
 emb|CAA55632.1| endosomal protein [Homo sapiens]
         Length = 1411

 Score = 44.3 bits (103), Expect = 0.009
 Identities = 51/217 (23%), Positives = 97/217 (44%), Gaps = 23/217 (10%)

Query: 128 LKAAKDQLTLEIEAFRNENGNLKTTAED---LEEQVSKLSEQLEALERINQLIQANAGDA 184
           L+  ++ L  +I+A   E   L    E    L+EQV++L+E+L+     ++  Q N  D
Sbjct: 542 LEKEREDLYAKIQAGEGETAVLNQLQEKNHTLQEQVTQLTEKLKNQSESHKQAQENLHD- 600

Query: 185 QEISSELKKLISGWDSKVVEQINTSIQALKVLLGQEWVQEAQTHVKAMQEQIQALQAEIL 244
            ++  +   L +  D   V  + TS+  L   L      E++  V  +  QI+A    +L
Sbjct: 601 -QVQEQKAHLRAAQDR--VLSLETSVNELNSQLN-----ESKEKVSQLDIQIKAKTELLL 652

Query: 245 GMHNQSTALQKSVENLL------VQD--QALTRVVGELLESENKLS---QACSALRQEIE 293
                TA +  ++N L      +QD  Q L ++  +L +   KL    + CS L   ++
Sbjct: 653 SAEAAKTAQRADLQNHLDTAQNALQDKQQELNKITTQLDQVTAKLQDKQEHCSQLESHLK 712

Query: 294 KLAQHETSLQQRIDAMLAQEQNLAEQVTALEKMKQEA 330
           +  +   SL+Q+ + +  Q + L      ++  K++A
Sbjct: 713 EYKEKYLSLEQKTEELEGQIKKLEADSLEVKASKEQA 749


*** BASTP QUERY:

>Q9Z8Z8|Q9Z8Z8_CHLPN Similarity to CT119 Inclusion membrane protein A (IncA)
```

Figure I-7. Example of mimetope identification - 2. (a) Partial sequence alignment coming from the output file of the BLASTP algorithm analysis between the exposed-protein database from *Pseudomonas aeruginosa* and the non-redundant protein database. A putative linear B-cell epitope is highlighted. **(b)** Partial sequence alignment coming from the output file of the BLASTP algorithm analysis between the exposed-protein database from *Chlamydia pneumoniae* and the non-redundant protein database. A putative linear B-cell epitope is highlighted (Amela et al, 2007).

These examples are emphasized because they clearly show why a protein should not be used as a vaccine candidate when working with *reverse vaccinology* techniques. As can be seen in the alignments above, the two pathogen proteins share a significant sequence-identity region with human proteins and that is why they are candidates for generating an antibody response against the human protein with which they share this stretch of sequence. Moreover, these human proteins are part of the FYVE domain containing proteins and are required for the formation of early-endosomal membranes that are the main resource used for phagocyte action [Bannantine et al., 2000; Vieira et al., 2001; Birkeland and Stenmark, 2004; Nguyen and Pieters, 2005; Lindmo and Stenmark, 2006]. For this reason, the probable antibody elicitation against these proteins might generate a serious auto-immune problem for the endosome-mediated action in the human defense system.

***Streptococcus pyogenes*:**

Regarding *Streptococcus pyogenes* BLASTP output file analysis, there was only one pathogen protein sharing a sequence identity-region with a human protein differently annotated from the pathogen protein (See Figure 3). This pathogen protein had all the desirable properties to be a vaccine candidate and, in fact, it was demonstrated to elicit antibody immune response [Banks et al., 2004; Seepersaud et al., 2005]. Moreover, the epitope-prediction servers corroborated that the sequence-identity region corresponds to a putative linear B-cell epitope and the transmembrane-prediction server showed that this sequence identity region does not coincide with the predicted transmembrane ones. As before, we checked that the stretches of shared sequence do not correspond to the signal peptide region of the protein.

I-8)

***Streptococcus pyogenes*:**

```
*** BLASTP SUBJECT:

>gb|AAH51330.1| DOCK6 protein [Homo sapiens]
          Length = 509

 Score = 34.7 bits (78), Expect = 4.4
 Identities = 19/58 (32%), Positives = 29/58 (50%)

Query: 181 AQSASEGPWLLAEGLPTVEDHRHLPIGLQVELMKAIGTIDNILISNQFISEEELAACT 238
           AQ        L+AE L  +EDHRHLP+G          +   ++   IS+  +S +E   C+
Sbjct: 79  AQCMVHAAALVAEYLALLEDHRHLPVGCVSFQNISSNVLEESAISDDILSPDEEGFCS 136

*** BLASTP QUERY:

>outer surface protein
```

Figure I-8. Example of mimetope identification - 3. Partial sequence alignment coming from the output file of the BLASTP algorithm analysis between the exposed-protein database from *Streptococcus pyogenes* and the non-redundant protein database. A putative linear B-cell epitope is highlighted (Amela et al, 2007).

As can be seen here, this pathogen protein has a stretch of shared sequence with a human protein called DOCK6. Consequently, the antibodies that might be generated against this exposed-pathogen protein could attack DOCK6, leading to auto-immune effects. DOCK6 promotes neurite outgrowth [Miyamoto et al., 2007] being a clue protein in neural development and it is easy to suppose wrong effects in human health if a lack of DOCK6 is present. That is why our *reverse vaccinology* approaches may recommend not using this outer-surface protein as a putative vaccine candidate.

It has to be taken into account that only the protein sequences sharing certain sequence identity regions were considered if these two proteins are different and do not have anything in common in their annotation. This decision was assumed due to the fact that there are many

proteins of completely different organisms grouped in the same family or being considered as protein-like proteins and obviously these proteins are condemned to share common regions, so they were not included in the study.

In summary, a protein sequence analysis that may be applied before a *reverse vaccinology* study has been proposed and several examples of the procedure have been shown. This proposal may be used in *reverse vaccinology* either to corroborate to or warn about the viability of a linear B-cell epitope as a putative vaccine candidate. Therefore, epitopes without any sequence identity with human proteins should be very good vaccine candidates, and the other way round.

CONCLUSIONS:

- None of the 7353 linear B-cell epitopes analyzed, which should be capable of generating antibodies, share any sequence identity region with human proteins. Moreover, only 1% of the 2175 exposed proteins analyzed contain a stretch of shared sequence with the human proteome. These facts suggest the existence of a mechanism to avoid autoimmunity.

- A strategy for corroborating or warning about the viability of a pathogen protein for vaccine design approaches has been proposed. Therefore, epitopes without any sequence identity with human proteins might be used as vaccine candidates, and the other way round.

CHAPTER II:

Analysis of Protein Interactions

INTRODUCTION:

As initially mentioned, in Chapter II of this thesis the focus is put on the elaboration of a bioinformatic program to treat the big quantity of local data that could be produced by PPD tools. The bioinformatic approaches used in this section comprise the use of several structural bioinformatic techniques over proteins, the management of PPI databases, the retrieval of data and literature mining of PPIs, the utilization of different PPD tools and the application of the R programming language and some clustering methods. In these two last approaches we tightly collaborated with a mathematician of the Universitat Politècnica de Catalunya (UPC) who basically programmed the algorithm. Among many other putative uses, this program will be employed in Chapter III.

Protein interactions

Proteins are not isolated, but they interact with each other to accomplish their functions. Therefore, PPI is the key process by which most of the proteins fulfill their roles and interactomics represents one of the current frontiers of biosciences [Gavin and Superti-Furga, 2003; Pache et al., 2008]. PPIs can be studied at different levels:

- Functional interactions → despite taking part in the same biological process, proteins that might never physically interact.
- Physical interactions → proteins that contact each other and participate in the same biological route to do a specific function.
    - Permanent interactions → proteins that are parts, called monomers, of a quaternary structure, a multimer. Monomers alone do not have a function but interacting they do have it [Jackson et al., 1993; Cohen et al., 2005].

- Transient interactions → some of the physical interactions are not permanent but they are temporary. This fact occurs in most of the biochemical reactions of the cell.

The experimental techniques currently used to study the interactome can be divided into two principal branches depending on the type of research:

- High-Throughput experimental methods → the most used are Yeast Two-Hybrid (Y2H) and Tandem Affinity Purification (TAP) [Fields and Song, 1989; Rigaut, 1999; Puig et al., 2001].
- Detailed interaction techniques → there three experimental techniques commonly used to study the protein interactions in detail are X-Ray Crystallography (XRC), Nuclear Magnetic Resonance (NMR) and High-Resolution Transmission Electron Microscopy (HRTEM).

These techniques produce a lot of PPI data and, consequently, many different PPI databases are currently available, like for example: IntAct, DIP, BioGRID, MIPS, HPRD, MINT or BOND (See Figure II-1). It is a difficult task to combine efficiently this information and some efforts have been put in this direction. Several tools that integrate most of these PPI databases and facilitate the search of PPI information are available [Prieto, 2006].

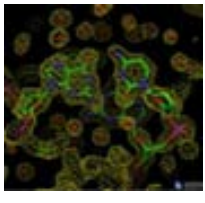
Figure II-1. Protein inteeaction network. A map of protein-protein interactions in *Saccharomyces cerevisiae* (Macmillan Magazines Ltd.).

PPIs can help us to predict protein function and, therefore, many protein function predictors have been developed using PPI databases [Vázquez et al., 2003; Chua et al., 2006; Sun et al., 2006; Chen et al., 2008; , Espadaler et al., 2008; Gabow et al., 2008; Jaeger et al., 2008]. Due to PPIs, it is expected that in the near future the number of solved protein complexes will surpass the number of proteins in some organisms. A lot of PPIs involve surface displacements among the members of the protein complex to achieve the required biological function.

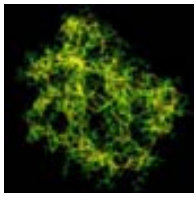
Three-dimensional structure determination of proteins

There are mainly three experimental techniques used to study protein structure: XRC, NMR and EM. Most of the currently available protein structures are solved by XRC in which a protein crystal is bombarded with a beam of X-rays to obtain a diffraction pattern exclusive for that protein. This pattern depends on the particular electron density of the

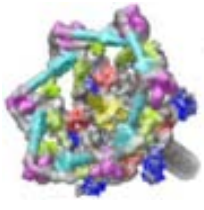
protein crystal and can be interpreted to generate a three-dimensional map of the atoms of the protein. Some protein structures, less than 10 %, are determined by NMR, which makes use of atomic nuclei and their resonance emission under magnetic conditions. Small proteins with this kind of atoms can be studied from a structural and dynamical point of view with this technique. EM is a method that obtains images of the shape of the protein at low resolution directly from the sample. It is used to define the structure of big protein complexes because it does not reach atomic levels. Usually, these three techniques are combined to obtain a proper protein structure. See the next figure:



a) XRC



b) NMR



c) EM

Figure II-2. Methods for Determining Protein Structures. The images show: (a) an electron density map obtained by XRC, (b) some of the restraints used to solve a structure by NMR, and (c) a surface rendering of Electron Microscopy (EM) data (Protein Data Bank, http://www.rcsb.org).

From those technologies, XRC and NMR have been the two mainly applied for relatively high resolution structure elucidation until the moment. Even so, these hi-tech methods are frequently constrained by the methodological requirements when dealing with protein complexes. 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 as the Protein Data Bank (PDB; http://www.rscb.org/pdb/; [Berman et al., 2000]).

Therefore, when trying to analyze the dynamics of the interaction process among the proteins of a protein complex, a NMR spectroscopic technique may not be feasible, and the data obtained of a XRC experiment may not be useful to represent the dynamic behavior. Consequently, despite the use of these two experimental technologies for protein structure determination being widely distributed, other complementary strategies may be useful to accurately model the dynamics of the interaction among the proteins of a protein complex.

Protein docking

In this context, some theoretical methods to study protein complexes at a structural level, such as docking, are now emerging. PPD is a computational method to predict the best way by which proteins interact [Ritchie, 2008; Vakser and Kundrotas, 2008] (See Figure II-3). It is important to say that PPD does not prognosticate which proteins could interact, but the method predicts how the proteins already known to interact do it.

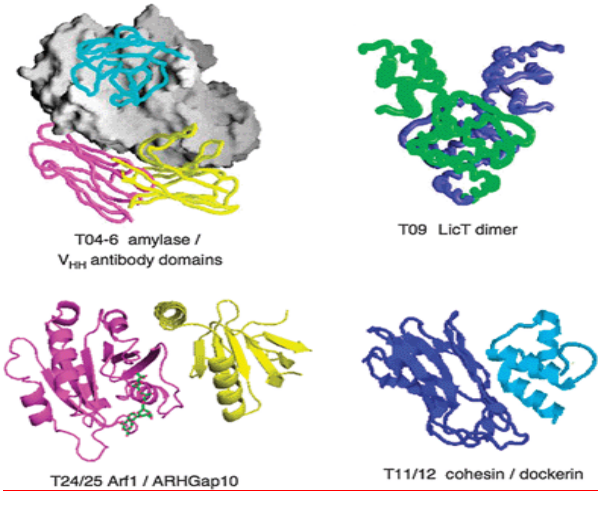


Figure II-3. Protein complexes postulated by PPD. Examples of the Critical Assessment of PRedicted Interactions (CAPRI) competition targets (Janin 2010).

There are two basic types of PPD methods:

- Rigid-body docking → the protein is considered as a fixed entity in which each residue is represented as a sphere allowing a degree of penetration between them. First, a surface docking is done producing all possible rotations between the proteins, and, after that, translations are done to get the proteins into contact. Conformational changes during the complex formation are not permitted, in order to save computation time and power. This technique may be appropriate when non-substantial conformational changes are expected to take place in the interacting proteins.
- Flexible docking → angles, lengths and torsions of bonds between atoms of the proteins are taken into account in the docking procedure. High levels of computation time and power are required because, apart from a rigid surface docking, a relocation of side-chains and some main-chain modifications are done, which is called structure refinement.

Some algorithms for PPD docking have been developed over the last few years in order to combine the advantages of rigid-body docking and flexible docking. These methods are called Soft-docking algorithms and employ in different ways reciprocal space rules, geometric hashing, shape complementarity, electrostatics and physical chemistry restraints between the proteins to be as much effective as possible. Moreover, these procedures tend to make use of Fast Fourier Transform (FFT) correlation algorithms, which define a cubic grid as a simplified model of the proteins, to speed up and enhance the process.

This method shows high speed in rigid–body docking and no restraints can be introduced initially.

Instead of taking into consideration all of the possible configurations the methods based on the Monte-Carlo simulated annealing algorithm explore only the best configurations by generating random movements and selecting them according to interaction energies. The genetic algorithms do the same but selecting the potentially useful configurations based on the best scoring positions being generated after a random position search. In these methods an initial docking conformation is taken from a scoring function after different steps. This should converge to the best possible structure but these methods do not cover all the docking solutions and depend a lot on the scoring functions.

Although these approaches allow for flexibility, more precise flexible methods are being developed combining Monte-Carlo algorithms and ensemble dockings of NMR structures or motion predictions. With these new programs not only the side chains movements but also those of the main chain may be studied. The success of flexible docking methods will depend on better scoring functions as much as faster algorithms..

In summary, an accurate, affordable and relatively fast PPD procedure is required and many docking protocols are being developed in order to achieve these premises. These new docking algorithms tend to simplify the docking process by dividing it into different steps where a different approach is made in each of the phases mixing rigid-body PPD, flexible PPD and docking solution evaluation and selection. It is like the “Divide and conquer” question! (See Figure II-4).

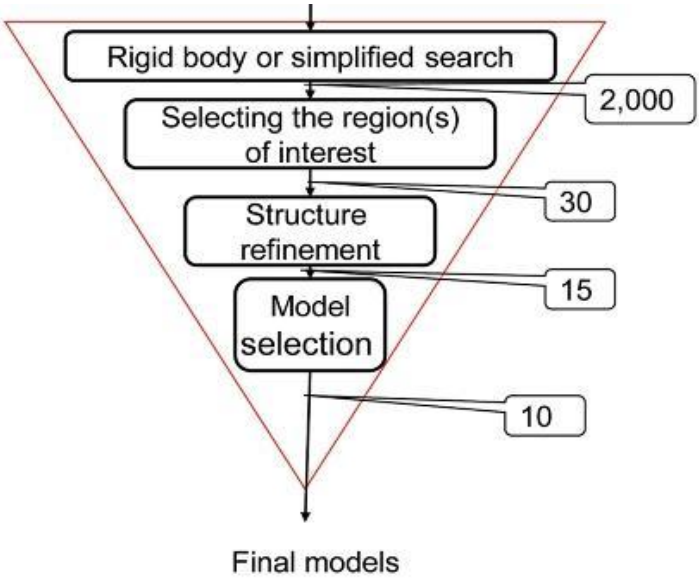


Figure II-4. A typical flowchart of a PPD procedure. Computational steps in multistage docking (Vajda 2009).

Some of the most commonly used PPD tools are:

- 3D-Dock Suite (BioMolecular Modeling, Cancer Research UK, http://www.sbg.bio.ic.ac.uk/docking).

Integrated approach to protein docking with FTDock, RPScore and MultiDock.

- 3D-Garden (Imperial College London, http://www.sbg.bio.ic.ac.uk/~3dgarden).

System for modeling protein complexes based on conformational refinement of ensembles generated with the marching cubes algorithm.

- Bielefeld Protein Docking (Bielefeld University, http://www.techfak.uni-bielefeld.de/~posch/DOCKING/install.html).

It detects geometrical and chemical complementarities between surfaces of proteins and estimates docking positions.

- BiGGER (BioTecnol, S.A., http://ww.cqfb.fct.unl.pt/bioin/chemera).

Protein docking algorithm integrated in Chemera, a molecular graphics and modeling program for studying protein structures and interactions.

- ClusPro (Boston University, http://cluspro.bu.edu/).

Integrated approach to protein docking with DOT and ZDOCK and PIPER.

- DOT (San Diego Supercomputer Center, http://www.sdsc.edu/CCMS/DOT/).

It computes the electrostatic potential energy between two given proteins or other charged molecules.

- ZDOCK (University of Massachusetts Medical School, http://zdock.umassmed.edu/software).

Performs a full rigid-body search of docking orientations between two proteins including performance optimization and a novel pairwise statistical energy potential.

- PIPER (Boston University, http://structure.bu.edu/content/protein-protein-docking)

FFT-based docking with pairwise potentials.

- Escher NG (Milan University, http://www.ddl.unimi.it/escherng/index.htm).

Enhanced version of the original ESCHER protein-protein automatic docking system developed in 1997.

- HADDOCK (Utrecht University Netherlands, http://www.nmr.chem.uu.nl/haddock).

High Ambiguity Driven biomolecular DOCKing that employs biochemical and/or biophysical interaction data.

- Hex (University of Aberdeen, http://hex.loria.fr).

Protein docking and molecular superposition program.

- RosettaDock (Johns Hopkins University,http://rosettadock.graylab.jhu.edu).

Predicts the structure of protein complexes given the structures of the individual components and an approximate binding orientation.

- DOCK (UCSF Molecular Design Institute, http://dock.compbio.ucsf.edu).

It uses a geometric matching algorithm to superimpose the swtructures.

- GRAMM (University of Kansas, http://vakser.bioinformatics.ku.edu/resources/gramm).

Global RAnge Molecular Matching.

- ICM-DISCO (MolSoft LLC, http://www.molsoft.com/icm_pro.html).

It is a direct stochastic global energy optimization from multiple starting positions.

- PatchDock (Tel Aviv University, http://bioinfo3d.cs.tau.ac.il/PatchDock/).

It is based on geometric hashing and shape complementarity principles

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 energies to finally form the protein complex. These displacements occur along the protein surface, generating multiple low-energy interaction complexes. In these cases, these low-energy interaction 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 PPIs coming from different orientations like, for instance, proteins that require movements among them when they act as a protein complex. Owing to all these facts, interaction among proteins seems to be a dynamic mechanism where there is not only one single solution with the best interaction energy, like most of the current PPD programs consider, but rather there are several solutions with more or less interaction energy, and not necessarily does the native form have the best theoretical solution [Halperin et al., 2002].

OBJECTIVES:

- Develop a new program to analyze and simplify the output-data of a docking essay.

- Make the interpretation of the docking solutions easier for the user guiding him to find out the best representative structures which do not always match those with the minimal energy.

MATERIALS AND METHODS:

Basis of *DockAnalyse*

The clustering algorithm used in the design of *DockAnalyse* was DBscan (density-based spatial clustering of applications with noise) [Ester et al., 1996]. It relies on a density based notion of clusters and finds a number of clusters starting from the estimated density distribution of corresponding nodes. Furthermore, it is designed to discover the clusters of arbitrary shape as well as to properly distinguish noise. DBscan is one of the most common clustering algorithms and also most cited in scientific literature, but, curiously, it has not been previously used for the analysis of protein docking results.

The algorithm is based on the definition of density connection, where two points in a dataset are density connected if a chain of points in the dataset that allows for the movement from one to the other if it exists. The connecting chain must verify two conditions: firstly, each point of the chain (except, probably, the first and the last one) has at least k observed data at a distance less than a determined radius (ε), and, secondly, the distance between two consecutive points in the chain is less than ε. This definition induces a partition in the set of observed points and, therefore, each group is defined as a subset of points which are density connected among each other. The definitive clusters provided by DBscan are those components in the partition with, at least, two or more elements. DBscan considers those non-density connected points as isolated ones (groups with only one member or outliers).

For example given the nodes shown in Figure II-5, it is clear that the blue point is a noise point (N), yellow points (B and C) are density-reachable or density-connected points and red points (A) are core points. Points A, B and C belong to the same cluster whereas

point N is a noise point because it is neither a core point nor density-reachable.

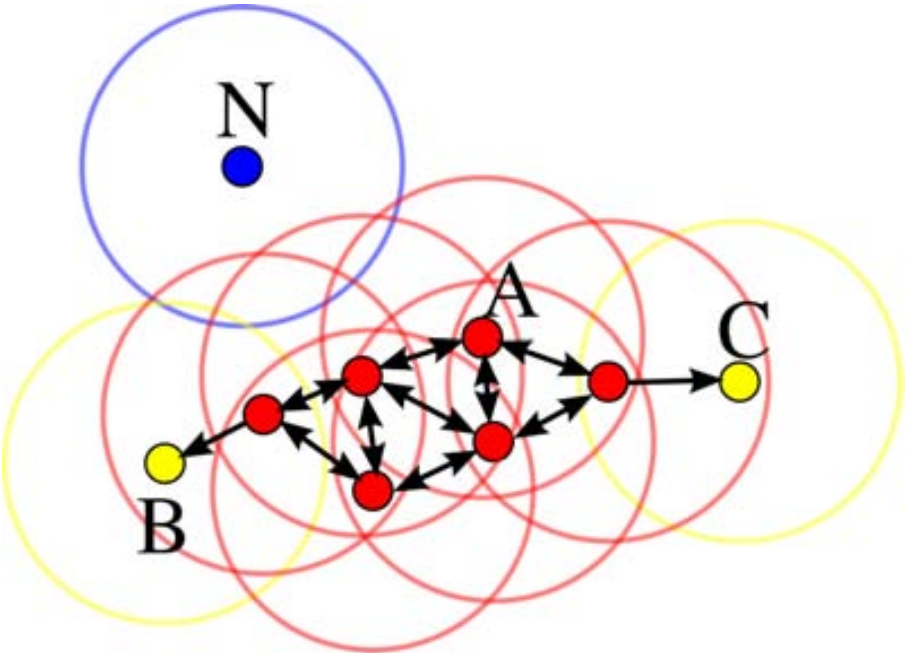


Figure II-5. Density connection. DBscan's definition of a cluster is based on the notion of density reachability (http://en.wikipedia.org/wiki/DBSCAN).

Why DBscan?

DBscan has been designed to discover clusters of arbitrary shape due to the fact that it is a density based algorithm. It identifies dense regions (clusters) which are separated by regions of low density (considered as outliers). The lack of an appropriate outlier is a well-known weakness of one of the classical clustering methods like k-means, where even very far points from the closest centroid are included in the same cluster without any additional criteria. This clustering method was chosen because it is extremely robust and it also solves some inconsistency

problems that usually appear when applying other clustering methods. In general, the classical clustering methods do not manage the outliers well, while DBscan tends to treat these isolated points much better and it allows for the finding of all cluster members independently of the cluster shape, discarding the outliers. One of the main problems of clustering is that the classical methods are dependent on the previously defined number of clusters, while DBscan is not. As can be seen in Figure II-6, DBSCAN can find non-linearly separable clusters but the same dataset cannot be adequately clustered with k-means or other broadly used clustering methods. This figure demonstrate that this algorithm is very resistant to noise and can handle clusters of various shapes and sizes.

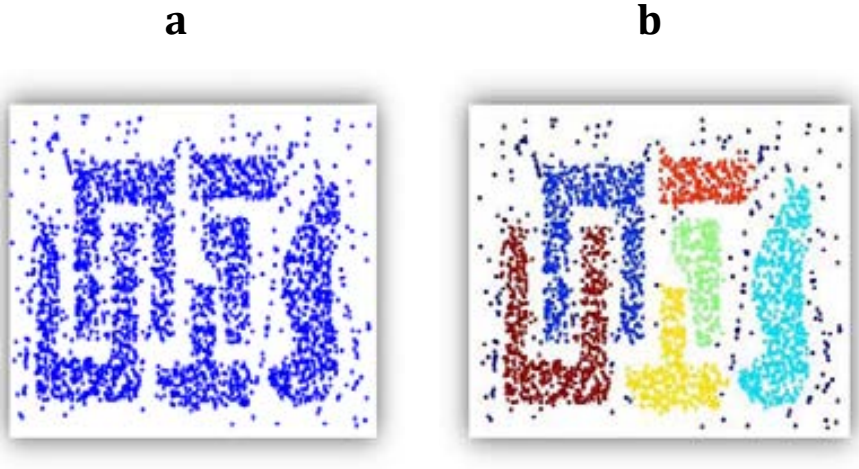


Figure II-6. DBscan clusters. A lot of noise is present in this example. In (a) the original data is shown and in (b) the clusters found with DBscan are depicted. The noise around the outer parameters are handled appropriately. A lot of clusters that DBSCAN can find K-mean would not be able to find.

DBscan parameters

The DBscan algorithm depends on two tuning parameters that define the density connection (k and ε). Ester et al. 1996 indicate that the choice of a good ε is much more important than the choice of k. The results in their datasets are quite similar for any k≥4 and, therefore, they proposed to fix k=4. In our experiments, better results for k values greater than 4 were verified and, consequently, k=15 was used in all computations. The classical purpose of executing the DBScan algorithm is usually to clusterize a cloud of points in order to group all of the points together according to their similarities, but in *DockAnalyse* we also used it to sort these clusters taking into account the number of members of each cluster. The more members a cluster has the better the solution is. This property allows for the sorting of the solution according to the number of members/solutions per cluster, and at the same time it allows for the removing of non-relevant docking solutions, which is the objective of fixing our k=15 (minimum number of points in a cluster). The best solutions, which are the only ones that will be checked by the user of *DockAnalyse*, are integrated by hundreds of points. Those clusters with less than 15 members are discarded, therefore, they have no effect on the final result. This is done because nobody is going to check a docking representative with such a limited number of docking solutions supporting the goodness of this region to be a binding surface. Besides, the user can always access the script source code to modify the k parameter (minimum number of points in a cluster), but this is only an optional possibility because, as previously commented, this value has no relevant effect on the final results.

Protein Dockings

DockAnalyse is currently designed to be used with whatever version of Escher NG or Hex [Ausiello et al., 1997; Ritchie and Kemp, 2000; Ritchie et al., 2008] PPD programs, but by modifying only a few parameters of the script source code most of the PPD programs could also be employed. The only premise is that the PPD program must generate an output datafile composed of a matrix with information of rotations, translations and interaction energies for each solution as shown in Figure II-7. Despite the existence of the possibility of reducing the number of given solutions in the PPD experiment for most PPD programs, it has to be considered that the more solutions obtained in the docking assay, the more robust the *DockAnalyse* results would be.

```
# Format: Representative_Line Solution Etotal x y z alpha beta gamma RMSD
#
# ---------------------------------------------------------
1    1   -845.37 -0.111143  0.008272  0.162905 0.865416  0.006480  5.417475      0.147
2    2   -829.63 -0.385231  0.317502 -0.002757 6.282684  0.000000  0.000000      0.487
3    3   -784.25  0.007804 -0.213867  0.731412 0.864956  0.026800  5.417015      0.705
4    4   -769.31 -0.064938  0.925024 -1.396958 4.008249  0.048199  2.277122      1.741
5    5   -725.62 -0.001598 -0.534265 1.162267  0.864617  0.041750  5.416676      1.233
6    6   -660.43  0.112566 -3.252725 -1.328348 2.102316  0.064851  4.240163      1.437
7    10  -592.82  0.518444 -0.746692  0.911540 0.864858  0.031121  5.416917      1.402
8    11  -589.73 -0.671570  1.022269 -0.480875 4.258525  0.025334  2.025773      1.633
9    12  -560.23 -0.141013 -3.315965  1.523314 1.332539  0.090863  5.005719      2.083
10   16  -532.79  0.965819 -4.249353  1.089213 1.236738  0.025995  5.163925      2.582
11   22  -515.29  1.107065 -4.934368 -1.942827 2.044183  0.093625  4.304260      3.307
12   26  -502.55  0.462168 -1.026089  1.683549 0.864083  0.065310  5.416141      2.259
13   33  -462.58 -0.622339  1.391158 -1.454892 4.342973  0.094249  1.944247      3.299
14   52  -437.29  1.540769 -3.487746 -6.108625 3.354122  0.130747  3.007737      1.833
15   62  -426.93  2.358576 -4.055119 -5.820873 3.344851  0.121573  3.016361      1.877
16   69  -423.79  0.886040 -4.336377 -4.088686 2.509778  0.094161  3.842551      2.339
17   87  -410.38  2.067793 -5.015933 -5.050265 2.802227  0.106572  3.553095      2.751
18   88  -408.96  2.716425  3.597665 -6.083304 3.324415  0.139389  2.891223      2.115
19   122 -379.25 -1.278899  0.634788  7.919492 6.244314  0.157015  0.028376      1.797
20   143 -367.15 -0.845114  0.466409  7.573457 6.189298  0.150616  0.083225      1.759
21   160 -356.94  0.435395  1.303410  3.847189 5.281776  0.127513  0.988776      2.903
22   7   -617.96  2.916360  6.834843 -6.706204 3.065673  0.162859  3.027229      4.263
23   0   -613.57  2.256320  6.797478 -5.379806 2.900978  0.131374  3.188243      4.435
24   9   -598.62  2.946232  6.669707 -6.403817 3.055504  0.156579  3.037026      4.272
```

Figure II-7. Docking output file format. This is an example of the initial part of a results file of the protein docking program Hex 5.1 (data of the 25 initial solutions out of 1000).

RESULTS:

Overview

Our approach attempts to deal with the particularities of PPD methods that were mentioned in the Introduction section of Chapter II of this thesis. By considering the global contribution of the calculated docking solutions and selecting those representative solutions, which are the centers of the clusters having high interaction energy, we could describe a general behavior of a subset of solutions without improving unrealistic ones. To extract those solutions that best describe the real dynamic mechanism of interaction from the output data-file of the current PPD programs, we have developed an application, called *DockAnalyse*, which is based on the already existing DBscan clustering method [Ester et al., 1996] and, moreover, is unsupervised and automatic. The aim of this new method 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. The way of choosing the representative solutions is made by searching for continuities among these clusters.

The real challenge of the newly developed application is the ability to identify significant structures from the huge amount of previously calculated docking solutions without requiring too many tuning parameters from the user in order to run the program. Normally, the decision about which of the docked structures is the most important is very difficult, but *DockAnalyse* guides the search for good docking candidates by reducing the huge amount of putative docking solutions to check. Furthermore, the use of *DockAnalyse* allows a global vision of many characteristics of the PPI process through different data, graphical representations and possible personalized searches which also guides the search for significant solutions. The exhaustive analysis of all

of the PPI structures obtained with *DockAnalyse* may help us to theoretically postulate the structure of the studied protein complex or to propose the way by which certain proteins interact together in a mobile fashion to execute a biological function. Moreover, the analysis made by *DockAnalyse* might guide future flexible PPD approaches, because of the important PPD information obtained from the use of this new application.

Details of the program

Along the paragraphs of this subsection, the mathematical specifics used during the programming of *DockAnsalyse* are explained. As mentioned in the introduction section of this chapter, to end up with this new application we tightly collaborated with a mathematician of the Universitat Politècnica de Catalunya (UPC) who mathematically programmed and designed it.

Summarizing, with the aim of elucidating which of the docked structures between two studied protein structures are the most important from a functional point of view, an unsupervised procedure, based on the already existing DBscan clustering method [Ester et al., 1996], was designed and implemented with the R package. The movement, expressed in rotations and translations described by the proteins, and the interaction energy were considered in the algorithm to finally obtain the cluster distribution with the best internal coherence among the clusters generated by the docking output data-file representation. An initial transformation of the angles is required in order to make them comparable to location information.

First, distances among angles of the different docking solutions are computed. For instance, assume that $A = (a_1, a_2, a_3)$ and $B = (b_1, b_2, b_3)$ are the angles corresponding to two docking solutions, then the distance between A and B is defined as:

$$d(A, B) = \sqrt{d_1^2 + d_2^2 + d_3^2}$$

where

$d_i = \min \{|b_i - a_i|, |b_i - a_i + 2\pi|, |b_i - a_i - 2\pi|\}$

for $i = 1, 2, 3$.

That is, d_i is the angular distance between angles a_i and b_i. Once the distance matrix between angles of docking solutions has been computed, Multidimensional Scaling (MDS) [Cox and Cox, 2001] is performed on this distance matrix**.** The resulting principal coordinates are then concatenated to position variables in the docking output data-file in order to have a new data matrix where the Euclidean distances between rows represent the joint distance between angles and location of docking solutions. Moreover, the weight (or variability) and the information of each angle/position are forced to be the same in the new data matrix. Then, an automatic pre-processing step finds the radius necessary to run the DBscan clustering method, and the best ε (density reachability distance) parameter possible, also necessary for a proper DBscan analysis, is chosen according to a battery of cluster quality measures. To be specific, high values for high quality clustering of the following indexes have been considered (See Walesiak and Dudek 2007 for more details): Davies-Bouldin (multiplied by -1), Calinski-Harabasz, Hubert-Levine (multiplied by -1) and Silhouette. DBscan is applied to several ε-candidate values and the resulting clusters are evaluated by these criteria. The ε-candidate values are ranked according to every

index, and the score of a ε-candidate value is then established as the mean of its ranks. Finally, the value with the highest score is taken as the final ε and the corresponding cluster is considered to be the right one.

DockAnalyse, was applied to interpret the results obtained from different PPD assays because it gives a lot of information that is produced by the PPD output data-file analysis. As well as this information, the shape, size and distribution of the clusters obtained along with the position of the outliers are shown in *DockAnalyse* result graphical representations (See Figure II-8 for an example).

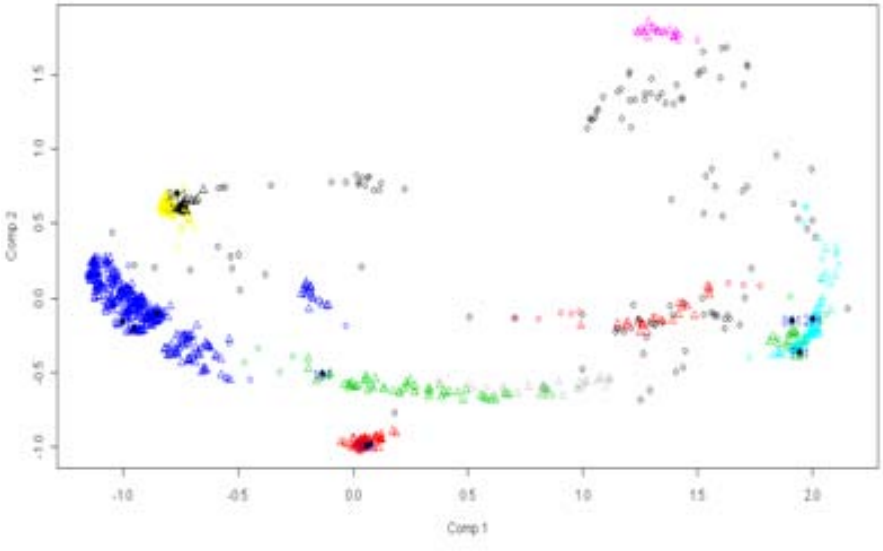


Figure II-8. Example of one of the *DockAnalyse* graphical output windows of a certain docking assay. This is one of the most important output windows of *DockAnalyse*, which shows the clustering graph of all of the docking solutions tested. The axes are the two extracted components of the computed Principal Components Analysis (PCA). The clusters found by the program are depicted in different colors, and the representative points of each cluster are highlighted. This, and all of the other *DockAnalyse* output representations, allow for an easy and visual interpretation of the docking procedure (Amela et al. 2010).

These graphs give a global vision of the PPI process, enabling a curated study of the most interesting PPD solutions. Figure II-8 represents a bidimensional graph that depicts the representation in a plane of a cloud of points in a multidimensional space (here 8 dimensions) after the application of a method to reduce dimensions (in this case Principal Component Analysis [Jolliffe, 2002]). As is well-known, the two first principal components explain the main part of the variability contained in the data, but usually it is impossible to reflect all of the distances among the points in a multidimensional space in only two dimensions, which is why outlier points might appear very close to other cluster centers. To check that point, the alternative representation in the script where all combinations of axes are depicted can be activated as is detailed in the "readme" file of the application that can be found at: http://bioinf.uab.es/rker/DockAnalyse/DockAnalyse.zip).

The representative solutions of each of the calculated clusters can be highlighted and they refer to the significant points among all of the docked structures tested (See again Figure II-8). These points represent the most relevant solutions obtained from the PPD calculation and they allow us to identify which solutions among all might be more directly involved in the PPI process. These representatives (or representative solutions) are central members of the clusters and they also have high interaction energy. A strong point of *DockAnalyse* is that it reduces the number of solutions to analyze after the PPD experiment and, therefore, the docking output-data analysis is facilitated because the number of solutions to check is reduced from a huge number (e.g., 1000) to approximately less than 10 in most of the cases. Some PPD programs do not incorporate a clustering process and the use of *DockAnalyse* in these cases is even more justified. Evidently, *DockAnalyse* gives researchers the possibility to use it with a greater or lesser number of docking

solutions although this characteristic has been proposed to guarantee the exhaustive exploration of the whole space of docking solutions.

The main advantage of *DockAnalyse* when trying to interpret the results of a docking procedure, is shown in Figure II-8 and in Figure II-9, where it can be seen that obtaining conclusions from the graphical representations given by *DockAnalyse* is much more intuitive than from the raw numerical data given by most of the PPD programs (See Figure II-7 of the Materials and Methods section).

Testing DockAnalyse

Through a set of 35 Enzyme/Inhibitor or Enzyme/Substrate protein complexes of the Protein-Protein Docking Benchmark 3.0 [Hwuang et al., 2008] (which are labeled with an "E" in the benchmark table), we have shown the way by which *DockAnalyse* can be applied in a systematic way to monitor the quality and type of docking predictions. This group of known protein complexes included homodimers and heterodimers, protein inhibitors and other enzyme complexes. The unbound structures of the interacting proteins were used when available; otherwise, the bound structures were extracted from the complex.

The percentage of satisfactory dockings detected was 51.43%, where a satisfactory docking is the one on which one or more *DockAnalyse* clusters are significant in terms of a high number of members and high average interaction energy (These can be easily seen through *DockAnalyse* graphical outputs). In comparison to the crystallographic protein complex structure, which was obtained from the benchmark set, all of these satisfactory solutions showed a very low RMSD. This means that in these cases only through *DockAnalyse* outputs could be seen that the dockings were credible before realizing that the RMSD was so low. On the contrary, the percentage of unsatisfactory dockings detected was

28.57%, where an unsatisfactory docking means that all of the clusters given by *DockAnalyse* are composed by few members and, moreover, have very low interaction energies. The RMSD values calculated here were all very high. Again, with *DockAnalyse* these unsatisfactory dockings could be detected before knowing their high RMSD values. In general, RMSD values could be calculated because we knew the crystallographic structure of the protein complex from the benchmark, but in real research it will almost never be known, so *DockAnalyse* might be used at this point to guide the researcher concerning the quality and credibility of the docking. In addition, 17.14% of the dockings were considered to be static interactions with a RMSD again very low in the considered solutions. The way by which *DockAnalyse* can detect this type of static interactions is explained below.

These kind of interactions could be detected with our program due to the possibility to perform personalized searches, introducing specific PPD solution values in the "marker" variable of the program source code (See again the "readme" file of the program that can be found at: http://bioinf.uab.es/rker/DockAnalyse/DockAnalyse.zip). In Part A of Figure II-9, some of the best solutions of the docking program (with optimal RMSD values and good interaction energies) are interpreted by *DockAnalyse* as a clear trajectory. These values do not belong to any cluster so, consequently, they are included initially in Cluster 0, but the graphical representations provided by this tool and the different information given could help the user to realize that he is in front of two proteins with a small binding site and without any permitted flexibility.

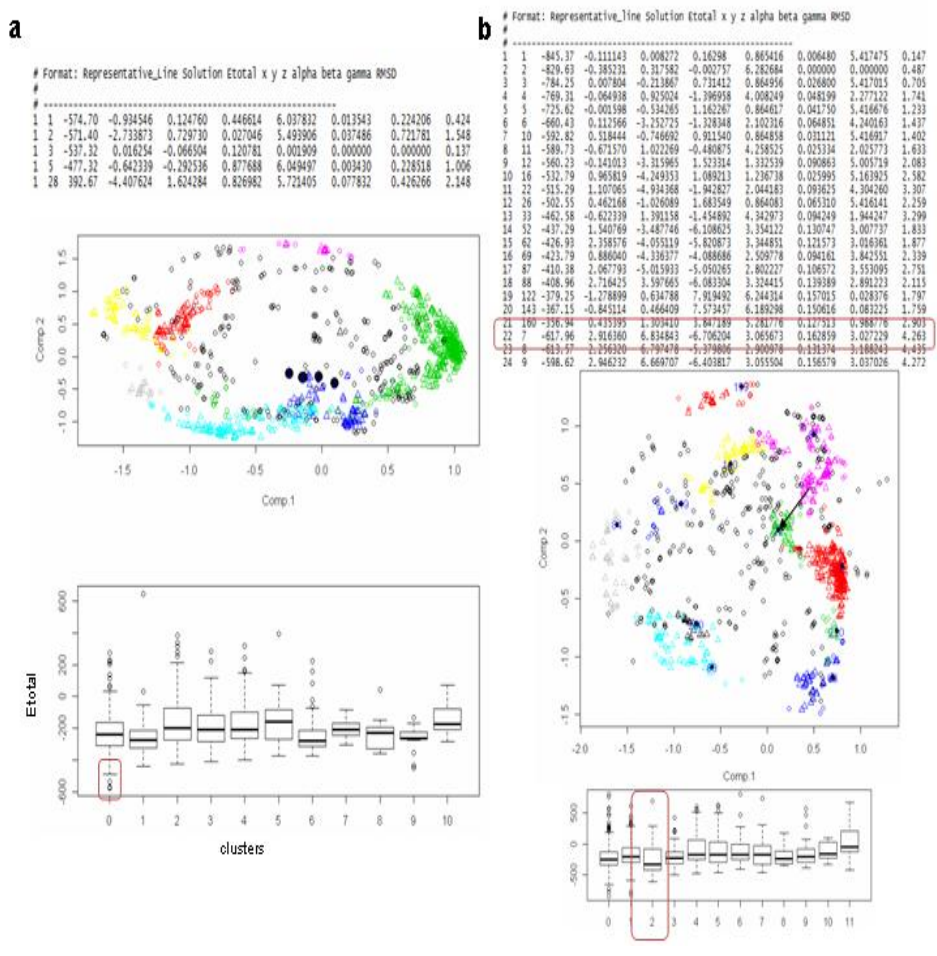


Figure II-9. Initially low-scored docking solutions might be important and considered with the use of *DockAnalyse.* Firstly, the best solutions of the original results of the PPD program with each RMS deviation values calculated are shown. Below, these optimal solutions are highlighted in the graphical representations provided by *DockAnalyse.* These configurations can be interpreted as a binding site with a high level of constraints that seem to describe an interaction pocket. Despìte not belonging to any representative cluster, their high interaction energies reveal this type of static contact. (b) A section of the raw output data-file of the PPD program as well as two of the graphical representation outputs obtained with *DockAnalyse* for these same data are depicted. The most representative *DockAnalyse* solution is highlighted in the three sections showing the ability of this new application to consider important alternative solutions. (Amela et al. 2010)

As has been reported thoroughly in previous sections, the main potential of this new method is the capability to explore the interaction space, making clusters that correspond to extensive contact regions. These graphical representations reproduce the movements that occur between the constituents of a protein complex. These pre-assumed non-optimal solutions described before have lower scores in the PPD program output-file; therefore, they are discarded by the PPD initial filter. As shown below in the example, these solutions could be rescued and their score improved with *DockAnalyse* application. In Part B of Figure II-9, Solution 7 has been included in Cluster 2 in *DockAnalyse* results, but in the PPD program it is ranked as Solution 22, far from the optimal solution although both RMSD and energy values are significant. It has to be considered that *DockAnalyse* highlights this solution as one of the most representative because it is at the center of the cluster with the highest interaction energy (Cluster 2). Moreover, as can also be seen in In Part B of Figure II-9, this cluster is in a highly connected interaction zone, demonstrating displacements among the two docked proteins. These are the types of results that could be obtained using *DockAnalyse*.

For Protein Complex 3 (PDB: 1BVN) of the Protein-Protein Docking Benchmark 3.0 [Hwuang et al., 2008], *DockAnalyse* outputs showed a satisfactory docking in which Cluster 14 was significant. Using the supplementary scripts that come with *DockAnalyse* and can be found at http://bioinf.uab.es/rker/DockAnalyse/DockAnalyse.zip, all of the ligand positions of the solutions of Cluster 14 were extracted as PDB files and then loaded in a protein modeling and visualization tool with the structure of the receptor. As can be seen in Figure II-10, all of the ligand positions contained in this cluster were very similar and, therefore, corroborated the robustness of *DockAnalyse*. Furthermore, that is another useful way to apply our program.

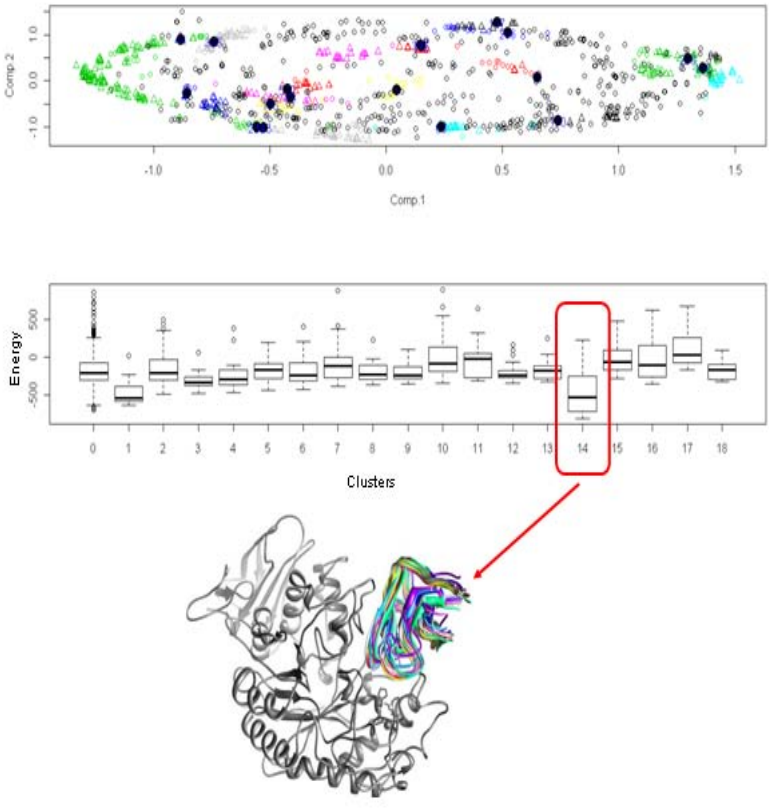


Figure II-10. Tridimensional visualization of the selected cluster. One of the graphical results obtained with *DockAnalyse* for the protein complex of PDB:1BVN of the Protein-Protein Docking Benchmark. Cluster 14 is significant in terms of cluster members and average interaction energy. All of the ligand structures of the previously selected cluster (depicted in different colors and displayed in the "ribbons" format) are viewed in 3D on the receptor (depicted in gray and displayed in the "ribbons" format) (Amela et al., 2010).

DISCUSSION:

Modeling a protein complex

As an example of a procedure where *DockAnalyse* can be applied to model the movements between the members of a protein complex is as follows (This might be similar, in general.):

Isu1 and Isu2 are two yeast mitochondrial proteins which perform a scaffolding function during the maturation of Iron-Sulfur Cluster (ISC) prosthetic groups [Gerber and Lill, 2002; Lill and Mühlenhoff, 2006]. These proteins physically and functionally interact, leading to the formation of a stable protein complex [Gerber et al., 2004]. To achieve the appropriate orientation between these two proteins, we have seen that Isu2 comes into contact with Isu1 and slips on it with the aim of reaching a certain orientation. In this appropriate position, the two proteins are situated one in front of the other and their tails might allow for the required stable interaction. Moreover, in this final conformation, three cysteine residues per protein (which typically conform an iron binding pocket) remain close enough to each other to be crucial for anchoring the ISC that is being generated while Isu1 and Isu2 tails facilitate their interaction [Mühlenhoff et al., 2003] (See Figure II-11). Most of the studies prompt the suggestion that the iron and sulfur atoms required for the ISC biogenesis on Isu1/Isu2 are donated by other proteins, named Frataxin and Nfs1 respectively [Lill and Mühlenhoff, 2006]. This ISC biogenesis machinery is not yet well understood and problems in it cause several human diseases linked to protein/enzyme deficits. That is why the study of this prosthetic group generation represents an important challenge from any point of view.

The sequence, structure, function, interaction and current literature of these proteins were analyzed in-depth. After that, PPD experiments were performed between the structures of the two proteins,

setting up a small rotation step to exhaustively explore a great number of solutions in a reasonable computing time. Finally, *DockAnalyse* was applied with the aim of reducing the huge amount of docking solutions obtained to several representative ones. These solutions were the 4th, 78th, 28th, and 1st initially ranked solutions of the Escher NG docking output data-file. Here, the main utility of *DockAnalyse* in reducing the number of solutions to analyze after a PPD calculation is demonstrated. For these four representative docking solutions, the protein structure (PDB) files were obtained, merged into a trajectory file and then subsequently loaded into a protein modeling and visualization tool with which we could analyze them. This procedure allowed us to build a point-to-point pseudo-trajectory with which we could postulate a model to explain the surface displacements between the given proteins (See again Figure II-11). This pseudo-trajectory could be reconstructed by means of the selection of other solutions along *DockAnalyse* clusters or by joining the different *DockAnalyse* cluster representatives. For this reason, the representative solutions could be considered to be static frames that describe the motion between the interacting proteins, and we could model/study the surface displacements of one protein on the other.

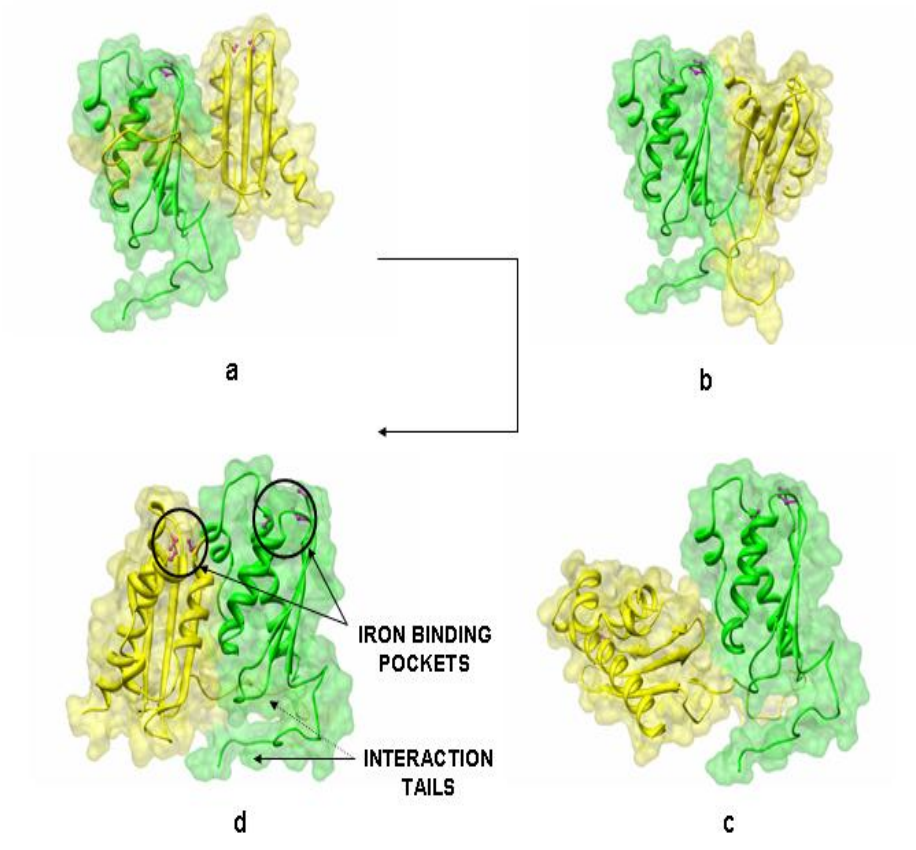


Figure II-11 Example of a protein complex modeling. The images (a) -> (b) -> (c) -> (d) represent the modeled structures of the solutions given by *DockAnalyse* for the docking between proteins Isu1 and Isu2. The structures are displayed in the "surface" and "ribbon" formats and colored in green for protein Isu1 and yellow in the case of Isu2. The iron binding pocket of each of the proteins, which is composed of 3 cysteine residues, is displayed in a "ball and stick" format and colored in magenta. Isu1 and Isu2 iron binding pockets and interaction tails are labeled. The edges attempt to show the trajectory that may occur when these proteins interact to finally acquire the desired configuration required for ISC biogenesis (Amela et al. 2010).

Use of DockAnalyse

A comprehensive tool for the analysis of PPIs has been designed. This new application permits a better interpretation of the obtained PPD solutions as well as the surface displacements that may occur during the interaction between the proteins of a protein complex. Therefore, this tool guides the modeling of a protein complex and can be applied in a systematic way to monitor the quality and type of docking predictions through global or local visions of the docking results that facilitate the decision making process regarding the docking characteristics. The simplicity in applying this tool and the ease in interpreting the PPD solutions makes it ideally suited to analyze the data obtained in a PPD experiment. Considering all of the facts stated above, to go further and propose new functional interpretations for the proteins of interest might be much easier. In terms of these new hypotheses, when the initially docked proteins are monomers, a proposal on the putative structure of a multimeric protein complex might be postulated [Jackson et al., 1993 Cohen et al., 2005]. Another procedure to visualize the expected surface displacements between two interacting proteins may be suggested. This last approach could be applied to pairs of proteins that require displacements between them to fulfill a specific function [Lill and Mühlenhoff, 2006].

As a whole, *DockAnalyse* could be used after a docking assay in the context of a more complex procedure where a model of the behavior between the proteins that take part in a biologically functional protein complex would be performed. A schematic description of how to use *DockAnalyse* in this whole bioinformatics procedure is shown in Figure II-12.

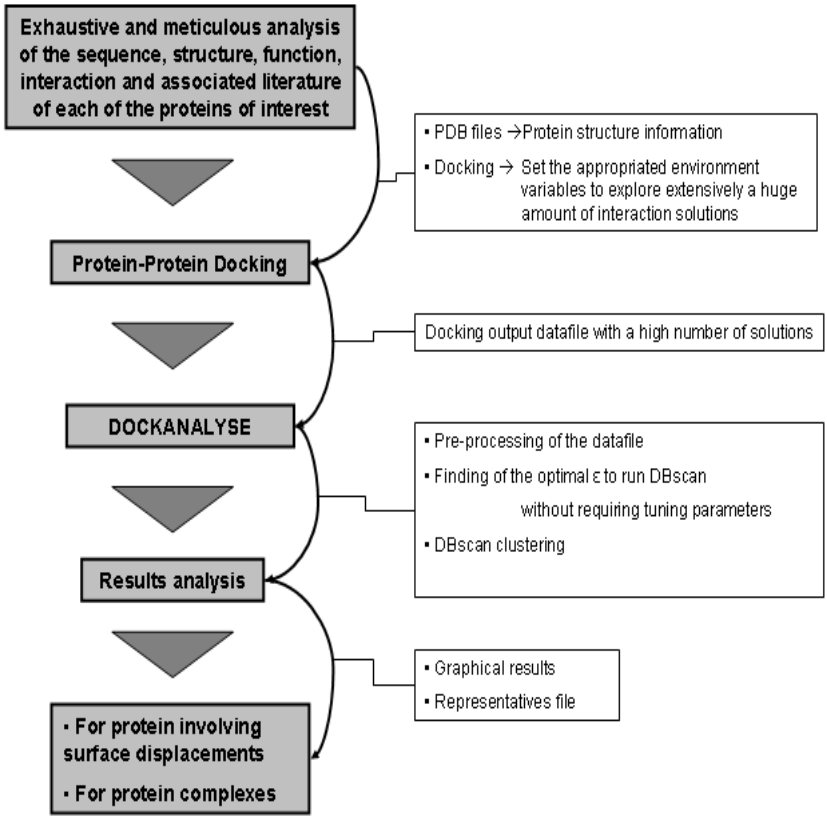


Figure II-12 DockAnalyse: how and when. *Schematic flowchart where the sequential steps of the bioinformatic study in which* DockAnalyse *might be used are described (Amela et al. 2010).*

First of all, an extensive literature mining analysis coupled with a profound study of the sequence, structure, function and interactions of the proteins of interest is required. Secondly, the required PPD 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 algorithm has to be applied to each of the docking output data-files to obtain the representative docking solutions among those thousands

calculated. Lastly, manual curation of the obtained docking representatives might be necessary to fit the solutions given with the appropriate biological function and to eliminate the putative aberrant results. The combination of theoretical docking procedures with the available experimental information is shown to greatly improve the modeling. *DockAnalyse* is accessible at:

http://bioinf.uab.es/rker/DockAnalyse/DockAnalyse.zip

CONCLUSIONS:

- A new program for the analysis of protein-protein interactions named *DockAnalyse*, which provides graphical and visual representations that facilitate the interpretation of docking results, has been created.

- This tool gives the user many possibilities and some representative docking solutions to allow for an easy protein complex modeling process. It is accessible at:
http://bioinf.uab.es/rker/DockAnalyse/DockAnalyse.zip .

- As an example, the modeling of the dynamic behavior of the interactions of a certain protein complex has been done applying our new approach.

CHAPTER III:

Iron-Sulfur Cluster Biogenesis and Friedreich's Ataxia

INTRODUCTION:

A combination of several protein bioinformatic tools and some of the previously mentioned PPD programs were used in this chapter. The new application designed in our lab, DockAnalyse, which is extensively described along Chapter II, was also employed here to select the representative solutions of the docking experiments. Moreover an exhaustive literature examination and the use of different structural bioinformatics programs were required in this section of the thesis with the aim of modeling the ISC assembly protein complex.

Iron-Sulfur Clusters

ISCs are prosthetic groups formed by iron ions and inorganic sulfide that are present in proteins of all the organisms along the evolution and represent one of the most flexible and ingenious metal cofactor. These structures are basically ligated to proteins by cysteine residues and perform many different functions such as mitochondrial respiration. The biosynthesis of ISC is carried out by complex protein machinery that, in eukaryotes, is placed in the mitochondria. In the initial ISC assembly step, a protein complex composed by an iron donor (Frataxin), a sulfur donor (Nfs1), and a scaffold protein (Isu) is formed. Problems affecting these proteins cause distinct diseases such as FRDA, which is due to Frataxin deficits. The principal forms of ISCs typically present in proteins are [2Fe-2S] and [4Fe24S] (See Figure III-1).

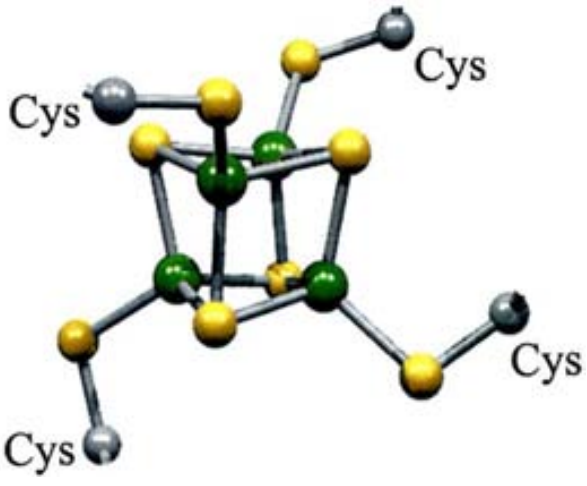


Figure III-1. Structure of a [4Fe-4S] iron-sulfur cluster. Iron atoms are shown in green, sulffur in yellow and those of cysteine residues in grey (Frazzon 2001).

The Friedreich's Ataxia syndrome

FRDA a human, neurological, progressive and hereditary disease which, through the nervous system, spinal bone marrow, neurons, and cortico-spinocerebellar routes, affects the equilibrium and movement coordination. Moreover, along with other symptoms, it causes muscle weakness and heart hypertrophy. This disease is the most common autosomal recessive ataxia in Caucasians and it is associated with a pronounced lack of a conserved mitochondrial protein of a not fully understood function, called Frataxin. This protein is encoded by the gene *fxn*, initially termed *x25*, which is located in the 9q13 chromosome region, is of 80 Kb and is constituted by seven exons, five of which encode the protein. FRDA is classified in the group of diseases that are caused by an expansion of a DNA triplet (like Huntington disease). Even so, while in the protein affected in Huntington disease (Huntingtin) the expanded DNA triplet (CAG) generates an aberrant polyglutamine protein, in this particular case the protein is correctly produced even in

low levels because the expansion of a GAA triplet in the first intron of the gene generates an aberrant structure of the DNA helix (sticky DNA) that impedes/reduces its transcription and expression. In the normal population, the GAA motif is polymorphic with a number of repetitions, varying from 6 to 36 while, in individuals affected by the mutation, the repetitions are increased up to 1000 [Campuzano et al., 1996; Pandolfo, 2009].

The protein Frataxin

Whereas the protein Frataxin in humans (PDB code: 1EKG) has 210 amino acids and does not belong to any characterized protein family, in yeast the protein is named Yeast Frataxin Homolog 1 (Yfh1, PDB code: 2GA5) and is composed of 174 residues. The protein is mitochondrial, even encoded in the nucleus and, therefore, contains a mitochondrial targeting sequence. Besides, it is ubiquitous in both organisms, but its expression is higher in tissues that require huge amounts of energy like, for instance, the spinal cord, muscles or heart. It is highly conserved during evolution, with homologs in mammals, yeast, prokaryotes and plants [Dhe-Paganon et al., 2000; He et al., 2004; Pandolfo and Pastore, 2009]. From a general point of view, Frataxin is a compact and globular protein with a well characterized tri-dimensional structure constituted by two α-helices and five to seven aligned anti-parallel β-sheets that form an α/β sandwich. The most conserved protein domains, from prokaryotes to human and from sequence to structure, correspond to these five β-sheets and one of the α-helices mentioned before that establish an acidic area capable of binding iron with low affinity [Cook et al., 2006; Bencze et al., 2006; Foury et al., 2007; Correia et al., 2010]. Due to the nature and size of the Frataxin conserved regions, key interaction functions are suggested for these protein zones, both with another protein or with a ligand. Situated at the core of Frataxin is a concentration of hydrophobic amino acids

which are essential for structure stabilization and, therefore, cannot be substituted [Correia et al., 2006; Prischi et al., 2009].

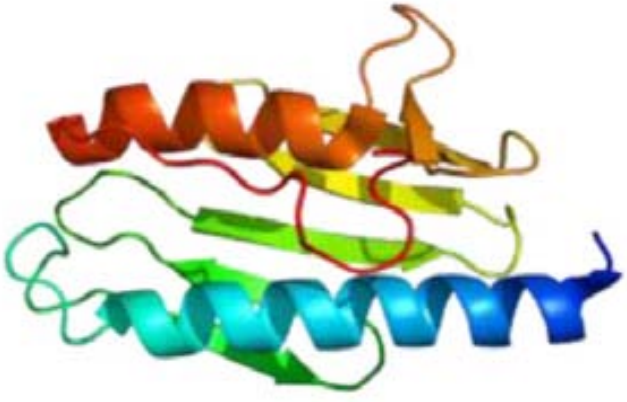


Figure III-2. Human Frataxin structure. PDB renderig based on 1EKG (source http://en.wikipedia.org/wiki/Frataxin).

Frataxin function

Several functions for Frataxin have been proposed, always associated with iron accumulation inside the mitochondria and increased sensitivity to oxidative stress [Delatycki et al., 1999; Seznec et al., 2005]. Frataxin functions are based on iron oxidation, iron binding, and, more recently, it has been suggested that it plays an important role in early stages of ISC assembly/maturation, avoiding the depletion of proteins like Aconitase and respiratory chain complexes I-II-III, inside and outside of the mitochondria [Chen et al., 2002; Bulteau et al., 2004; González-Cabo et al., 2005; Martinelli et al., 2007]. Some evidence indicates that Frataxin's main function is in ISC biosynthesis, and that is when iron deregulation/accumulation occurs inside the mitochondria [Puccio et al., 2001]. The yeast ISC assembly machinery is basically constituted by scaffold proteins (Isu-type proteins) and other different proteins which donate the iron and sulfur atoms. Yfh1 is the protein that has been postulated as contributing with

the iron, while the protein Nfs1 has been proposed as being the donor of sulfur. Recently, it has been demonstrated that the small protein Isd11 is coupled with Nfs1 and might mediate the interaction between Nfs1 and Isu and has been shown to be essential for Nfs1 action [Adam et al., 2006; Wiedemann et al., 2006; Shi et al., 2009] (See Figure III-3).

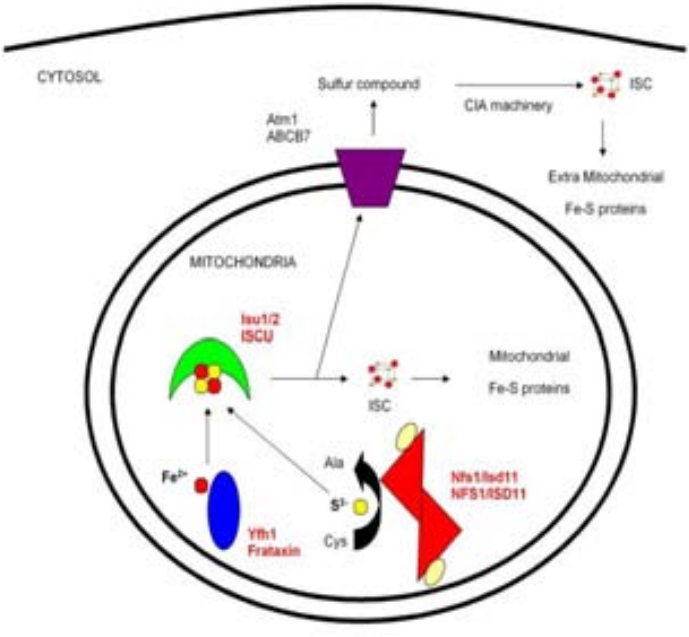


Figure III-3. Iron-Sulfur cluster biogenesis key proteins. Schematic representation of the proteins involved in the initial ISC assembly process inside the mitochondria (Amela et al. 2013).

This machinery is located inside the mitochondria and it is evolutionarily very well conserved. Frataxin can handle iron atoms that might be transferred to the scaffold ISC protein Isu, suggesting that it acts as a chaperone in the initial part of ISC formation [Lill and Muhlenhoff, 2006; Lill, 2009]. In addition, there are other studies proposing that

Frataxin oligomerizes to act for example as an iron storage protein with the aim of avoiding oxidation, free radical generation and, therefore, toxicity [Karlberg et al., 2006; Li et al., 2009].

Yeast as a human model

Not only is yeast one of the eukaryote organisms in which more functional genomics and interactomics data have been reported, but it also is an excellent model for gaining insights into the molecular basis of human rare mitochondrial disorders. Moreover, it is a very well characterized system from which most of our current knowledge about mitochondrial biogenesis genetics and biochemistry is derived [Barrientos, 2003]. This makes yeast ideally suited to propose different mutants to better understand the ISC biogenesis system inside of the mitochondria, both in yeast and human. Many mitochondrial proteins have been used for the later identification and characterization of their human homologs. As the genetic manipulations are more difficult in complex cells, as in human, model organisms such as yeast have been used to facilitate these experiments because of the genetically easy treatment and because many features of eukaryote physiology are evolutionarily conserved in it. Furthermore, the patient clinical trials are arduous to be held and the yeast model enables new putative drug testing to find new therapeutic candidates against these mitochondrial diseases [Smith and Snyder, 2006; Schimmer et al., 2006]. Normal human Frataxin is able to restore the defects of yeast Frataxin deficient cells, while the human mutant Frataxin is unable to do so, strongly suggesting that the function of Yfh1 is conserved in human Frataxin. All of the required protein sequences and some protein structures have been elucidated in *Saccharomyces cerevisiae,* which, moreover, possesses the human protein homologs

needed to study our purpose. There is clear evidence that the yeast and human Frataxins are orthologous proteins [Knight et al., 1999]. In yeast, it has been demonstrated by co-immunoprecipitation that Yfh1 interacts with its protein partner Isu which, as mentioned above, is the matrix mitochondrial scaffold protein for ISC assembly, and this may be similar in the human counterparts. Moreover, it has been shown that Isu interacts with the protein complex Nfs1/Isd11, and all of these mentioned proteins generate the central platform for ISC assembly [Rawat and Stemmler, 2010] (See again Figure III-3). To summarize, Isu is the scaffold protein, Nfs1 is the sulfur source and Frataxin donates the iron. This represents an interesting question: How exactly does Yfh1 interact with its protein partners to generate the central platform for ISC assembly? Can bioinformatic tools help us to predict a model regarding how this group of proteins interacts? Taking these premises into account, the approach has been performed using the yeast protein model because it provides a complete molecular system to study, in detail, all of the pieces of the ISC biogenesis process in general.

OBJECTIVES:

- Characterize the proteins Frataxin, Nfs1, Isu and Isd11 from the sequence, structure, function and interaction point of view.

- Improve the current model of ISC biogenesis protein complex and study the dynamic behavior of its components to propose a new dynamic model of the ISC assembly process in yeast.

- Have a better knowledge about the molecular pathology of the ISC deficits occurring in FRDA.

MATERIALS AND METHODS:

General bioinformatic analyses of the proteins

The sequences of Frataxin, Isu and Nfs1 from evolutionarily distinct organisms were retrieved and used to perform different sequence multi-alignment analyses in each of the cases [Higgins and Sharp, 1988; Chenna et al., 2003]. As found in the literature, we realized that these proteins have a high score of sequential homology, indicating that they are much conserved during evolution. In addition to these analyses, several classical bioinformatics studies were made to set up some characteristics for each of the proteins, which were also contrasted with the bibliography. Although Isu and Nfs1 have no already solved tri-dimensional structures available in the PDB [Berman et al., 2000], the sequential and structural homology to some already-solved protein family members, allowed for the modeling of these proteins. Secondary and tri-dimensional structure models for these proteins were obtained by applying three widely used applications designed for this purpose. On the one hand, the secondary structure was predicted using PsiPred [McGuffin et al., 2000], which incorporates neuronal networks to the outputs of PSI-BLAST. On the other hand, tri-dimensional structures were built using ESyPred3D [Lambert et al., 2002], which is a homology-based application that uses the Modeller package [Fiser and Sali, 2003]. 3D-PSSM and Phyre were also used to corroborate the results obtained [Fischer et al., 1999; Kelley et al., 2000; Bennet-Lovsev et al., 2008]. Regarding protein Isd11, there was not enough sequence homology to any existing protein to model its structure in any of the cases so, therefore, we applied the Robetta full-chain protein structure prediction server that uses the Rosetta *de novo* method and enables protein modeling without any detectable homolog

[Chivian et al., 2003; Kim et al., 2004; Chivian et al., 2005]. Both the 2D and 3D predicted information for all of the proteins was combined with the current knowledge about the proteins in order to obtain feasible structures. In the case of Frataxin, the tri-dimensional solved structure was found in the PDB under PDB code: 2GA5. With the purpose of identifying/corroborating the protein interaction regions in Frataxin, Isu, Nfs1 and Isd11, ProMate and meta-PPISP, which are protein structure based programs, were used [Neuvirth et al., 2004; Qin and Zhou, 2007]. After that, PPI-Pred, which is a support vector machine based program, was applied to corroborate the previous results [Bradford and Westhead, 2005]. To elucidate the experimentally found protein interaction partners of these proteins, searches in different interactomics databases were performed [Prieto and De La Rivas, 2006]. The situations of Frataxin iron atoms were predicted with the Autodock force field after computing molecule charge with the Gasteiger method. This was performed using the VEGA ZZ 2.3.2 Molecular Modeling Toolkit [Pedretti et al., 2002; Pedretti et al., 2003; Pedretti et al., 2004] . To evaluate the feasibility of the added ions, ArgusLab 4.0.1 was employed [Thompson; www.arguslab.com]. Special emphasis has to be made in remarking that all the above mentioned studies were always complemented with the proper literature information to contrast the data obtained.

Protein docking tools used

Docking essays among the proteins Frataxin, Isu, Nfs1 and Isd11 were performed with the Escher NG protein-protein automatic docking system of the VEGA ZZ project and Hex as the main tools [Ausiello et al., 1997; Ritchie et al., 2008]. Regarding Escher NG and Hex, several docking control values were tested to achieve proper docking executions and to obtain good output files to then be analyzed. With the aim of validating the

results obtained, other docking protocols, like BiGGER of the Chemera 3.0 package and HADDOCK, were also used [Palma et al., 2000; Dominguez et al., 2003; de Vries et al., 2007; de Vries et al., 2010]. HADDOCK is a docking tool that employs biochemical and/or biophysical interaction data such as bioinformatics predictions and, therefore, the previous data obtained from the bioinformatics analysis of the proteins could be used to dock all of the proteins together and see how they are predicted to interact when forming the ISC biogenesis protein complex from a general point of view. As previously emphasized, it must be taken into account that some expert knowledge about the living-protein context during the ISC biogenesis in yeast was necessary to finally propose a coherent model. Most of the protein docking programs used in our experiments, such as Hex, BiGGER and HADDOCK, have taken part in different rounds of the CAPRI (Critical Assessment of PRediction of Interactions) experiment, showing satisfactory results in protein docking structure prediction. To select the most representative solutions for each of the docking output files from Escher NG and Hex, DockAnalyse was applied [Amela et al., 2010]. This program relies on an algorithm based on the DBscan clustering method, which searches for continuities between clusters, generated by the output docking data representation and, moreover, solves some of the inconsistency problems of the classical clustering methods. In addition to the interaction energy, the program considers the density of solutions around the representatives and, furthermore, does not need any tuning parameter from the user. The structures for the most representative docking solutions retrieved from DockAnalyse were obtained and loaded in several modeling or visualization tools commonly used in structural bioinformatics. This procedure allowed us to postulate a model with which the putative surface and rotation displacements between the initially docked proteins were monitored.

Finally, manual curation was necessary to link the solutions obtained with the proper biological function, discarding the putative aberrant results.

The final protein complex modeling

Escherichia coli is the most studied model organism and it has crystallographic structures for the protein complexes IscS/IscS (PDB code: 3LVM) and IscS/IscU (3LVL), homologous to Nfs1/Nfs1 and Nfs1/Isu. Thus, the Nfs1/Isu tetrameric protein complex was basically modeled from *Escherichia coli* 3LVM and 3LVL structures, using Modeller, and then combining them with the DeepView-Swiss-PdbViewer application [Guex and Peitsch, 1997; Fiser and Sali, 2003] (See Part A of Figure III-4).

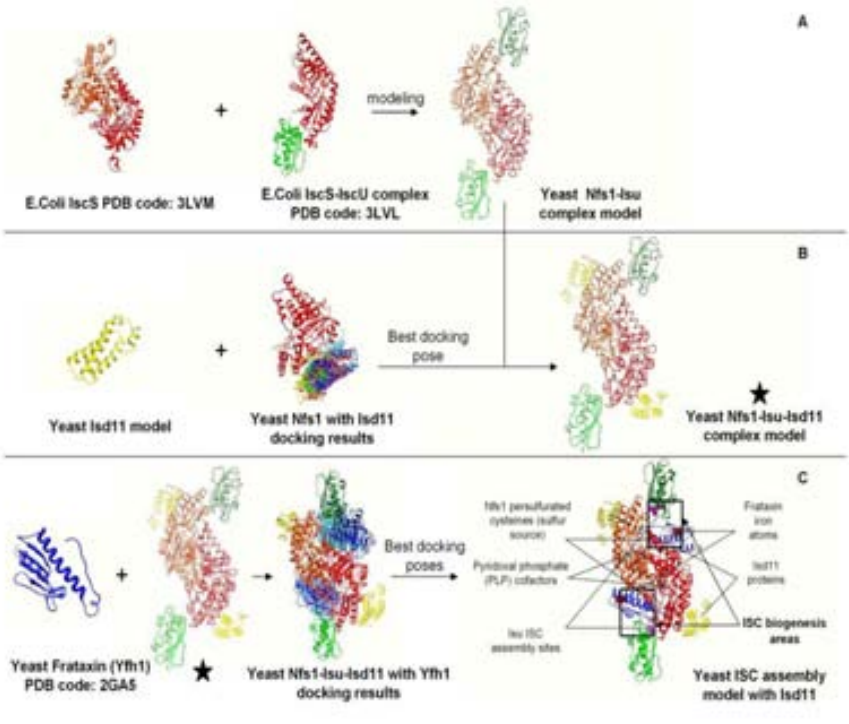


Figure III-4. The modeling process of the yeast Iron-Sulfur cluster assembly protein complex. As can be seen in (A), the Nfs1-Isu protein complex was modeled from *Escherichia coli* already solved structures under PDB codes: 3LVM and 3LVL (IscS and IscS-IscU dimers, respectively). In (B), the novel protein structure of Isd11, which was modeled by means of the Robetta *ab initio* modeling server, is shown. Moreover, the Nfs1-Isd11 interaction structure was obtained from the results of the docking assays between these proteins. Finally, the Nfs1-Isu-Isd11 protein complex was created combining the structures of (B) with those of (A). Regarding (C), the entire Nfs1-Isu-Isd11-Frataxin protein complex was completed docking the (B) protein complex with yeast Frataxin (PDB code: 2GA5). The docking results were consistent with the current literature (Amela et al, 2013).

The Isd11 yeast protein was, for the first time, modeled using the *ab initio* structure prediction server Robetta [Chivian et al., 2003; Kim et al., 2004; Chivian et al., 2005]. In the case of the *ab initio* predictions, four feasible models for Isd11 structure were obtained. From these models, it was easy to discriminate an almost certainly right fold because the confidence mean was high in all cases. It must be taken into account that confidence means equal to or greater than three means likely correct parents. Additionally, a structural fitting that superimposes the models has been made in order to show the similarity between those Isd11 predicted structures (See Figure III-5).

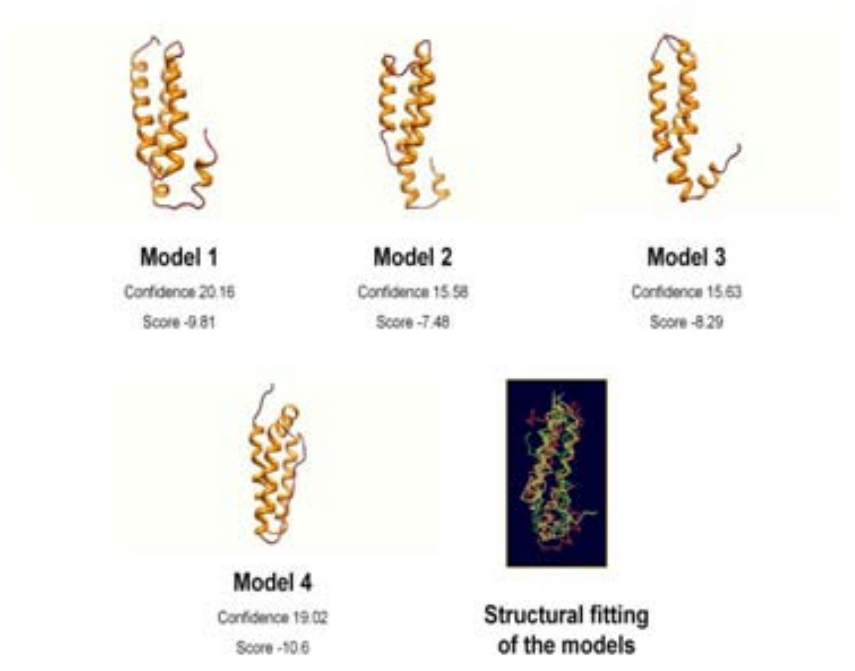


Figure III-5. Proposed models of the protein Isd11. This figure depicts the structure of each of the four Isd11 models obtained from the Robetta method along with their confidence and global score. As can be seen, all models have very similar structures demonstrating the reliability of the modeling process. A structural fitting is shown to corroborate the rightness of the models. (Amela et al. 2013).

According to these data, no matter what of the models could be used for our studies. All of these novel protein structures were then docked to Nfs1 and, therefore, Isd11 could be added to the entire protein complex representing one of the novelties of our model. The same process was performed as before with DeepView-Swiss-PdbViewer to combine the models obtained (See Part B of Figure 2). Frataxin was also docked to the intermediate protein complex and we confirmed that it preferably

binds the protein complex in the regions where the current literature already proposes [Shi et al., 2010; Cook et al., 2010], thus the final yeast ISC assembly static model obtained in our process is in concordance with the currently proposed (See Part C of Figure 2). The Nfs1 structure indicates conformational plasticity both of the protein and of a long loop containing a cysteine essential for its fuction (See Results and Discussion section). These putative conformational changes were examined with several hinge prediction algorithms, and the expected movements were obtained both for the whole protein and the loop [Krebs and M. Gerstein, 2000; Emekli et al., 2008] (See Part A of Figure III-6).

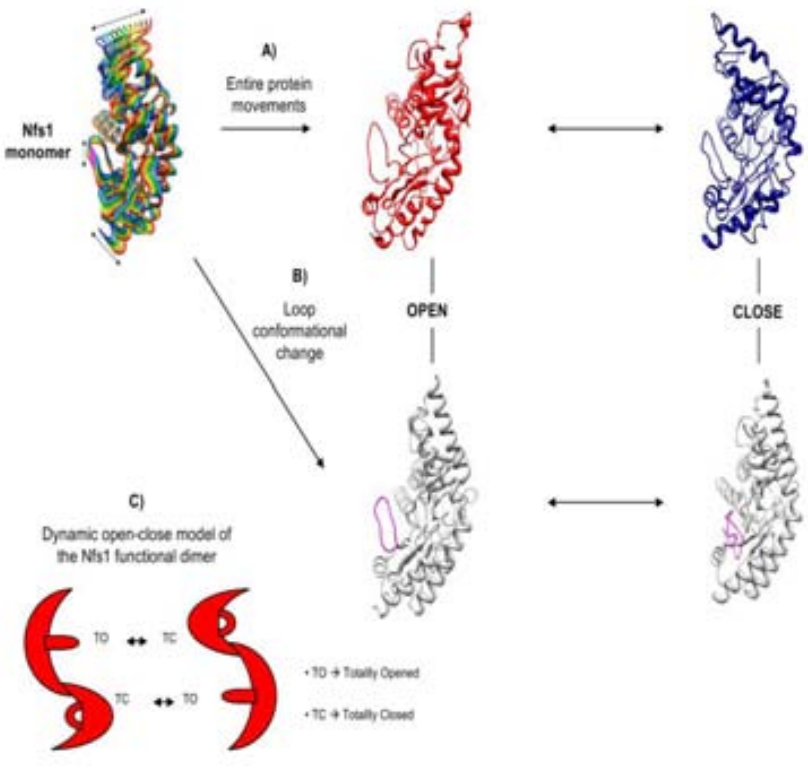


Figure III-6. Nfs1 and its open-closed loop conformational changes. The predicted hinges and putative macromolecular movements of protein Nfs1 are shown. The modeled open-closed conformational changes of the entire protein are presented in Part A, while those of the flexible loop are displayed in Part B. In Part C of the figure, the dynamic behavior proposed for the functional Nfs1 dimer is depicted (Amela et al. 2013).

Regarding the loop, the structure for both the closed and the extended states was modeled. The loop is closed when its cysteine residue is buried near the Nfs1 catalytic area where the PLP cofactor is situated, whereas the loop is extended when this already persulfurated cysteine residue becomes exposed. The extended loop state was modeled directly from the *Escherichia coli* solved structure deposited in the PDB under code 3LVL, while the modeling in the case of the closed state was not as trivial as before and required several steps. Starting from the protein IscS of *Escherichia coli* (PDB code: 3LVM), VAST was used in order to find homolog structures where the loop had a closed position. The VAST service allows searching for structural neighbors, starting with an initial structure [Gibrat et al., 1996]. Taking a look at the VAST output, the structure showing the closest loop and the best structural homology was 1T3I_A. This structure was remodeled with Modeller, forcing the loop to be more closed, and ten models were obtained. The best model was selected after corroborating its plausibility with the EasyModeller application [Fiser and Sali, 2003; Kuntal et al., 2010]. Automated loop models were also generated; however, worse models resulted in this case, thus discarding this automatic method. Regarding the substrate (PLP cofactor), the structure under PDB code: 1N31_A was found with the Ligand option of the PDB service and it was superimposed on the previous model with SuperPose [Maiti et al., 2004]. After that, the DeepView-Swiss-PdbViewer application was used to fit

the substrate and calculate the distance from the cysteine residue of the loop to the PLP cofactor, which is required for cysteine persulfuration [Guex and Peitsch, 1997]. Still, the distance seemed not to be appropriate, so HingeProt was used to obtain a structure with a more closed loop [Emekli et al., 2008]. The desired distance was now achieved, so the entire model was rebuilt once again with the DeepView-Swiss-PdbViewer application, and finally side chains were added to the model with PulChRa [Guex and Peitsch, 1997:, Fiser and Sali, 2003; Rotkiewicz and Skolnick, 2008]. When the model for the Nfs1 with the closed loop was done, a VAST search was performed again with this new model and the results demonstrated an excellent structural homology with the already solved structure with PDB code: 1KMJ_A. This fact demonstrates the feasibility of the model with the closed loop (See Part B of Figure III-6).

RESULTS:

Putative structure and function of Isd11

Isd11 is one of the proteins recently proposed to take part in ISC biogenesis in eukaryotes along with Nfs1, Isu and Frataxin [Schmucker et al., 2011]. This protein is essential for Nfs1 activity. However, no bacterial homolog of Isd11 has been found suggesting that Isd11 is eukaryotic specific for ISC assembly [Wiedemann et al., 2006]. It has been shown that Nfs1 tends to aggregate, but Isd11 prevents this behavior [Adam et al., 2006]. Using the protein-structure prediction program Robetta, mentioned in the Materials and Methods section, models for Isd11 structure were obtained [Chivian et al., 2003; Kim et al., 2004; Chivian et al., 2005]. Apart from the confidence and score provided by the method, the result of a structural fitting of the obtained models corroborated the feasibility of the different predictions as it shows very high similarity among the tridimensional structure of the models (See Figure III-5 of the Materials and Methods section). Docking of all Isd11 models on Nfs1 showed a clear interaction preference for a delimited area of Nfs1 This fact guided and allowed for the incorporation of Isd11 in our ISC assembly protein complex (See Part B of Figure III-4 of the Materials and Methods section and the last picture of Part C of the same Figure). Until now, no structural data for Isd11 exist, so this is the first time where a tridimensional model of Isd11 has been proposed. Curiously, Isd11 binds to Nfs1 in the same area where Nfs1 docks itself to oligomerize. That is why our hypothesis is that Isd11 interacts with Nfs1 with the aim of preventing its oligomerization and, therefore, allowing for the correct activity of Nfs1. Some studies postulate the interaction of two Isd11 proteins with each Nfs1 monomer, and some others report that Isd11 facilitate Isu and/or Frataxin interaction to Nfs1 [Shan et al.,

2007; Li et al., 2009]. We saw that when introducing two Isd11 proteins in our protein complex model, the second Isd11 and Frataxin are very close together, suggesting implications to favor Frataxin interaction with Nfs1. This was corroborated by the fact that Frataxin preferably binds to Nfs1 when Isd11 is present, and the other way around. Further investigations will be needed to exactly elucidate how many Isd11 proteins interact with Nfs1. Generally speaking; Isd11 may alter the Nfs1 structure in eukaryotes, suggesting crucial differences between ISC biogenesis mechanisms of prokaryotes and eukaryotes.

A hinge on Nfs1 allows for an open-closed conformational change

As shown in the literature, Nfs1 acts as a dimer in the ISC biogenesis mechanism [Prischi et al., 2010; Shi et al., 2010]. Owing to the structural properties of the Nfs1 monomer, we had the idea of searching for putative conformational change zones and we clearly detected the existence of an axis describing an evident hinge in Nfs1 (See Part A of Figure III-6 of the Materials and Methods section). Together with the Nfs1 loop flexibility detailed in the next subsection, these two facts suggest an open-close conformational change between the two monomers of the Nfs1 active dimer (See Part C of Figure III-6 of the Materials and Methods section). This open-closed conformational change may promote the correct interaction with Isu and Frataxin and, together with the loop flexibility, put the cysteine-containing loop within reach of the Isu ISC binding pocket.

The Nf1 cysteine-containing loop is extremely flexible

Many articles published over the last few years have speculated about the flexibility of the Nfs1 cysteine-containing loop, which has been proposed as being the sulfur donor to the ISC assembly on Isu through

the desulfuration of a cysteine to alanine and the generation of a persulfide [Cook et al., 2010; Selbach et al. 2010; Shi et al., 2010]. This process occurs in different steps: (1) The binding of a free cysteine to the Nfs1 pyridoxal phosphate (PLP) cofactor, which is not on the protein surface but rather in a catalytic pocket inside Nfs1; (2) The formation of a PLP-cysteine adduct in this Nfs1 pocket; (3) The transfer of sulfur from this adduct in the Nfs1 pocket to the cysteine of the Nfs1 loop generating a persulfide bond and the release of alanine; and, (4) The persulfurated cysteine of the Nfs1 loop is redirected to the Isu ISC assembly pocket and its extra-sulfur is given to Isu for ISC assembly. As can be seen, Steps 3 and 4 require very big Nfs1 cysteine-containing loop displacements because the loop needs to be in contact with the PLP site on Nfs1 and then with the Isu ISC assembly site, which is quite far away (around 20 Å). Therefore, a lot of flexibility to this loop is supposed [Tirupati et al., 2004]. Nfs1 loop movements are thought to allow not only for the required contacts between the loop cysteine and the PLP-cysteine adduct, but also for the reaching of the Isu ISC binding pocket by the persulfurated cysteine of the Nfs1 loop after contact with the PLP cofactor. In the current models, based exclusively on the available crystallographic structures, a big distance is supposed to be covered by the loop of Nfs1 in order to reach the Nfs1 PLP area and then the Isu ISC assembly site, as explained before. Apart from being an extremely big distance to be covered, it should be taken into account that it is taking place inside of a protein complex with little freedom for the loop movements. From our point of view, it is so difficult task to be only explained by the Nfs1 loop flexibility and, therefore, some additional conformational changes on Nfs1, which are quite controversial from a structural point of view, should be taking place (See the last picture of Part C of Figure III-4 of the Materials and Methods section) [Shi et al., 2010].

On account of these loop movements, we modeled the two extreme loop states, the opened and the closed ones (See Part B of Figure III-6 of the Materials and Methods section). The conformational change in the Nfs1 loop to adopt the closed state permits a better position of Frataxin in our docking results where the residues that carry iron are much closer to the Isu cysteine residues that form the ISC assembly pocket. When docking Frataxin to a Nfs1 dimer where one monomer is in the opened loop conformation and the other is in the closed loop conformation, all of the docking solutions are situated at the Nfs1 monomer with the closed loop showing that Frataxin prefers the closed loop state. That is why we propose an open-close dynamic model for the Nfs1 dimer, which might facilitate the required contacts and reactions for ISC assembly (See Part C of Figure III-6 of the Materials and Methods section).

Iron and sulfur donation

The fact of what is first to be donated, iron or sulfur, to assemble the ISC on Isu is still under debate [Shimomura et al., 2008]. We propose a solution to this question that is derived from our model. The obtained results indicate that Frataxin preferably binds the closed loop conformation. In this state is when iron loaded Frataxin positions its acidic conserved residues closest to Isu ISC binding pocket. Thus, in this closed loop conformation iron can easily be recruited by Isu. Furthermore, in this closed loop conformation is also when the cysteine persulfuration can be done because the cysteine of the Nfs1 loop can contact with the PLP cofactor (See Part A of Figure III-7).

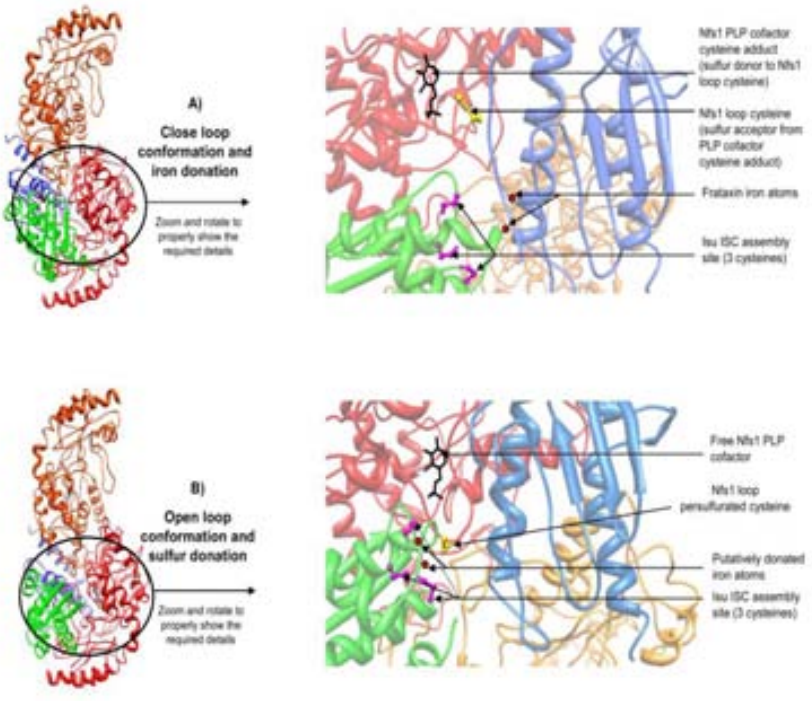


Figure III-7. Structural details of iron and sulfur donation. Following the same colors as the previous figures, (A) shows our Nfs1 loop "close state" model for iron donation where the proximity between Frataxin iron atoms and the Isu iron-sulfur cluster (ISC) assembly site is very high. In addition, the cysteine of the Nfs1 loop in this "close state" model is adjacent to the pyridoxal phosphate (PLP) to be persulfurated. (B) displays our model for sulfur donation where the Nfs1 loop is now in the "open state" and the already persulfurated cysteine of the Nfs1 loop is very close to the Isu ISC assembly site where the previously donated irons are (Amela et al. 2013).

On the contrary, in the opened loop conformation Frataxin binding is hampered, and is in this conformation when the already persulfurated cysteine of the Nfs1 loop approaches the Isu ISC binding pocket. Thus, in this opened loop conformation sulfur can readily approach the Isu ISC assembly site (See Part B of Figure III-7). Take into account that:

(a) sulfur cannot be donated without a previous contact with PLP and the persulfide bond formation (the opened loop conformation requires the pass through the closed loop conformation), (b) the closed loop conformation is necessarily previous to opened loop conformation to allow for sulfur arrangements indispensable for the process, (c) iron can not be donated in the opened loop conformation because the persulfurated cysteine of the Nfs1 loop is blocking this action and is also much closer to the Isu ISC assembly site (See Figure III-7). That is why we propose that iron participation is previous to sulfur one. In general, is to the presence of iron that the process is initiated and many references supporting the idea that ISC biogenesis depends on iron availability have been published over the last years.

Yeast Frataxin tail

Yeast Frataxin has been shown to have a tail connected to the first α-helix that, from our results, seems to be important for the protein behavior. As can be seen in the structure solved by Nuclear Magnetic Resonance (NMR), this tail undergoes from a remarkable rotation movement that covers an angle range of up to 90º (PDB code: 2GA5). The existing hinge between the initial part of the first α-helix of the protein and the already mentioned tail allows these large displacements. This tail has been studied in detail, identifying several negatively charged amino acids in it and certifying that the docking results are altered due to its presence. Owing to these facts, our hypothesis is that the negatively charged amino acids of the tail are helping the first α-helix ones to fulfill the Frataxin iron handling function. Taken together, the tail should be closed in front of the first α-helix of Frataxin to collaborate in the iron binding and to permit a proper interaction with the protein complex (iron-loaded Yfh1 → closed tail → good interaction).

On the contrary, the tail is opened in the absence of iron, thus preventing the interaction with the complex (iron-free Yfh1 → opened tail → bad interaction). Hence, the donation of iron from Yfh1 to Isu for ISC biogenesis might promote the Yfh1 tail opening and, consequently, the expulsion of Yfh1 of the protein complex to then be reloaded again with iron. In our results, dockings with the 20 NMR structures of 2GA5, in which the tail is in the opened state, resulted in bad solutions, while dockings with folded conformations resulted in good solutions.

DISCUSSION:

Structure of the initial ISC biogenesis protein complex and its dynamics

Although ISC biogenesis is a complex process where many proteins take part in different steps: (a) an initial 2Fe-2S cluster assembly, (b) a maturation of this cluster, and (c) its transfer to Apo-proteins, in this paper we have been focusing our studies on the primary and essential protein machinery necessary to assemble two irons and two sulfurs in a premature ISC. As previously stated, in eukaryotes the proteins Frataxin (iron donor), Nfs1 (sulfur donor), Isd11 (essential for Nfs1 activity) and Isu (ISC scaffold) compose the central platform for this machinery [Rawat and Stemmler, 2011] (See Figure III-3 of the Introduction section). Nfs1 is bigger than the other proteins and acts biologically as a dimer. It donates the sulfur though a cysteine desulfurase reaction in Nfs1 that implies the passing of a free cysteine to alanine with a PLP cofactor intervention. The persufuration of a Nfs1 conserved cysteine residue situated in a mobile loop of Nfs1 follows this previous reaction [Selbach et al., 2010]. This big Nfs1 dimer seems to act as the anchorage for all of the other proteins and, therefore, on each monomer of this Nfs1 dimer, one Isu, one Frataxin and one Isd11 interact to generate a big protein complex composed of eight proteins, which constitute the initial ISC assembly machinery [Cook et al., 2010; Prischi et al., 2010] (See the final image of Part C of Figure III-4 of the Materials and Methods section). A crucial part of our hypothesis deals with the idea of an “open-close” Nfs1 dimer alternative conformational change. These conformational changes, together with the Nfs1 loop flexibility, generate two types of conformations in each of the Nfs1 monomers of the dimer: On the one hand, a “totally opened” (TO) conformation, in which both the

monomer and the loop are opened and, on the other hand, a "totally closed" (TC) conformation, where both the monomer and the loop are closed (See Figure III-6 of the Materials and Methods section). According to our results, the proteins Frataxin and Isu interact preferably and better on the Nfs1 TC monomer than on the Nfs1 TO monomer. The area on Nfs1 for Frataxin and Isu interaction is in a region near to the catalytic area of Nfs1 where the PLP cofactor and the cysteine loop are situated. In the TC monomer, this region is more delimited than is the TO monomer and, therefore, this fact enables a proper distance of the iron binding residues of Frataxin and the three cysteine residues that compose the ISC scaffold pocket of Isu to favor iron donation [Correia et al., 2010; Rawat and Stemmler, 2011] (See Part A of Figure III-7 of the Results section). The extension of the Nfs1 cysteine-containing loop, the Nfs1 conformational changes to the opened state, the expulsion of Frataxin both by this loop conformational change and Frataxin tail opening after iron donation, and the surface displacements of Isu towards the Nfs1 loop, allow for the proper distances required for sulfur donation [Maiti et al., 2004; Mansy and Cowan, 2004] (See Part B of Figure III-7 of the Results section). Our docking results also show that in the TO monomer the interacting surface is too wide and the space for the proteins is not as well delimited as before, so Frataxin and Isu tend not to interact in the desired orientation to initiate the ISC assembly cycle. We therefore believe that Frataxin is not interacting with Nfs1 and that Isu is in its initial position in this TO monomer and is in this state when Frataxin undergoes the iron loading procedure. In summary, we propose ISC assembly only in the TC monomer of the Nfs1 dimer, while in the TO monomer the unbound Frataxin is being reloaded with iron and a free cysteine have the required facilities to contact the PLP cofactor (See Figure III-8).

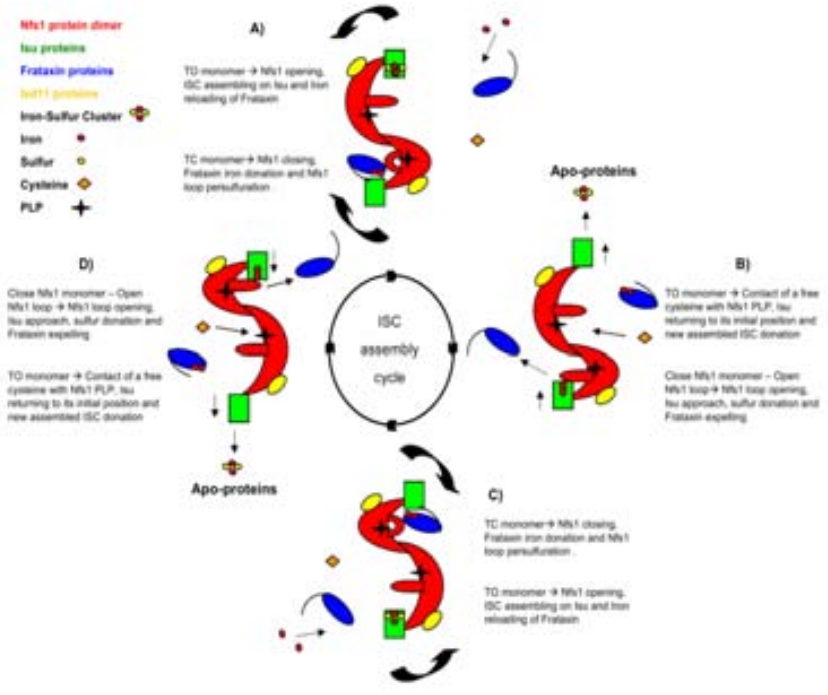


Figure III-8. Dynamics of the iron-sulfur cluster assembly. This is the representation of our proposal for the iron-sulfur cluster (ISC) assembly cycle. The images are mimicking those proteins of the protein complex depicted in the final image of Part C of Figure 2. Along the cycle, the open-close conformational changes of the Nfs1 dimer and its loop are clearly appreciated. In the bottom TC monomer of the Nfs1 dimer: (A) iron-bound Frataxin is very close to the Isu ISC assembly site, enabling for iron recruitment while the cysteine of the Nfs1 loop is persulfurated by the pyridoxal phosphate (PLP) cofactor (black star); (B) Frataxin losses its iron and is being expelled from the complex while the Nfs1 loop, which contains the just persulfurated cysteine, is being opened and Isu undergoes some surface displacements on Nfs1 to get closer to that loop. These facts allow for sulfur donation; (C) free Frataxin is reloaded with iron while the Nfs1 monomomer tends be opened; (D) unbound Frataxin is now iron-loaded again, a free cysteine (orange rhomb) contacts the PLP cofactor of Nfs1 to be desulfurated and form alanine while Isu returns to its initial position with a pre-ISC that will be assembled to a mature ISC and prepared for maturation and donation to Apo-proteins. Then, cycle starts again. The above monomer of the Nfs1 dimer follows the same behavior, but in an opposite state (in this case this explanation should begin in C)
(Amela et al, 2013).

Moreover, this is in agreement with the latest investigations, showing that Frataxin triggers Nfs1 function and its binding stimulates sulfur delivery from Nfs1 to Isu for ISC assemly [Tsai and Barondeau, 2010; Bridwell-Rabb et al., 2011; Tsai et al., 2011]. If Frataxin interacts with the TC monomer of Nfs1, the sequence of events represented in Figure III-8 are initiated, thus allowing for ISC assembly. Taking these premises and the previous subsections into account, a summary of the proposed procedure is as follows (beginning from the TC Nfs1 monomer of Part A of Figure III-8):

1. Frataxin donates iron atoms to Isu while cysteine of the Nfs1 loop is being persulfurated due to the contact with the PLP cofactor in Nfs1.
2. The Nfs1 loop changes to an open conformation while Isu moves towards it. This favors sulfur donation from the already persulfurated cysteine of the Nfs1 loop to Isu. Iron-free Frataxin is expelled from the complex due to its tail opening and also due to the Nfs1 loop opening.
3. The Nfs1 monomer changes to a TO state while unbound Frataxin is being reloaded with iron.
4. A free cysteine contacts the PLP cofactor of Nfs1 to be desulfurated to alanine. Isu returns to its initial position and the new assembled ISC can be matured and given to Apo-proteins.

Although the model is theoretical, based on results obtained from different computational approaches, other models that use similar bioinformatics tools have recently been proposed [Gerber et al., 2003; Maiti et al., 2004]. But these models do not present the novelties of the dynamical behavior of the entire complex to achieve its function as well as the incorporation of the putative structure and function of the protein Isd11, which is explained in a previous paragraph. Moreover, this behavior depends on Nfs1 protein flexibility that is, for the first time,

proposed here. Our protein complex was modeled considering the current PDB crystallographic structures, but not only this. In the already proposed models, which are exclusively based on the available crystallographic structures, the distance between the catalytic Nfs1 cysteine-containing loop and the active ISC biogenesis site is big. Therefore, we think are controversial models from a functional point of view and we propose that of the manuscript in which Nfs1 loop displacements are assisted by Nfs1 protein open-close movements with which the small distances that can better explain the ISC assembly process are achieved. The structure and dynamics of our model emulates an ISC molecular stapler in which Frataxin seems to act as a staple holder. In this type of model if the staple (iron) is not charged in the staple holder (Frataxin), the proper interaction between Frataxin and the stapler (ISC assembly protein complex) will be hampered impeding Nfs1 cysteine loop persulfuration, iron donation, sulfur donation and, thus, ISC generation. The general structure of the complex is feasible according to the computer simulation state of the art and from reported experimental data. As previously mentioned, these bioinformatics approaches give us a basic model with which we can help in the understanding of ISC biogenesis proteins as well as to design new useful experiments. Finally, it must be taken into account that proposing new functional models is important to, at least, validate or refuse it but the absence of models hinders to tackle the complex issue of ISC biogenesis. In addition, the fact of having at least a basic model of Isd11 structure and function and trying to incorporate it to the complex will be helpful for further investigations.

Data supporting the model

In our yeast Open/Close models, Frataxin residues currently proposed to participate in the iron binding and donation, the three Isu cysteine residues that act as the ISC assembly site, the catalytic cysteine residue of Nfs1, and the internal PLP cofactor of Nfs1 are close enough to permit the specific contacts and reactions required for ISC biogenesis. The relevance of those residues and that of the PLP cofactor is shown by Rawat & Stemmler in their paper entitled "*Key players and their role during mitochondrial iron-sulfur cluster biogenesis*", [Rawat and Stemmler, 2011]. Regarding the interaction energies, not only considering the Haddock results but also those from other programs designed for this purpose, like FastContact [Camacho and Zhang, 2005; Champ and Camacho, 2007], we saw that the interaction energy of our proposed protein complex is high enough to explain a dynamic process, and at the same time we could follow how the strength implicated in the interaction changes depending on the considered step of the cluster formation process. That sequence of events allows for the appropriate formation of the cluster. In that sense, Frataxin seems to guide the interaction and starts out the action of the whole complex, which is in agreement with elsewhere reported works [Tsai and Barondeau, 2010; Bridwell-Rabb et al., 2011; Tsai et al., 2011].

The prokaryotic paradox

As recently reported by Yoon et al. [Yoon et al., 2012], the main difference of an eukaryotic form of Isu and a prokaryotic protein seems to rely only on a mutation in residue 107 from Methionine to Isoleucine. This amino acid is surface exposed in he protein and very close to one of the cysteines of the Iron-Sulfur Cluster (ISC) assembly site of Isu. This change avoids the necessity of Frataxin as iron donor for ISC biogenesis in yeast, although many articles suggest that effectively plays an

important role in ISC biogenesis in eukaryotes. Thus, in eukaryotes ISC biogenesis is thought to be a Frataxin-dependent process, Frataxin activates it and the protein Isd11 is present. On the contrary, in prokaryotes ISC biogenesis is a Frataxin-independent process [Yoon et al., 2012], Frataxin inactivates it [Adinolfi et al., 2009] and in this case Isd11 is not present.

Isoleucine is a branched-chain amino acid that might difficult the entry of Frataxin and, therefore, the iron donation by this protein. In this case an alternative iron donor might be recruited in prokaryotes to act as the iron donor for the ISC biogenesis mechanism as pointed out by Yoon et al, and discussed below. On the contrary, a methionine residue in this position is present in eukaryotes where the mechanism requires Frataxin as the ISC assembly iron donor. Here, methionine might not be hampering the iron donation by Frataxin

This previous hypothesis converges with another idea in which iron chelation is relevant. In eukaryotes, the methionine residue might retain the excess of iron outside the active center due to its ability to chelate iron and, thus, Frataxin would be forcing their recruited iron atoms to enter to the active site. This may also serve to synchronize the input of iron that remains from the cluster formation mechanism. In prokaryotes, the isoleucine residue might allow for the free enter of iron atoms because this amino acid does not have chelation properties. Here the concentration of iron around the active site would be smaller without the participation of Frataxin and, moreover, there is not an excess of iron to interfere in the generation of the complex in its active form.

Yoon et al. [Yoon et al. 2012] proposes in their paper that an alternative iron donor might be recruited in prokaryotes to act as the iron donor for the ISC biogenesis mechanism and this is in agreement with the fact that

prokaryotic Frataxin (CyaY) was found as an inhibitor of ISC biogenesis [Adinolfi et al., 2009]. A few years ago, Pastore et al. described a new small protein, termed YfhJ, whose function remains unknown and seems to be a Frataxin-like protein because it has a negatively charged surface, binds iron with low affinity and interacts with IscS (prokaryotic Nfs1) in an iron-dependant manner [Pastore et al., 2006]. There are no physical interaction between YfhJ and CyaY. Considering the structural homology between Yfh1 and YfhJ, the equivalent properties of these two proteins found by Pastore et al. [Pastore et al., 2006], and the similar region where YfhJ docks on the IscS(Nfs1)/IscU(Isu) protein complex (See Figure III-9), YfhJ could perfectly carry out this function in prokaryotes.

Figure III-9. YfhJ docks similar to Yfh1. The figure shows the best pose for the docking between IscS/IscU and YfhJ. The residues of YfhJ implicated in the iron binding and IscS interaction are depicted in blue (Pastore et al, Structure 2006, 14, 857-867). These results suggest the Frataxin-like function of the protein YfhJ as proposed by Pastore et al. (Amela et al. 2013).

Another intriguing hypothesis derived from the use of structural homology services in NCBI, is that we found high structural similarity between our model for the protein Isd11 and a Ferritin monomer of *Escherichia coli* (See Figure III-10).

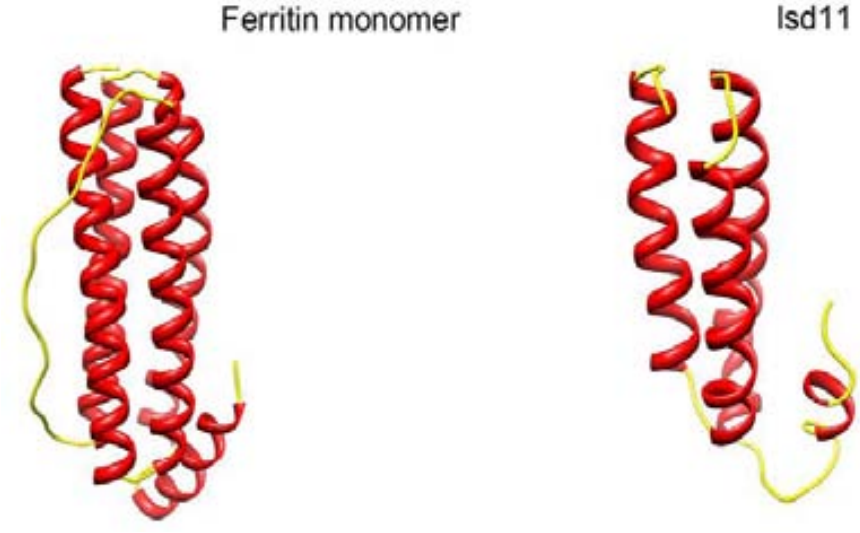


Figure III-10. A Ferritin monomer is homologous to Isd11. By using structural homology searches in NCBI, we found that our model for the protein Isd11 is very similar to a Ferritin monomer of *Escherichia coli*. (http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?uid=1EUM) (Amela et al, 2013).

Therefore, we speculate that a mitochondrial Ferritin monomer in prokaryotes, apart from oligomerize and form large structures to store iron and detoxify the mitochondria, could be acting as Isd11 in eukaryotes, instead of YfhJ. This phenomenon could represent a non-orthologous gene displacement (NOD) during evolution, which describes a variant form of a system in which an expected protein is replaced by a functional equivalent that differs in its evolutionary origin.

In this particular case, Isd11 might evolutionary appear in eukaryotes to do the alternative function proposed for a mitochondrial Ferritin monomer in prokaryotes. This alternative function for a mitochondrial Feritin monomer might partially explain the protective role of mitochondrial Ferritin in FRDA [Campanella et al., 2009].

Further investigations will be needed to confirm these hypotheses.

CONCLUSIONS:

- The sequence, structure, function and interaction of Frataxin, Nfs1 and Isu have been deeply studied. A specific structure and function for the eukaryotic protein Isd11 has been proposed.

- A new dynamic model of the ISC assembly protein complex in yeast as well as the details concerning the iron and sulfur donation to the process have been suggested.

- A speculative hypothesis has been postulated regarding the role of Frataxin in the prokaryotic ISC biogenesis system.

- This approach should help not only in the understanding of the function and molecular properties of the FRDA causing protein (Frataxin) and its protein partners, but also in increasing the knowledge about FRDA being helpful for a possible future treatment of FRDA.

LIST OF FIGURES:

REFERENCES:

1. *The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.*
2. Abdissa, A., et al., *High diversity of group A streptococcal emm types among healthy schoolchildren in Ethiopia.* Clin Infect Dis, 2006. **42**(10): p. 1362-7.
3. Adam, A.C., et al., *The Nfs1 interacting protein Isd11 has an essential role in Fe/S cluster biogenesis in mitochondria.* EMBO J, 2006. **25**(1): p. 174-83.
4. Adamou, J.E., et al., *Identification and characterization of a novel family of pneumococcal proteins that are protective against sepsis.* Infect Immun, 2001. **69**(2): p. 949-58.
5. Adinolfi, S., et al., *Bacterial frataxin CyaY is the gatekeeper of iron-sulfur cluster formation catalyzed by IscS.* Nat Struct Mol Biol, 2009. **16**(4): p. 390-6.
6. Adinolfi, S., et al., *A structural approach to understanding the iron-binding properties of phylogenetically different frataxins.* Hum Mol Genet, 2002. **11**(16): p. 1865-77.
7. Agar, J.N., et al., *IscU as a scaffold for iron-sulfur cluster biosynthesis: sequential assembly of [2Fe-2S] and [4Fe-4S] clusters in IscU.* Biochemistry, 2000. **39**(27): p. 7856-62.
8. Aguilar, D., et al., *TranScout: prediction of gene expression regulatory proteins from their sequences.* Bioinformatics, 2002. **18**(4): p. 597-607.
9. Ala'Aldeen, D.A. and S.P. Borriello, *The meningococcal transferrin-binding proteins 1 and 2 are both surface exposed and generate bactericidal antibodies capable of killing homologous and heterologous strains.* Vaccine, 1996. **14**(1): p. 49-53.
10. Ala'Aldeen, D.A., et al., *Immune responses in humans and animals to meningococcal transferrin-binding proteins: implications for vaccine design.* Infect Immun, 1994. **62**(7): p. 2984-90.
11. Aloria, K., et al., *Iron-induced oligomerization of yeast frataxin homologue Yfh1 is dispensable in vivo.* EMBO Reports, 2004. **5**(11): p. 1096-1101.
12. Aloy, P., et al., *'TransMem': a neural network implemented in Excel spreadsheets for predicting transmembrane domains of proteins.* Comput Appl Biosci, 1997. **13**(3): p. 231-4.
13. Altschul, S., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucl. Acids Res., 1997. **25**(17): p. 3389-3402.
14. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

15. Amela, I., et al., *Pathogen proteins eliciting antibodies do not share epitopes with host proteins: a bioinformatics approach.* PLoS ONE, 2007. **2**(6): p. e512.
16. Amela, I., et al., *DockAnalyse: an application for the analysis of protein-protein interactions.* BMC Struct Biol, 2010. **10**(37).
17. Amela, I., et al., *A dynamic model of the proteins that form the initial iron-sulfur cluster biogenesis machinery in yeast mitochondria.* Protein J, 2013. **32**(3): p. 183-96.
18. Armstrong, J.S., et al., *Does oxidative stress contribute to the pathology of Friedreich's ataxia? A radical question.* FASEB J, 2010. **24**(7): p. 2152-63.
19. Askelof, P., et al., *Protective immunogenicity of two synthetic peptides selected from the amino acid sequence of Bordetella pertussis toxin subunit S1.* Proc Natl Acad Sci U S A, 1990. **87**(4): p. 1347-51.
20. Ausiello, G., et al., *ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure.* Proteins, 1997. **28**(4): p. 556-67.
21. Banks, D.J., et al., *Progress toward characterization of the group A Streptococcus metagenome: complete genome sequence of a macrolide-resistant serotype M6 strain.* J Infect Dis, 2004. **190**(4): p. 727-38.
22. Bannantine, J.P., et al., *A secondary structure motif predictive of protein localization to the chlamydial inclusion membrane.* Cell Microbiol, 2000. **2**(1): p. 35-47.
23. Barras, F.d.r., et al., *How Escherichia coli and Saccharomyces cerevisiae build Fe/S proteins.* Adv Microb Physiol, 2005. **50**: p. 41-101.
24. Barrientos, A., *Yeast models of human mitochondrial diseases.* IUBMB Life, 2003. **55**(2): p. 83-95.
25. Beghetto, E., et al., *Discovery of novel Streptococcus pneumoniae antigens by screening a whole-genome lambda-display library.* FEMS Microbiol Lett, 2006. **262**(1): p. 14-21.
26. Bencze, K.Z., et al., *The structure and function of frataxin.* Crit Rev Biochem Mol Biol, 2006. **41**(5): p. 269-91.
27. Bencze, K.Z., et al., *Human frataxin: iron and ferrochelatase binding surface.* Chem Commun (Camb), 2007(18): p. 1798-800.
28. Bendtsen, J.D., et al., *Improved prediction of signal peptides: SignalP 3.0.* J Mol Biol, 2004. **340**(4): p. 783-95.
29. Benner, G.E., et al., *Immune response to Yersinia outer proteins and other Yersinia pestis antigens after experimental plague infection in mice.* Infect Immun, 1999. **67**(4): p. 1922-8.
30. Bennett-Lovsey, R.M., et al., *Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre.* Proteins, 2008. **70**(3): p. 611-625.

31. Benoist, C. and D. Mathis, *Autoimmunity provoked by infection: how good is the case for T cell epitope mimicry?* Nat Immunol, 2001. **2**(9): p. 797-801.
32. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
33. Biederbick, A., et al., *Role of Human Mitochondrial Nfs1 in Cytosolic Iron-Sulfur Protein Biogenesis and Iron Regulation.* Mol Cell Biol, 2006. **26**(15): p. 5675-87.
34. Birkeland, H.C. and H. Stenmark, *Protein targeting to endosomes and phagosomes via FYVE and PX domains.* Curr Top Microbiol Immunol, 2004. **282**: p. 89-115.
35. Borchardt, J.K., *The history of bacterial meningitis treatment.* Drug News Perspect, 2004. **17**(3): p. 219-24.
36. 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.
37. Branda, S.S., et al., *Yeast and human frataxin are processed to mature form in two sequential steps by the mitochondrial processing peptidase.* J Biol Chem, 1999. **274**(32): p. 22763-9.
38. Breitkreutz, B.J., et al., *The GRID: the General Repository for Interaction Datasets.* Genome Biol, 2003. **4**(3): p. R23.
39. Bridwell-Rabb, J., et al., *Effector role reversal during evolution: the case of frataxin in Fe-S cluster biosynthesis.* Biochemistry, 2012. **51**(12): p. 2506-14.
40. Bridwell-Rabb, J., et al., *Structure-function analysis of Friedreich's ataxia mutants reveals determinants of frataxin binding and activation of the Fe-S assembly complex.* Biochemistry, 2011. **50**(33): p. 7265-74.
41. Bulteau, A.L., et al., *Frataxin acts as an iron chaperone protein to modulate mitochondrial aconitase activity.* Science, 2004. **305**(5681): p. 242-5.
42. Bulteau, A.L., et al., *Changes in mitochondrial glutathione levels and protein thiol oxidation in âˆ†yfh1 yeast cells and the lymphoblasts of patients with Friedreich's ataxia.* Biochim Et Biophys Acta, 2012. **1822**(2): p. 212-225.
43. Busi, M.V. and D.F. Gomez-Casati, *Exploring frataxin function.* IUBMB Life, 2011. **64**(1): p. 56-63.
44. Camacho, C.J. and C. Zhang, *FastContact: rapid estimate of contact and binding free energies.* Bioinformatics (Oxford, England), 2005. **21**(10): p. 2534-6.
45. Campanella, A., et al., *Mitochondrial ferritin limits oxidative damage regulating mitochondrial iron availability: hypothesis for a protective role in Friedreich ataxia.* Hum Mol Genet, 2009. **18**(1): p. 1-11.

46. Campuzano, V., et al., *Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion.* Science (New York, N.Y.), 1996. **271**(5254): p. 1423-7.
47. Cazalet, C., et al., *Evidence in the Legionella pneumophila genome for exploitation of host cell functions and high genome plasticity.* Nat Genet, 2004. **36**(11): p. 1165-73.
48. Cedano, J., et al., *Relation between amino acid composition and cellular location of proteins.* J Mol Biol, 1997. **266**(3): p. 594-600.
49. Champ, P.C. and C.J. Camacho, *FastContact: a free energy scoring tool for protein-protein complex structures.* Nucl. Acids Res., 2007. **35**(Web Server issue): p. W556-560.
50. Chandramouli, K., et al., *Formation and properties of [4Fe-4S] clusters on the IscU scaffold protein.* Biochemistry, 2007. **46**(23): p. 6804-11.
51. Chang, B., et al., *Identification of a novel adhesion molecule involved in the virulence of Legionella pneumophila.* Infect Immun, 2005. **73**(7): p. 4272-80.
52. Chatenoud, L., *[Immunotolerance and autoimmunity].* Rev Prat, 2000. **50**(13): p. 1497-505.
53. Chen, O.S., et al., *Inhibition of Fe-S cluster biosynthesis decreases mitochondrial iron export: evidence that Yfh1p affects Fe-S cluster synthesis.* Proc Natl Acad Sci U S A, 2002. **99**(19): p. 12321-6.
54. Chen, X.W., et al., *Protein function assignment through mining cross-species protein-protein interactions.* PLoS ONE, 2008. **3**(2): p. e1562.
55. Chenna, R., et al., *Multiple sequence alignment with the Clustal series of programs.* Nucl. Acids Res., 2003. **31**(13): p. 3497-3500.
56. Chiavolini, D., et al., *The three extra-cellular zinc metalloproteinases of Streptococcus pneumoniae have a different impact on virulence in mice.* BMC Microbiol, 2003. **3**: p. 14.
57. Chivian, D., et al., *Automated prediction of CASP-5 structures using the Robetta server.* Proteins, 2003. **53 Suppl 6**: p. 524-33.
58. Chivian, D., et al., *Prediction of CASP6 structures using automated Robetta protocols.* Proteins, 2005. **61 Suppl 7**: p. 157-66.
59. Chua, H.N., et al., *Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions.* Bioinformatics, 2006. **22**(13): p. 1623-30.
60. Claros, M.G. and P. Vincens, *Computational method to predict mitochondrially imported proteins and their targeting sequences.* Eur J Biochem, 1996. **241**(3): p. 779-86.

61. 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.
62. Cook, J.D., et al., *Monomeric yeast frataxin is an iron-binding protein.* Biochemistry, 2006. **45**(25): p. 7767-77.
63. Cook, J.D., et al., *Molecular details of the yeast frataxin-Isu1 interaction during mitochondrial Fe-S cluster assembly.* Biochemistry, 2010. **49**(40): p. 8756-65.
64. Corech, R., et al., *Early immune response to the components of the type III system of Pseudomonas aeruginosa in children with cystic fibrosis.* J Clin Microbiol, 2005. **43**(8): p. 3956-62.
65. Correia, A.R., et al., *Conformational stability of human frataxin and effect of Friedreich's ataxia-related mutations on protein folding.* Biochem J, 2006. **398**(3): p. 605-11.
66. Correia, A.R., et al., *Iron-binding activity in yeast frataxin entails a trade off with stability in the alpha1/beta1 acidic ridge region.* Biochem J, 2010. **426**(2): p. 197-203.
67. Crooks, D.R., et al., *Posttranslational stability of the heme biosynthetic enzyme ferrochelatase is dependent on iron availability and intact iron-sulfur cluster assembly machinery.* Blood, 2010. **115**(4): p. 860-69.
68. de Alava, E., et al., *Adenovirus E1A and Ewing tumors.* Nat Med, 2000. **6**(1): p. 4.
69. de Vries, S.J., et al., *HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets.* Proteins, 2007. **69**(4): p. 726-33.
70. de Vries, S.J., et al., *The HADDOCK web server for data-driven biomolecular docking.* Nat Protoc, 2010. **5**(5): p. 883-97.
71. Delatycki, M.B., et al., *Direct evidence that mitochondrial iron accumulation occurs in Friedreich ataxia.* Ann Neurol, 1999. **45**(5): p. 673-5.
72. Dhe-Paganon, S., et al., *Crystal structure of human frataxin.* J Biol Chem, 2000. **275**(40): p. 30753-56.
73. Dias, W.O., et al., *A Bordetella pertussis acellular vaccine candidate: antigenic characterization and antibody induction.* Braz J Med Biol Res, 1994. **27**(11): p. 2607-11.
74. Dominguez, C., et al., *HADDOCK: a protein-protein docking approach based on biochemical or biophysical information.* J Am Chem Soc, 2003. **125**(7): p. 1731-7.
75. Duby, G., et al., *A non-essential function for yeast frataxin in iron-sulfur cluster assembly.* Hum Mol Genet, 2002. **11**(21): p. 2635-43.

76. Duchene, M., et al., *Pseudomonas aeruginosa outer membrane lipoprotein I gene: molecular cloning, sequence, and expression in Escherichia coli.* J Bacteriol, 1989. **171**(8): p. 4130-7.
77. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucl. Acids Res., 2004. **32**(5): p. 1792-7.
78. Ekins, S., et al., *In silico repositioning of approved drugs for rare and neglected diseases.* Drug Discov Today, 2011. **16**(7-8): p. 298-310.
79. El-Adhami, W., et al., *Characterization of the gene encoding a 26-kilodalton protein (OMP26) from nontypeable Haemophilus influenzae and immune responses to the recombinant protein.* Infect Immun, 1999. **67**(4): p. 1935-42.
80. El-Manzalawy, Y., et al., *Predicting linear B-cell epitopes using string kernels.* J Mol Recognit, 2008. **21**(4): p. 243-55.
81. Emekli, U., et al., *HingeProt: automated prediction of hinges in protein structures.* Proteins, 2008. **70**(4): p. 1219-27.
82. Erkkila, L., et al., *Heat shock protein 60 autoimmunity and early lipid lesions in cholesterol-fed C57BL/6JBom mice during Chlamydia pneumoniae infection.* Atherosclerosis, 2004. **177**(2): p. 321-8.
83. Espadaler, J., et al., *Prediction of enzyme function by combining sequence similarity and protein interactions.* BMC Bioinformatics, 2008. **9**: p. 249.
84. Espinos-Armero, C., et al., *[Autosomal recessive cerebellar ataxias. Their classification, genetic features and pathophysiology].* Rev Neurol, 2005. **41**(7): p. 409-22.
85. 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.
86. Fields, S. and O.-k. Song, *A novel genetic system to detect protein–protein interactions.* Nature, 1989. **340**(6230): p. 245-6.
87. Finco, O., et al., *Identification of new potential vaccine candidates against Chlamydia pneumoniae by multiple screenings.* Vaccine, 2005. **23**(9): p. 1178-88.
88. Fischer, D., et al., *CAFASP-1: critical assessment of fully automated structure prediction methods.* Proteins, 1999. **Suppl 3**: p. 209-17.
89. Fiser, A. and A. Sali, *Modeller: generation and refinement of homology-based protein structure models.* Methods Enzymol, 2003. **374**: p. 461-91.
90. Fitzpatrick, D.A., et al., *Evidence of positive Darwinian selection in putative meningococcal vaccine antigens.* J Mol Evol, 2005. **61**(1): p. 90-8.
91. Foury, F., et al., *Acidic residues of yeast frataxin have an essential role in Fe-S cluster assembly.* EMBO Rep, 2007. **8**(2): p. 194-9.

92. Frazzon, J., et al., *Biosynthesis of iron-sulphur clusters is a complex and highly conserved process.* Biochem Soc Trans, 2002. **30**(4): p. 680-5.
93. Gabow, A.P., et al., *Improving protein function prediction methods with integrated literature data.* BMC Bioinformatics, 2008. **9**: p. 198.
94. Gakh, O., et al., *Normal and Friedreich ataxia cells express different isoforms of frataxin with complementary roles in iron-sulfur cluster assembly.* J Biol Chem, 2010. **285**(49): p. 38486-501.
95. Gakh, O., et al., *Assembly of the iron-binding protein frataxin in S. cerevisiae responds to dynamic changes in mitochondrial iron influx and stress level.* J Biol Chem, 2008. **283**(46): p. 31500-10.
96. Gasteiger, E., et al., *ExPASy: The proteomics server for in-depth protein knowledge and analysis.* Nucleic Acids Res, 2003. **31**(13): p. 3784-8.
97. Gattiker, A., et al., *Automated annotation of microbial proteomes in SWISS-PROT.* Comput Biol Chem, 2003. **27**(1): p. 49-58.
98. Gautam, A.M., et al., *A polyalanine peptide with only five native myelin basic protein residues induces autoimmune encephalomyelitis.* J Exp Med, 1992. **176**(2): p. 605-9.
99. 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.
100. Gerber, J. and R. Lill, *Biogenesis of iron-sulfur proteins in eukaryotes: components, mechanism and pathology.* Mitochondrion, 2002. **2**(1-2): p. 71-86.
101. Gerber, J., et al., *An interaction between frataxin and Isu1/Nfs1 that is crucial for Fe/S cluster synthesis on Isu1.* EMBO Rep, 2003. **4**(9): p. 906-11.
102. 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.
103. Gibrat, J.F., et al., *Surprising similarities in structure comparison.* Curr Opin Struct Biol, 1996. **6**(3): p. 377-85.
104. Gilbert, D., *Biomolecular interaction network database.* Brief Bioinform, 2005. **6**(2): p. 194-8.
105. Goldschneider, I., et al., *Human immunity to the meningococcus. I. The role of humoral antibodies.* J Exp Med, 1969. **129**(6): p. 1307-26.
106. Gómez, A., et al., *Prediction of protein function improving sequence remote alignment search by a fuzzy logic algorithm.* Protein J, 2008. **27**(2): p. 130-9.

107. Gómez, A., et al., *Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins?* Bioinformatics, 2003. **19**(7): p. 895-6.
108. González-Cabo, P., et al., *Friedreich ataxia: an update on animal models, frataxin function and therapies.* Adv Exp Med Biol, 2009. **652**: p. 247-61.
109. González-Cabo, P., et al., *Frataxin interacts functionally with mitochondrial electron transport chain proteins.* Hum Mol Genet, 2005. **14**(15): p. 2091-8.
110. Goure, J., et al., *Protective anti-V antibodies inhibit Pseudomonas and Yersinia translocon assembly within host membranes.* J Infect Dis, 2005. **192**(2): p. 218-25.
111. Guex, N. and M.C. Peitsch, *SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.* Electrophoresis, 1997. **18**(15): p. 2714-23.
112. Guldener, U., et al., *MPact: the MIPS protein interaction resource on yeast.* Nucleic Acids Res, 2006. **34**(Database issue): p. D436-41.
113. Guzman, C.A., et al., *Protective immune response against Streptococcus pyogenes in mice after intranasal vaccination with the fibronectin-binding protein SfbI.* J Infect Dis, 1999. **179**(4): p. 901-6.
114. 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.
115. Hausman, S.Z. and D.L. Burns, *Use of pertussis toxin encoded by ptx genes from Bordetella bronchiseptica to model the effects of antigenic drift of pertussis toxin on antibody neutralization.* Infect Immun, 2000. **68**(6): p. 3763-7.
116. He, Y., et al., *Yeast frataxin solution structure, iron binding, and ferrochelatase interaction.* Biochemistry, 2004. **43**(51): p. 16254-62.
117. Hermjakob, H., et al., *IntAct: an open source molecular interaction database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D452-5.
118. Higgins, D.G. and P.M. Sharp, *CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.* Gene, 1988. **73**(1): p. 237-44.
119. Huang, J. and J.A. Cowan, *Iron-sulfur cluster biosynthesis: role of a semi-conserved histidine.* Chem Commun (Camb), 2009(21): p. 3071-3.
120. Huang, M.L.-H., et al., *Elucidation of the mechanism of mitochondrial iron loading in Friedreich's ataxia by analysis of a mouse mutant.* Proc Natl Acad Sci U S A, 2009. **106**(38): p. 16381-6.

121. Hwang, H., et al., *Protein-protein docking benchmark version 3.0.* Proteins, 2008. **73**(3): p. 705-9.
122. Iannuzzi, C., et al., *The Role of CyaY in Iron Sulfur Cluster Assembly on the E. coli IscU Scaffold Protein.* PLoS ONE, 2011. **6**(7): p. e21992.
123. 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.
124. Jacobson, D.L., et al., *Epidemiology and estimated population burden of selected autoimmune diseases in the United States.* Clin Immunol Immunopathol, 1997. **84**(3): p. 223-43.
125. Jaeger, S., et al., *Integrating protein-protein interactions and text mining for protein function prediction.* BMC Bioinformatics, 2008. **9 Suppl 8**: p. S2.
126. Janin, J., *Protein-protein docking tested in blind predictions: the CAPRI experiment.* Mol Biosyst, 2010. **6**(12): p. 2351-62.
127. Janson, H., et al., *Limited diversity of the protein D gene (hpd) among encapsulated and nonencapsulated Haemophilus influenzae strains.* Infect Immun, 1993. **61**(11): p. 4546-52.
128. Johnson, D.C., et al., *Structure, function, and formation of biological iron-sulfur clusters.* Annu Rev Biochem, 2005. **74**: p. 247-81.
129. Kanduc, D., *Epitopic peptides with low similarity to the host proteome: towards biological therapies without side effects.* Expert Opin Biol Ther, 2009. **9**(1): p. 45-53.
130. Karlberg, T., et al., *The Structures of Frataxin Oligomers Reveal the Mechanism for the Delivery and Detoxification of Iron.* Structure, 2006. **14**(10): p. 1535-46.
131. Kaut, A., et al., *Isa1p is a component of the mitochondrial machinery for maturation of cellular iron-sulfur proteins and requires conserved cysteine residues for function.* J Biol Chem, 2000. **275**(21): p. 15955-61.
132. Keil, D.J., et al., *Cloning and immunologic characterization of a truncated Bordetella bronchiseptica filamentous hemagglutinin fusion protein.* Vaccine, 1999. **18**(9-10): p. 860-7.
133. Kelley, L.A., et al., *Enhanced genome annotation using structural profiles in the program 3D-PSSM.* J Mol Biol, 2000. **299**(2): p. 499-520.
134. Kim, D.E., et al., *Protein structure prediction and analysis using the Robetta server.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W526-31.
135. Kim, J.H., et al., *Disordered form of the scaffold protein IscU is the substrate for iron-sulfur cluster assembly on cysteine desulfurase.* Proc Natl Acad Sci U S A, 2011. **109**(2): p. 454-9.

136. Kim, J.-H., et al., *Dynamics of Protein Damage in Yeast Frataxin Mutant Exposed to Oxidative Stress.* OMICS, 2010. **14**(6): p. 689-99.
137. Knight, S.A., et al., *The yeast connection to Friedreich ataxia.* Am J Hum Genet, 1999. **64**(2): p. 365-71.
138. Koeppen, A.H., *Friedreich's ataxia: Pathology, pathogenesis, and molecular genetics.* J Neurol Sci, 2011. **303**(1-2): p. 1-12.
139. Kolaskar, A.S. and P.C. Tongaonkar, *A semi-empirical method for prediction of antigenic determinants on protein antigens.* FEBS Lett, 1990. **276**(1-2): p. 172-4.
140. Krebs, W.G. and M. Gerstein, *The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework.* Nucl. Acids Res., 2000. **28**(8): p. 1665-75.
141. Kuntal, B.K., et al., *EasyModeller: A graphical interface to MODELLER.* BMC Res Notes, 2010. **3**(226): p. doi: 10.1186/1756-0500-3-226.
142. Lambert, C., et al., *ESyPred3D: Prediction of proteins 3D structures.* Bioinformatics, 2002. **18**(9): p. 1250-6.
143. Lane, D.J.R. and D.R. Richardson, *Frataxin, a molecule of mystery: trading stability for function in its iron-binding site.* Biochem J, 2010. **426**(2): p. e1-3.
144. Lange, H., et al., *A mitochondrial ferredoxin is essential for biogenesis of cellular iron-sulfur proteins.* Proc Natl Acad Sci U S A, 2000. **97**(3): p. 1050-5.
145. Leen, A., et al., *Differential immunogenicity of Epstein-Barr virus latent-cycle proteins for human CD4(+) T-helper 1 responses.* J Virol, 2001. **75**(18): p. 8649-59.
146. Leidgens, S., et al., *Frataxin interacts with Isu1 through a conserved tryptophan in its {beta}-sheet.* Hum Mol Genet, 2010. **19**(2): p. 276-86.
147. Levi, S. and E. Rovida, *The role of iron in mitochondrial function.* Biochim Biophys Acta, 2009. **1790**(7): p. 629-36.
148. Li, B., et al., *Protein microarray for profiling antibody responses to Yersinia pestis live vaccine.* Infect Immun, 2005. **73**(6): p. 3734-9.
149. Li, H., et al., *Oligomeric yeast frataxin drives assembly of core machinery for mitochondrial iron-sulfur cluster synthesis.* J Biol Chem, 2009. **284**(33): p. 21971-80.
150. Lill, R., *Function and biogenesis of iron-sulphur proteins.* Nature, 2009. **460**(7257): p. 831-8.
151. Lill, R., et al., *Mechanisms of iron-sulfur protein maturation in mitochondria, cytosol and nucleus of eukaryotes.* Biochim Biophys Acta, 2006. **1763**(7): p. 652-67.
152. Lill, R., et al., *The role of mitochondria in cellular iron-sulfur protein biogenesis and iron metabolism.* Biochim Biophys Acta, 2012. **1823**(9): p. 1491-508.

153. Lill, R. and U. Mühlenhoff, *Iron-sulfur-protein biogenesis in eukaryotes.* Trends Biochem Sci, 2005. **30**(3): p. 133-41.
154. Lill, R. and U. Mühlenhoff, *Iron-sulfur protein biogenesis in eukaryotes: components and mechanisms.* Annu Rev Cell Dev Biol, 2006. **22**: p. 457-86.
155. Lill, R. and U. Mühlenhoff, *Maturation of iron-sulfur proteins in eukaryotes: mechanisms, connected processes, and diseases.* Annu Rev Biochem, 2008. **77**: p. 669-700.
156. Lindmo, K. and H. Stenmark, *Regulation of membrane traffic by phosphoinositide 3-kinases.* J Cell Sci, 2006. **119**(Pt 4): p. 605-14.
157. Liu, D.F., et al., *The C-terminal fragment of the internal 110-kilodalton passenger domain of the Hap protein of nontypeable Haemophilus influenzae is a potential vaccine candidate.* Infect Immun, 2004. **72**(12): p. 6961-8.
158. Lucchese, G., et al., *Peptidology: short amino acid modules in cell biology and immunology.* Amino Acids, 2007. **33**(4): p. 703-7.
159. Lugtenberg, B., et al., *Biochemical and immunological characterization of cell surface proteins of Pasteurella multocida strains causing atrophic rhinitis in swine.* Infect Immun, 1986. **52**(1): p. 175-82.
160. Maione, D., et al., *Identification of a universal Group B streptococcus vaccine by multiple genome screen.* Science, 2005. **309**(5731): p. 148-50.
161. Maiti, R., et al., *SuperPose: a simple server for sophisticated structural superposition.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W590-W4.
162. Mansy, S.S. and J.A. Cowan, *Iron-sulfur cluster biosynthesis: toward an understanding of cellular machinery and molecular mechanism.* Acc Chem Res, 2004. **37**(9): p. 719-25.
163. Marmolino, D., *Friedreich's ataxia: Past, present and future.* Brain Res Rev, 2011. **67**(1-2): p. 311-30.
164. Martelli, A., et al., *Understanding the genetic and molecular pathogenesis of Friedreich's ataxia through animal and cellular models.* Dis Model Mech, 2012. **5**(2): p. 165-76.
165. Martelli, A., et al., *Frataxin is essential for extramitochondrial Fe-S cluster proteins in mammalian tissues.* Hum Mol Genet, 2007. **16**(22): p. 2651-8.
166. Mascarell, L., et al., *Induction of neutralizing antibodies and Th1-polarized and CD4-independent CD8+ T-cell responses following delivery of human immunodeficiency virus type 1 Tat protein by recombinant adenylate cyclase of Bordetella pertussis.* J Virol, 2005. **79**(15): p. 9872-84.
167. Mattoo, S., et al., *Role of Bordetella bronchiseptica fimbriae in tracheal colonization and development of a humoral immune response.* Infect Immun, 2000. **68**(4): p. 2024-33.

168. Matzinger, P., *The danger model: a renewed sense of self.* Science, 2002. **296**(5566): p. 301-5.
169. McArthur, J.D. and M.J. Walker, *Domains of group A streptococcal M protein that confer resistance to phagocytosis, opsonization and protection: implications for vaccine development.* Mol Microbiol, 2006. **59**(1): p. 1-4.
170. McGuffin, L.J., et al., *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.
171. McMillan, D.J., et al., *Immune response to superoxide dismutase in group A streptococcal infection.* FEMS Immunol Med Microbiol, 2004. **40**(3): p. 249-56.
172. Miao, R., et al., *Biophysical Characterization of the Iron in Mitochondria from Atm1p-Depleted Saccharomyces cerevisiae.* Biochemistry, 2009. **48**(40): p. 9556-68.
173. Mishra, G.R., et al., *Human protein reference database--2006 update.* Nucleic Acids Res, 2006. **34**(Database issue): p. D411-4.
174. Miyamoto, Y., et al., *Dock6, a Dock-C subfamily guanine nucleotide exchanger, has the dual specificity for Rac1 and Cdc42 and regulates neurite outgrowth.* Exp Cell Res, 2007. **313**(4): p. 791-804.
175. Montigiani, S., et al., *Genomic approach for analysis of surface proteins in Chlamydia pneumoniae.* Infect Immun, 2002. **70**(1): p. 368-79.
176. Moreno-Cermeno, A., et al., *Frataxin depletion in yeast triggers upregulation of iron transport systems before affecting iron-sulfur enzyme activities.* J Biol Chem, 2010. **285**(53): p. 41653-64.
177. Moss, J., et al., *Sera from adult patients with cystic fibrosis contain antibodies to Pseudomonas aeruginosa type III apparatus.* Infect Immun, 2001. **69**(2): p. 1185-8.
178. Moyle, P.M., et al., *Method for the synthesis of multi-epitopic Streptococcus pyogenes lipopeptide vaccines using native chemical ligation.* J Org Chem, 2006. **71**(18): p. 6846-50.
179. Mühlenhoff, U., et al., *Functional characterization of the eukaryotic cysteine desulfurase Nfs1p from Saccharomyces cerevisiae.* J Biol Chem, 2004. **279**(35): p. 36906-15.
180. Mühlenhoff, 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.
181. Mühlenhoff, U. and R. Lill, *Biogenesis of iron-sulfur proteins in eukaryotes: a novel task of mitochondria that is inherited from bacteria.* Biochim Biophys Acta, 2000. **1459**(2-3): p. 370-82.
182. Mühlenhoff, U., et al., *Characterization of iron-sulfur protein assembly in isolated mitochondria. A requirement for ATP, NADH, and reduced iron.* J Biol Chem, 2002. **277**(33): p. 29810-6.

183. Mühlenhoff, U., et al., *The yeast frataxin homolog Yfh1p plays a specific role in the maturation of cellular Fe/S proteins.* Hum Mol Genet, 2002. **11**(17): p. 2025-36.
184. Mühlenhoff, U., et al., *Specialized Function of Yeast Isa1 and Isa2 Proteins in the Maturation of Mitochondrial [4Fe-4S] Proteins.* J Biol Chem, 2011. **286**(48): p. 41205-16.
185. Murphy, T.F., et al., *Construction of a mutant and characterization of the role of the vaccine antigen P6 in outer membrane integrity of nontypeable Haemophilus influenzae.* Infect Immun, 2006. **74**(9): p. 5169-76.
186. Neuvirth, H., et al., *ProMate: a structure based prediction program to identify the location of protein-protein binding sites.* J Mol Biol, 2004. **338**(1): p. 181-99.
187. Nguyen, L. and J. Pieters, *The Trojan horse: survival tactics of pathogenic mycobacteria in macrophages.* Trends Cell Biol, 2005. **15**(5): p. 269-76.
188. Notredame, C.d., et al., *T-coffee: a novel method for fast and accurate multiple sequence alignment.* J Mol Biol, 2000. **302**(1): p. 205-17.
189. Novotny, P., et al., *Biologic and protective properties of the 69-kDa outer membrane protein of Bordetella pertussis: a novel formulation for an acellular pertussis vaccine.* J Infect Dis, 1991. **164**(1): p. 114-22.
190. Novotny, P., et al., *A novel bivalent acellular pertussis vaccine based on the 69 kDa protein and FHA.* Dev Biol Stand, 1991. **73**: p. 243-9.
191. Okamoto, S., et al., *Systemic immunization with streptococcal immunoglobulin-binding protein Sib 35 induces protective immunity against group: a Streptococcus challenge in mice.* Vaccine, 2005. **23**(40): p. 4852-9.
192. Oldstone, M.B., *Molecular mimicry and immune-mediated diseases.* FASEB J, 1998. **12**(13): p. 1255-65.
193. Oldstone, M.B.A., *Molecular mimicry, microbial infection, and autoimmune disease: evolution of the concept.* Curr Top Microbiol Immunol, 2005. **296**: p. 1-17.
194. Pache, R.A., et al., *Towards a molecular characterisation of pathological pathways.* FEBS Lett, 2008. **582**(8): p. 1259-65.
195. Palma, P.N., et al., *BiGGER: a new (soft) docking algorithm for predicting protein interactions.* Proteins, 2000. **39**(4): p. 372-84.
196. Pandolfo, M., *Molecular genetics and pathogenesis of Friedreich ataxia.* Neuromuscul Disord, 1998. **8**(6): p. 409-15.
197. Pandolfo, M., *Molecular pathogenesis of Friedreich ataxia.* Arch Neurol, 1999. **56**(10): p. 1201-8.
198. Pandolfo, M., *Friedreich's ataxia: clinical aspects and pathogenesis.* Semin Neurol, 1999. **19**(3): p. 311-21.

199. Pandolfo, M., *Friedreich ataxia.* Arch Neurol, 2008. **65**(10): p. 1296-303.
200. Pandolfo, M., *Friedreich ataxia: the clinical picture.* J Neurol, 2009. **256 Suppl 1**: p. 3-8.
201. Pandolfo, M. and A. Pastore, *The pathogenesis of Friedreich ataxia and the structure and function of frataxin.* J Neurol, 2009. **256 Suppl 1**: p. 9-17.
202. Paris, Z.k., et al., *The Fe/S Cluster Assembly Protein Isd11 Is Essential for tRNA Thiolation in Trypanosoma brucei.* J Biol Chem, 2010. **285**(29): p. 22394-402.
203. Pastore, C., et al., *YfhJ, a molecular adaptor in iron-sulfur cluster formation or a frataxin-like protein?* Structure, 2006. **14**(5): p. 857-67.
204. Pedretti, A., et al., *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.
205. Pedretti, A., et al., *Atom-type description language: a universal language to recognize atom types implemented in the VEGA program.* Theor Chem Acc, 2003. **109**(4): p. 229-32.
206. Pedretti, A., et al., *VEGA - An open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming.* J C A M D, 2004. **18**(3): p. 167-73.
207. Peetermans, W.E. and P. Lacante, *Pneumococcal vaccination by general practitioners: an evaluation of current practice.* Vaccine, 1999. **18**(7-8): p. 612-7.
208. Peters, B., et al., *The immune epitope database and analysis resource: from vision to blueprint.* PLoS Biol, 2005. **3**(3): p. e91.
209. Pettersen, E.F., et al., *UCSF Chimera-a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.
210. Philipovskiy, A.V., et al., *Antibody against V antigen prevents Yop-dependent growth of Yersinia pestis.* Infect Immun, 2005. **73**(3): p. 1532-42.
211. Pilon, M., et al., *Biogenesis of iron-sulfur cluster proteins in plastids.* Genet Eng (N Y), 2006. **27**: p. 101-17.
212. Pizza, M., et al., *Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing.* Science, 2000. **287**(5459): p. 1816-20.
213. Poolman, J.T., *Development of a meningococcal vaccine.* Infect Agents Dis, 1995. **4**(1): p. 13-28.
214. Price, B.M., et al., *Protection against Pseudomonas aeruginosa chronic lung infection in mice by genetic immunization against outer membrane protein F (OprF) of P. aeruginosa.* Infect Immun, 2001. **69**(5): p. 3510-5.

215. Prieto, C. and J. De Las Rivas, *APID: Agile Protein Interaction DataAnalyzer.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W298-302.
216. Prischi, F., et al., *The N-terminus of mature human frataxin is intrinsically unfolded.* FEBS J, 2009. **276**(22): p. 6669-76.
217. Prischi, F., et al., *Structural bases for the interaction of frataxin with the central components of iron-sulphur cluster assembly.* Nat Commun, 2010. **1**(7): p. 95.
218. Puccio, H., et al., *Mouse models for Friedreich ataxia exhibit cardiomyopathy, sensory nerve defect and Fe-S enzyme deficiency followed by intramitochondrial iron deposits.* Nat Genet, 2001. **27**(2): p. 181-6.
219. Py, B.a. and F.d.r. Barras, *Building Fe-S proteins: bacterial strategies.* Nat Rev Microbiol, 2010. **8**(6): p. 436-46.
220. Pynyaha, Y., et al., *Deficiency in frataxin homologue YFH1 in the yeast Pichia guilliermondii leads to missregulation of iron acquisition and riboflavin biosynthesis and affects sulfate assimilation.* Biometals, 2009. **22**(6): p. 1051-61.
221. Qi, W. and J.A. Cowan, *Structural; Mechanistic and Coordination Chemistry of Relevance to the Biosynthesis of Iron-Sulfur and Related Iron Cofactors.* Coord Chem Rev, 2011. **255**(7-8): p. 688-99.
222. Qin, S. and H.-X. Zhou, *meta-PPISP: a meta web server for protein-protein interaction site prediction.* Bioinformatics, 2007. **23**(24): p. 3386-87.
223. Rajnavölgyi, E., et al., *A repetitive sequence of Epstein-Barr virus nuclear antigen 6 comprises overlapping T cell epitopes which induce HLA-DR-restricted CD4(+) T lymphocytes.* Int Immunol, 2000. **12**(3): p. 281-93.
224. Ramazzotti, A., et al., *Mitochondrial functional interactions between frataxin and Isu1p, the iron-sulfur cluster scaffold protein, in Saccharomyces cerevisiae.* FEBS Lett, 2004. **557**(1-3): p. 215-20.
225. Rappuoli, R., *Reverse vaccinology.* Curr Opin Microbiol, 2000. **3**(5): p. 445-50.
226. Rappuoli, R. and A. Covacci, *Reverse vaccinology and genomics.* Science, 2003. **302**(5645): p. 602.
227. Raulfs, E.C., et al., *In vivo iron-sulfur cluster formation.* Proc Natl Acad Sci U.S.A., 2008. **105**(25): p. 8591-6.
228. Rawat, S. and T.L. Stemmler, *Key players and their role during mitochondrial iron-sulfur cluster biosynthesis.* Chemistry, 2011. **17**(3): p. 746-53.
229. Reddehase, M.J., et al., *A pentapeptide as minimal antigenic determinant for MHC class I-restricted T lymphocytes.* Nature, 1989. **337**(6208): p. 651-3.

230. Richardson, D., et al., *The ins and outs of mitochondrial iron-loading: the metabolic defect in Friedreich's ataxia.* J Mol Med (Berl), 2010. **88**(4): p. 323-9.
231. Rigaut, G., et al., *A generic protein purification method for protein complex characterization and proteome exploration.* Nat Biotechnol, 1999. **17**(10): p. 1030-2.
232. Ring, G.O., *Ocular Disease, Notably of the Uveal Tract, Induced by Streptococcus Viridans Infection.* Trans Am Ophthalmol Soc, 1927. **25**: p. 93-104.
233. Ritchie, D.W., *Recent progress and future directions in protein-protein docking.* Curr Protein Pept Sci, 2008. **9**(1): p. 1-15.
234. Ritchie, D.W., et al., *Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational FFT generating functions.* Bioinformatics, 2008. **24**(17): p. 1865-73.
235. Rodríguez-Ortega, M.J., et al., *Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome.* Nat Biotechnol, 2006. **24**(2): p. 191-7.
236. Rotkiewicz, P. and J. Skolnick, *Fast procedure for reconstruction of full-atom protein models from reduced representations.* J Comput Chem, 2008. **29**(9): p. 1460-5.
237. Rouault, T.A. and W.H. Tong, *Iron-sulphur cluster biogenesis and mitochondrial iron homeostasis.* Nat Rev Mol Cell Biol, 2005. **6**(4): p. 345-51.
238. Rouault, T.A. and W.H. Tong, *Iron-sulfur cluster biogenesis and human disease.* Trends Genet, 2008. **24**(8): p. 398-407.
239. Rouhier, N., et al., *Glutaredoxins: roles in iron homeostasis.* Trends Biochem Sci, 2010. **35**(1): p. 43-52.
240. Saha, S., et al., *Bcipep: a database of B-cell epitopes.* BMC Genomics, 2005. **6**(1): p. 79.
241. Saha, S. and G.P. Raghava, *BcePred: Prediction of Continuous B-Cell Epitopes in Antigenic Sequences Using Physico-chemical Properties*. In Nicosia,G., Cutello,V., Bentley,P.J. and Timis,J. (eds), Artificial Immune Systems, Third International Conference (ICARIS 2004), LNCS 3239. 2004: Springer. 197-204.
242. Saha, S. and G.P. Raghava, *Prediction of continuous B-cell epitopes in an antigen using recurrent neural network.* Proteins, 2006. **65**(1): p. 40-8.
243. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
244. Sambri, V., et al., *Experimental infection by Chlamydia pneumoniae in the hamster.* Vaccine, 2004. **22**(9-10): p. 1131-7.
245. Santos, R., et al., *Friedreich ataxia: molecular mechanisms, redox considerations, and therapeutic opportunities.* Antioxid Redox Signal, 2010. **13**(5): p. 651-90.

246. Sardana, D., et al., *Drug Repositioning for Orphan Diseases.* Brief Bioinform, 2011. **12**(4): p. 346-56.
247. Sayle, R.A. and E.J. Milner-White, *RASMOL: biomolecular graphics for all.* Trends Biochem Sci, 1995. **20**(9): p. 374.
248. Schilke, B., et al., *Evolution of mitochondrial chaperones utilized in Fe-S cluster biogenesis.* Curr Biol, 2006. **16**(16): p. 1660-5.
249. Schmucker, S. and H. Puccio, *Understanding the molecular mechanisms of Friedreich's ataxia to develop therapeutic approaches.* Hum Mol Genet, 2010. **19**(R1): p. R103-110.
250. Schmucker, S.p., et al., *Mammalian Frataxin: An Essential Function for Cellular Viability through an Interaction with a Preformed ISCU/NFS1/ISD11 Iron-Sulfur Assembly Complex.* PLoS ONE, 2011. **6**(1): p. e16199.
251. Schwimmer, C., et al., *Yeast models of human mitochondrial diseases: from molecular mechanisms to drug screening.* Biotech J, 2006. **1**(3): p. 270-81.
252. Seepersaud, R., et al., *Characterization of a novel leucine-rich repeat protein antigen from group B streptococci that elicits protective immunity.* Infect Immun, 2005. **73**(3): p. 1671-83.
253. Seguin, A., et al., *Evidence that yeast frataxin is not an iron storage protein In Vivo.* Biochim Biophys Acta, 2010. **1802**(6): p. 531-8.
254. Selbach, B., et al., *Kinetic analysis of the bisubstrate cysteine desulfurase SufS from Bacillus subtilis.* Biochemistry, 2010. **49**(40): p. 8794-802.
255. Sette, A. and R. Rappuoli, *Reverse vaccinology: developing vaccines in the era of genomics.* Immunity, 2010. **33**(4): p. 530-41.
256. Seznec, H., et al., *Friedreich ataxia: the oxidative stress paradox.* Hum Mol Genet, 2005. **14**(4): p. 463-74.
257. Shan, Y., et al., *Mitochondrial frataxin interacts with ISD11 of the NFS1/ISCU complex and multiple mitochondrial chaperones.* Hum Mol Genet, 2007. **16**(8): p. 929-41.
258. Sheftel, A., et al., *Iron-sulfur proteins in health and disease.* Trends Endocrinol Metab, 2010. **21**(5): p. 302-14.
259. Shet, A., et al., *Human immunogenicity studies on group A streptococcal C5a peptidase (SCPA) as a potential vaccine against group A streptococcal infections.* Indian J Med Res, 2004. **119 Suppl**: p. 95-8.
260. Shi, R., et al., *Structural basis for Fe-S cluster assembly and tRNA thiolation mediated by IscS protein-protein interactions.* PLoS Biol, 2010. **8**(4): p. e1000354.
261. Shi, Y., et al., *Human ISD11 is essential for both iron-sulfur cluster assembly and maintenance of normal cellular iron homeostasis.* Hum Mol Genet, 2009. **18**(16): p. 3014-25.

262. Shimomura, Y., et al., *The asymmetric trimeric architecture of [2Fe-2S] IscU: implications for its scaffolding during iron-sulfur cluster biosynthesis.* J Mol Biol, 2008. **383**(1): p. 133-43.
263. Smith, M.G. and M. Snyder, *Yeast as a model for human disease*. Current Protocols in Human Genetics, ed. J.L.H. Editorial Board. Vol. Chapter 15. 2006. Unit 15.6.1-15.6.8.
264. Sollner, J., et al., *Analysis and prediction of protective continuous B-cell epitopes on pathogen proteins.* Immunome Res, 2008. **4**: p. 1.
265. Stehling, O., et al., *Iron-sulfur protein maturation in human cells: evidence for a function of frataxin.* Hum Mol Genet, 2004. **13**(23): p. 3007-15.
266. Stehling, O., et al., *Investigation of iron-sulfur protein maturation in eukaryotes.* Methods Mol Biol, 2007. **372**: p. 325-42.
267. Stemmler, T.L., et al., *Frataxin and Mitochondrial FeS Cluster Biogenesis.* J Biol Chem, 2010. **285**(35): p. 26737-43.
268. 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.
269. Thomas, L.D., et al., *Immunisation with non-integral OMPs promotes pulmonary clearance of Pseudomonas aeruginosa.* FEMS Immunol Med Microbiol, 2003. **37**(2-3): p. 155-60.
270. Thompson, M.A. *ArgusLab 4.0.1*. 4.0.1:[Available from: http://www.arguslab.com.
271. Tirupati, B., et al., *Kinetic and structural characterization of Slr0077/SufS, the essential cysteine desulfurase from Synechocystis sp. PCC 6803.* Biochemistry, 2004. **43**(38): p. 12210-9.
272. Toseland, C.P., et al., *AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data.* Immunome Res, 2005. **1**(1): p. 4.
273. Trost, B., et al., *Bacterial peptides are intensively present throughout the human proteome.* Self/Nonself, 2010. **1**(1): p. 71-4.
274. Trost, B., et al., *No human protein is exempt from bacterial motifs, not even one.* Self/Nonself, 2010. **1**(4): p. 328-34.
275. Tsai, C.-L. and D.P. Barondeau, *Human frataxin is an allosteric switch that activates the Fe-S cluster biosynthetic complex.* Biochemistry, 2010. **49**(43): p. 9132-9.
276. Tsai, C.-L., et al., *Friedreich's ataxia variants I154F and W155R diminish frataxin-based activation of the iron-sulfur cluster assembly complex.* Biochemistry, 2011. **50**(29): p. 6478-87.

277. Turner, P.C., et al., *Neisserial TonB-dependent outer-membrane proteins: detection, regulation and distribution of three putative candidates identified from the genome sequences.* Microbiology, 2001. **147**(Pt 5): p. 1277-90.
278. Umelo-Njaka, E., et al., *Expression and testing of Pseudomonas aeruginosa vaccine candidate proteins prepared with the Caulobacter crescentus S-layer protein expression system.* Vaccine, 2001. **19**(11-12): p. 1406-15.
279. Vajda, S. and D. Kozakov, *Convergence and combination of methods in protein-protein docking.* Curr Opin Struct Biol, 2009. **19**(2): p. 164-170.
280. Vakser, I.A. and P. Kundrotas, *Predicting 3D structures of protein-protein complexes.* Curr Pharm Biotechnol, 2008. **9**(2): p. 57-66.
281. Valentin-Weigand, P., et al., *The fibronectin binding domain of the Sfb protein adhesin of Streptococcus pyogenes occurs in many group A streptococci and does not cross-react with heart myosin.* Microb Pathog, 1994. **17**(2): p. 111-20.
282. van den Elsen, J., et al., *Bactericidal antibody recognition of meningococcal PorA by induced fit. Comparison of liganded and unliganded Fab structures.* J Biol Chem, 1999. **274**(3): p. 1495-501.
283. Van Regenmortel, M.H.V., *What is a B-cell epitope?* Methods Mol Biol, 2009. **524**: p. 3-20.
284. Vázquez, A., et al., *Global protein function prediction from protein-protein interaction networks.* Nat Biotechnol, 2003. **21**(6): p. 697-700.
285. Veatch, J.R., et al., *Mitochondrial Dysfunction Leads to Nuclear Genome Instability via an Iron-Sulfur Cluster Defect.* Cell, 2009. **137**(7): p. 1247-58.
286. Vermont, C. and G. van den Dobbelsteen, *Neisseria meningitidis serogroup B: laboratory correlates of protection.* FEMS Immunol Med Microbiol, 2002. **34**(2): p. 89-96.
287. Viau, M. and M. Zouali, *Effect of the B cell superantigen protein A from S. aureus on the early lupus disease of (NZBxNZW) F1 mice.* Mol Immunol, 2005. **42**(7): p. 849-55.
288. Vieira, O.V., et al., *Distinct roles of class I and class III phosphatidylinositol 3-kinases in phagosome formation and maturation.* J Cell Biol, 2001. **155**(1): p. 19-25.
289. Vohra, H., et al., *M protein conserved region antibodies opsonise multiple strains of Streptococcus pyogenes with sequence variations in C-repeats.* Res Microbiol, 2005. **156**(4): p. 575-82.
290. Vytvytska, O., et al., *Identification of vaccine candidate antigens of Staphylococcus aureus by serological proteome analysis.* Proteomics, 2002. **2**(5): p. 580-90.

291. Walesiak, M. and A. Dudek. *clusterSim: Searching for Optimal Procedure for a Data Set*. 2007; Available from: http://CRAN.R-project.org/package=clusterSim.
292. Wang, T. and E.A. Craig, *Binding of yeast frataxin to the scaffold for Fe-S cluster biogenesis, Isu.* J Biol Chem, 2008. **283**(18): p. 12674-9.
293. Weeratna, R., et al., *Human and guinea pig immune responses to Legionella pneumophila protein antigens OmpS and Hsp60.* Infect Immun, 1994. **62**(8): p. 3454-62.
294. Weichhart, T., et al., *Functional selection of vaccine candidate peptides from Staphylococcus aureus whole-genome expression libraries in vitro.* Infect Immun, 2003. **71**(8): p. 4633-41.
295. Welsch, J.A., et al., *Protective activity of monoclonal antibodies to genome-derived neisserial antigen 1870, a Neisseria meningitidis candidate vaccine.* J Immunol, 2004. **172**(9): p. 5606-15.
296. Wiedemann, N., et al., *Essential role of Isd11 in mitochondrial iron-sulfur cluster synthesis on Isu scaffold proteins.* EMBO J, 2006. **25**(1): p. 184-95.
297. Wizemann, T.M., et al., *Use of a whole genome approach to identify vaccine molecules affording protection against Streptococcus pneumoniae infection.* Infect Immun, 2001. **69**(3): p. 1593-8.
298. Worgall, S., et al., *Protection against P. aeruginosa with an adenovirus vector containing an OprF epitope in the capsid.* J Clin Invest, 2005. **115**(5): p. 1281-9.
299. Wu, G. and L. Li, *Biochemical Characterization of Iron-Sulfur Cluster Assembly in the Scaffold IscU of Escherichia coli.* Biochemistry (Mosc), 2012. **77**(2): p. 135-42.
300. Wyllie, S., et al., *Single channel analysis of recombinant major outer membrane protein porins from Chlamydia psittaci and Chlamydia pneumoniae.* FEBS Lett, 1999. **445**(1): p. 192-6.
301. Xu, X.M. and S.G. Möller, *Iron-Sulfur Clusters: Biogenesis, Molecular Mechanisms and Their Functional Significance.* Antioxid Redox Signal, 2011. **15**(1): p. 271-307.
302. Ye, H. and T.A. Rouault, *Human iron-sulfur cluster assembly, cellular iron homeostasis and disease.* Biochemistry, 2010. **49**(24): p. 4945-56.
303. Yoon, H., et al., *Mutation in the Fe-S scaffold protein Isu bypasses frataxin deletion.* Biochem J, 2012. **441**(1): p. 473-80.
304. Zagursky, R.J., et al., *Identification of a Haemophilus influenzae 5'-nucleotidase protein: cloning of the nucA gene and immunogenicity and characterization of the NucA protein.* Infect Immun, 2000. **68**(5): p. 2525-34.

305. Zhang, Y., et al., *Frataxin and mitochondrial carrier proteins, Mrs3p and Mrs4p, cooperate in providing iron for heme synthesis.* J Biol Chem, 2005. **280**(20): p. 19794-807.
306. Zhang, Y., et al., *Mrs3p, Mrs4p, and frataxin provide iron for Fe-S cluster synthesis in mitochondria.* J Biol Chem, 2006. **281**(32): p. 22493-502.