Data Science Research to Support Stem Cell Therapy for Muscular Dystrophy

by

Karen Nathaly Hernández Salas

A thesis submitted to the
Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Master of Computer Science with Data Science Specialization

Ottawa-Carleton Institute for Computer Science
School of Computer Science
Carleton University
Ottawa, Ontario

November, 2017

©Copyright
Karen Nathaly Hernández Salas, 2017

The undersigned hereby recommends to the
Faculty of Graduate and Postdoctoral Affairs
acceptance of the thesis

Data Science Research to Support Stem Cell Therapy for Muscular Dystrophy

submitted by **Karen Nathaly Hernández Salas**

in partial fulfillment of the requirements for the degree of

Master of Computer Science with Data Science Specialization

Professor Marcel Turcotte, Ph.D., External Examiner

Professor Olga Baysal, Ph.D., Examiner

Professor Frank Dehne, Ph.D., Thesis Supervisor

Professor Michiel Smid, Ph.D., Chair,
School of Computer Science

Ottawa-Carleton Institute for Computer Science
School of Computer Science
Carleton University
November, 2017

Abstract

This thesis project focused on using a sequence-based, high-performance computational tool to design synthetic proteins and is part of current collaborative research on Duchenne Muscular Dystrophy (DMD). A possible treatment for DMD consists of injecting patients with healthy muscle satellite cells grown in tissue culture. However, such cells cannot currently be produced in quantity because they convert to muscle cells (differentiate) prematurely. Using InSiPS, the **In**-**Si***lico* **P***rotein* **S***ynthesizer*, protein sequences were designed to interact with target proteins and inhibit the protein-protein interaction proposed to regulate the premature differentiation. The resulting sequences were predicted to interact with the target proteins with high specificity (99.98%). Complementary biochemistry experiments indicated interactions with the intended target for two out of ten synthetic proteins. These results are being studied as part of the ongoing research seeking to develop a treatment for DMD.

My driving force is, was and always will be, the people I care about the most, that is my lovely family. Husband, mom, dad, siblings, nephew, those who have already left, and all those yet to come, my life has a sense of purpose because of you.
Thank you from the bottom of my heart for being there for me, always.
This work is dedicated to each one of you.

Acknowledgments

I would like to acknowledge and give my sincerest appreciation to the person who made possible my participation in this research project, Dr. Frank Dehne. I will always be grateful for having such a recognized investigator as my thesis supervisor, who was always happy to help and give advice with his contagious passion for research that motivated me to improve myself.

An enormous thank you to Dr. Andrew Schoenrock, for his infinite patience in helping me at every step of the project, and whose expertise and knowledge of PIPE and InSiPS were crucial for my progress.

Also, a big thank to Daniel Burnside for kindly helping me to understand and to explain biochemistry concepts. I also thank Dr. Ashkan Golshani and everyone in the Carleton Bioinformatics group. I was very lucky to be among dedicated investigators who were always happy to provide ideas to improve the present work.

Special thanks go to Dr. Alexandre Blais, who developed the initial research proposal, who designed the biochemical experimental validation and who led the group that performed it — all with great commitment and a high sense of collaboration.

My infinite appreciation goes also to Carole Love, and all the hard-work she put into proofreading and helping me with suggestions that make me feel proud of my thesis writing.

I want to express my deepest gratitude to the Mexican National Council for Science and Technology (CONACyT) for sponsoring my graduate studies on behalf of my country — México — and for helping me to pursue this major milestone in my life.

I am very fortunate to have in my life a special colleague who happens to be my husband and love of my life. Thank you, Pablo for your unending support and for sharing with me this adventure.

This research work would have been impossible to achieve without the IBM Blue Gene/Q supercomputer — a Southern Ontario Smart Computing Innovation Platform (SOSCIP).

Contents

List of Tables

List of Figures

List of Abbreviations

FN	False Negative
FP	False Positive
GA	Genetic Algorithm
InSiPS	In-Silico Protein Synthesizer
MP-PIPE	Massively Parallel Protein-Protein Interaction Prediction Engine
PIPE	Protein-Protein Interaction Prediction Engine
PPI	Protein-Protein Interaction
ROC Curve	Receiver Operating Characteristic Curve
TN	True Negative
TP	True Positive

Chapter 1

Introduction

1.1 Overview of the Thesis

The work presented aims to contribute to an ongoing research effort looking to move forward with a treatment for Duchenne muscular dystrophy (DMD). DMD is the most common form of muscular dystrophy in childhood. Patients with this condition present with progressive muscle weakness, resulting in a shortened life span [2]. The Canadian Neuromuscular Disease Registry (CNDR) collected information on patients with neuromuscular diseases across Canada [3], among which 253 with muscular dystrophy were included. Via the CNDR registry, a study based on surveys to families of 176 children — between 4 and 18 years of age — with DMD reported that fatigue and the need for wheelchair use severely impacted health-related quality of life in patients [4]. One attempt to combat this degenerative disease involves a joint effort along two independent research lines, with investigators drawn from the disciplines of data science and biochemistry. The research lines are described below.

On the biochemistry side, there is research taking place at the Ottawa Institute of Systems Biology, as part of a continuous effort to study muscle cells [5, 6]. This research seeks to grow satellite cells (muscle stem cells) in tissue culture and then inject them into DMD patients as a potential treatment for DMD. However, satellite cells cannot currently be grown in quantity because they convert to muscle cells too quickly. Investigators working on this research aim to produce an inhibitory protein with the characteristic of arresting the early conversion of satellite cells so that they can proliferate and be of help in DMD treatment.

On the data science side, there is an **In-Si***lico* (i.e., computational) **P***rotein* **S***ynthesizer*, namely InSiPS [7]. InSiPS was developed at Carleton University and is a

result of years of research on sequence-based computational tools for the prediction of protein-protein interactions (PPIs). The objective is to design protein sequences predicted to interact with a target protein without affecting other non-target proteins. InSiPS combines PIPE [8], a protein-protein interaction prediction engine, with a genetic algorithm (GA) that performs thousands of high processing iterations. This computational method requires large computational resources and has been implemented for the IBM Blue Gene/Q (BGQ), a supercomputer ranked among the most powerful high-performance computers in the TOP500 project (www.top500.org).

The disruption of a particular PPI has been identified as a potential target to avoid the early conversion of muscle satellite cells. The basic idea of this research was to use InSiPS to design a synthetic protein sequence (of characters representing amino acids) able to function as an inhibitory protein. Predicted to interact with a given protein without affecting other proteins, the novel synthetic protein would potentially be able to disrupt the PPI proposed to regulate the early conversion of muscle satellite cells. Consequently, it could help researchers to move forward with investigating the prospective treatment for DMD.

1.2 Scope of the Thesis

The scope of this thesis work was to collaborate with researchers working on a treatment for DMD by providing data that would allow them to continue with their investigation. The author's contribution consisted in running InSiPS on the BGQ supercomputer and performing several computational experiments, looking for synthetic protein sequences predicted to interact with a target protein and not to interact with the rest of the proteins. This work involved, firstly, analyzing and selecting the input data; next, performing the computational experiments; and ultimately, analyzing and selecting relevant output data.

The main tasks carried out throughout this research are summarized as follows: (1) execution of preliminary tests with human proteins data on two computer clusters with different characteristics; (2) analysis and selection of data for the formal execution of experiments, using two recently collected data sets of human proteins; (3) execution of different tests with different input data, especially on the BGQ; (4) analysis and selection of the obtained results in the collaborative work, to decide which of the new protein sequences would be produced as actual synthetic proteins

for wet-lab experiments. Neither enhancements nor additional code development were designed or implemented to perform the experiments with InSiPS. It is also worth noting that the wet-lab experiments were done on the biochemistry side; however, the experimental procedure is mentioned in this thesis to support and validate the obtained results.

1.3 Main Contributions

The preliminary analysis and selection of the input data followed by the formal execution of experiments with InSiPS, led to the generation of several synthetic protein sequences. These sequences were predicted to interact with the intended target protein and not to interact with the defined non-target proteins. Further analysis of the generated data allowed identification of the highest ranked protein sequences. Furthermore, the InSiPS results were supported by complementary biochemistry experiments where two out of ten synthetic proteins were indicated to interact with the target protein. Based on this, a summary of the two major contributions is described below.

- The fact that the synthetic proteins showed a sign of interaction with the intended target through experimental validation was the first major contribution of this thesis. These synthetic proteins are now considered as potential protein inhibitors of the unwanted differentiation of satellite cells and could allow the production of muscle stem cells — which could then be further investigated as a potential treatment for DMD.

- InSiPS was successfully used in the past to design protein inhibitors for *S. cerevisiae* (a species of yeast), but it had not been used for human proteins. Another contribution was that this study represented the first time that InSiPS was used with *H. sapiens* proteins data. Additionally it was collaborative research, where InSiPS was a determinant factor. Should the ultimate results be good, this will likely increase the reliance on using InSiPS to design inhibitory proteins for other types of chronic diseases.

1.4 Motivation of the Thesis

The main reason that motivated and inspired the presented work was the opportunity to use a data science approach as a part of interdisciplinary collaboration to achieve a common objective. Specifically, there was the aspiration to contribute in the health-care field via a state-of-the-art subject matter — namely computational *de novo* protein design — by using algorithms running in parallel on large computational resources. This could help to bridge the gap that currently prevents medical science from progressing in the investigation of a potential treatment to combat a severe disease for which no cure has yet been found.

Most of the research in protein engineering so far has been directed towards the modification of naturally occurring proteins. Computational methods represent great potential in the *de novo* design of proteins, where the aim is to explore a full sequence space among 20^n possibilities to form a protein sequence of n amino acid residues [9]. Using conventional computational resources, it would be a titanic — or better said frankly impossible — task to design a new protein sequence to accomplish a specific goal, even for the most committed data scientist in the world. This could only be achievable through the application of a data science approach (i.e., InSiPS) by means of a wide-ranging selection of input data, the execution of computational experiments, and data analysis of results; in addition to joint interdisciplinary work.

1.5 Thesis Organization

1.5.1 Background: A Small Measure of Biochemistry Information

Chapter 2 contains biochemistry information and outlines the background behind this research in more detail. First, the importance of protein-protein interactions (PPI) is described. Next, information about the research on muscular dystrophy, the therapy treatment researchers are working on, and the obstacles that prevent them from moving forward with this research are outlined. Also, biochemistry terms that are mentioned often in this thesis are explained.

1.5.2 Review: Methods for the Prediction of Protein-Protein Interactions and Protein Design

Chapter 3 presents a survey of the literature focused on computational methods related to InSiPS. Firstly, the different categories of methods for the prediction of PPIs are mentioned, with methods that fall into two of the categories being described. Secondly, highlights of different reviews on computational protein design are provided, with some applications in this field being mentioned.

1.5.3 InSiPS, the *In-Silico* Protein Synthesizer

Chapter 4 gives a review of InSiPS, the computational method used for this research. As InSiPS relies on PIPE for the prediction of protein-protein interactions, this tool is also described. For both InSiPS and PIPE, aspects such as the required input data, key parameters, basics of the main algorithms, the output data and the process for validation of results are explained.

1.5.4 Problem Statement

Chapter 5 summarizes the problem this thesis aimed to solve. It also gives the reader a concise description of the justification and reasons that motivated this thesis. Thus, a bird's-eye view of the problem is given along with the major challenge and how this problem could be tackled by the present research work.

1.5.5 InSiPS Experiments with Human Proteins Data

Chapter 6 summarizes how the proposed solution for the problem was implemented. This recap includes the protein data used and the specific input values that led to the ultimate results. Also, the technical characteristics of the computer clusters used are described. The details and interpretation of the results, and how these results were validated, are also included. Also given is the description of the last round of experiments which focused on designing protein sequences to bind to specific sites of a target protein.

1.5.6 Data Analysis of Results

The data analysis of the results obtained in the formal execution of experiments with InSiPS is reviewed in Chapter 7. This chapter includes a summary of the analysis of specific protein sequences, such as the homology with other proteins and the predicted interaction sites. This analysis was fundamental to determine which protein sequences would be used for the subsequent wet-lab experimental validation phase carried out by biochemistry investigators.

1.5.7 Wet-Lab Experimental Validation

Chapter 8 describes the complementary wet-lab experimental validation process that helped to support the results obtained in the formal execution of the computational experiments. Although it was outside the scope of data science, a summary of the process that led to the finding that two out of ten of the synthetic proteins designed with InSiPS were reported to interact with the target protein is included. The chapter also includes a description of the future work needed to continue with this research project, especially from the biochemistry side.

1.5.8 Conclusions

Chapter 9 concentrates on the details of the conclusions that came out of the work performed in this research project. Discussed are the feasibility of using InSiPS to design human proteins, the importance of interdisciplinary collaboration, and a more detailed summary of the contributions. Finally, the chapter presents the future work regarding the use of InSiPS for other research projects similar to the one reported in this thesis.

Chapter 2

Background: A Small Measure of Biochemistry Information

This chapter gives some general background on the field this research study aimed to support, namely biochemistry. Section 2.1 describes what protein-protein interactions (PPIs) are and outlines some laboratory techniques used to study them. Section 2.2 outlines the research project from a biochemistry perspective. Finally, Section 2.3 includes definitions of biochemistry terminology used in subsequent chapters.

2.1 Protein-Protein Interactions

Proteins are large and complex organic compounds consisting of a unique sequence of amino acids that folds into a determined three-dimensional shape. Proteins carry out biological processes in organisms by physically interacting with other proteins or molecules (e.g., DNA and RNA). The region of a protein that binds to another molecule is known as a binding site, which is a portion of the protein surface. All proteins bind to other molecules with different affinity, and in most cases with high specificity, as they only adhere to one or just a few molecules from the thousands of proteins available in a given proteome [10].

The order of the amino acids in a protein sequence is fundamental to predicting the 3D structure of the protein, and hence its biological function [11] because of the close relationship between structure and function [12]. For several decades protein interactions have been the focus of numerous studies aiming to investigate the function of proteins; other studies have been carried out to predict and identify the structure and shape of proteins. This has allowed greater understanding of the diverse biological

aspects of cellular functions. More recently protein interactions have been studied to understand the mechanisms that lead to disease. As a consequence of this work, they have also been studied as potential targets for therapeutic treatments [13].

2.1.1 Laboratory Techniques for Experimental Validation

Techniques for experimental identification of PPIs can be grouped into (i) biophysical methods for characterization and (ii) high-throughput methods for detection. Biophysical methods allow the characterization of binding partners and thus provide details of the biochemical features of the interactions; examples of this kind of experiment are *X-ray crystallography, NMR spectroscopy, fluorescence polarization/anisotropy* and *atomic force microscopy.* High-throughput methods are of help in detecting protein interactions at a larger scale; examples of these methods are *yeast two-hybrid, mass spectrometry*, and *affinity purification* [13]. What follows is a description of two of the most common high-throughput methods for PPI detection.

Yeast two-hybrid

This method is a powerful technique for *in vivo* identification of PPIs and protein-DNA interactions. The main idea behind this method is that it uses a transcription factor (GAL4 proteins) or viral proteins (VP16) and two distinct domains: the DNA-binding domain (DBD) and the activation domain (AD). Then two proteins, bait and prey, are fused to each of the domains. If the co-expression of these two hybrid proteins reconstitutes the GAL4 or VP16 domains, it means that the proteins interact. This method is widely used, although it has reported a low specificity (true negative rate) and hence a high rate of false positives [14].

Western blotting

This method allows the separation and identification of specific proteins from a mixture of proteins extracted from cells. The separation of proteins is through gel electrophoresis (SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis), a technique that separates molecules based on the differences in their molecular weight. The proteins are then transferred to a membrane and the proteins of interest are detected by using specific antibodies. Although some problems in troubleshooting have been identified, this technique is widely used [15].

2.1.2 Homology and Alignment in Protein Sequences

Homology between two protein sequences or structures is an indication that they share a common ancestor in the evolutionary tree of life. Similarity in protein sequence or structure is a good indicator to infer homology. For example, by looking at homologous proteins, a protein's shape and function might be inferred, or the specific protein's sites to attach inhibitor proteins. In bioinformatics, sequence alignment between two or more proteins is a widely used technique and several algorithms have been developed looking to optimize this process. Two frequently used matrices for aligning amino acid sequences are PAM (Percent Accepted Mutation) and BLOSUM (Blocks Substitution Matrix). These matrices were designed to assess the similarity of each amino acid by taking its evolution into consideration [16].

2.2 Duchenne Muscular Dystrophy Research Project

This section describes the fundamentals of ongoing research seeking a therapeutic treatment for muscular dystrophy. Dr. Alexandre Blais, from the Ottawa Institute of Systems Biology, came up with the initial proposal and defined the research problem summarized in this section. There is first a general description of the disease; next, the proposed therapy and main problem impacting the research; and finally, a possible solution to address this problem.

2.2.1 Duchenne Muscular Dystrophy

Duchenne muscular dystrophy (DMD) is a genetic disorder characterized by a progressive loss of muscle tissue strength and density. This pathology is linked to the X-chromosome (i.e., it is expressed in males) and is caused by mutations in the gene encoding dystrophin, which is an essential protein in muscle fibers. The absence of normally functioning dystrophin leads to gradual muscle degeneration and weakness, ultimately causing death by the second or third decade of life. DMD affects approximately one in 3,500 male children. There is no curative treatment yet, although numerous experimental approaches have been studied over the years [2, 17].

DMD has several associated complications such as cardiac manifestations and

chronic respiratory insufficiency. Symptomatic and rehabilitative approaches, particularly the use of corticosteroids and respiratory support, have improved the quality of life for patients, as well as increased life expectancy [18].

2.2.2 Muscle Stem Cell Therapy

Muscle cells are unique in that they are multinucleate (i.e., they have multiple nuclei per muscle cell). When muscle cells are damaged by injury or extensive exercise, they use muscle stem cells to repair the damaged tissue. Muscle stem cells, also known as satellite cells, are essential for muscle growth and regeneration. Satellite cells remain quiescent or dormant until they are needed to regenerate tissue. Once activated, they begin to proliferate and set in motion the transcription factors responsible for the formation of skeletal muscle tissue. They then differentiate into muscle fibers and work to regenerate the muscle [19]. In other words, they repair the tissue and fuse to the muscle cell as another nucleus. Also, satellite cells can fuse to each other and form new muscle fibers.

The role of dystrophin in satellite cells was studied by Dumont *et al.* at the Ottawa Hospital Research Institute and the University of Ottawa [20]. This study represented a ground-breaking discovery as it was found that dystrophin is highly expressed in activated muscle stem cells, and that it has a fundamental role in satellite cell regulation. The authors concluded that muscle wasting in DMD is greatly aggravated by deficient muscle tissue regeneration caused by the dysfunction of satellite cells. The role of dystrophin is also very important for maintaining a population of stem cells. In DMD the constant deterioration of muscle fibers requires constant regeneration by satellite cells. Mutated dystrophin causes the number of stem cells to decrease over time making patients less able to repair their muscles, as described by Keefe and Kardon [21]. While current therapies target already differentiated muscle cells, the role of dystrophin in satellite cells is of great importance for future gene therapy that may consider the function of dystrophin in both muscle fibers and satellite cells.

Potentially satellite cells could be used in therapy, as they clearly contribute to muscle generation [22]. A possible treatment for DMD proposed by researchers at the Ottawa Institute of Systems Biology consists in growing large quantities of satellite cells in tissue culture, so that they could be injected into dystrophic patients to regenerate muscle tissue. The first step would be to perform a muscle biopsy, a procedure to extract a sample of muscle tissue from a healthy donor. Next the

muscle satellite cells would be isolated for *in vitro* proliferation; that is growing them in large quantities in tissue culture. Lastly, the satellite cells would be transplanted into the dystrophic patient.

Differentiation of Muscle Stem Cells

The use of satellite cells for cell-based transplantation therapy for diseases related to muscle degeneration such as DMD has been studied. However, some major obstacles to transplantation need to be resolved before proceeding to clinical trials. When attempting to grow satellite cells in tissue culture, they do not grow well under *in vitro* conditions and tend to convert (differentiate) into muscle cells prematurely. At this point medical science is unable to control the differentiation (conversion to muscle cells) of these satellite cells. If satellite cells could be grown in tissue culture, tightly controlling for the premature differentiation, they could potentially be considered as a therapy for DMD. The inability to grow sufficient quantities of satellite cells in tissue culture without losing their regenerative properties and converting into muscle fibers still remains the major challenge [23].

2.2.3 Focus on the SIX-EYA Interaction to Address the Cell Differentiation Challenge

A group of proteins known as the SIX (sine oculis homeobox) family of transcription factors was recently implicated as playing an important role in muscle cell differentiation. The SIX family of proteins consists of six homeodomain transcription factors, from SIX1 to SIX6, and is found in diverse organisms, including humans. These proteins have common structural characteristics for the conserved SIX domain (SD) and for the SIX-type homeodomain (HD). This homology for SD and HD in SIX proteins can be observed by looking at the identity percentage when aligning the amino acid sequences [24]. SIX proteins are known to play a major role in the development of cell populations for several tissue types [25]. Dr. Blais *et al.* have proposed that SIX family members play an essential function in differentiation during adult muscle regeneration, with regulatory factors for skeletal myogenesis, which is the process of forming muscle tissue from stem cells. In particular, in studying SIX1 and SIX4, they reported the function of SIX factors as essential to the process of myogenic differentiation (conversion from satellite cells to muscle cells) [5, 6]. It was also shown

in another study [26] that proteins of the SIX family are involved in the muscle regeneration process, identifying this family as a potential target in the manipulation of satellite cells for therapeutic use.

Several transcription factors that bind together induce the production of key proteins that control the duplication and differentiation of satellite cells. In particular, the EYA family of proteins (EYA1 to EYA4), a family of cofactors for SIX1, has been reported to activate the SIX1 transcriptional function [27].

Of special interest for the present research was the protein-protein interaction between the SIX and EYA transcription factors. The disruption of the SIX-EYA interaction, by competitively binding the interaction motif of either protein, has been identified as a potential target to avoid the premature differentiation of satellite cells; consequently, it is a potential target in the search for a therapeutic treatment for DMD.

SIX-EYA Interaction Site of Interest

Protein-protein interactions only occur at specific regions of each of the proteins involved in the interaction. In order to maximize the likelihood of blocking certain PPIs, it is crucial to know the binding sites mediating that protein interaction. Patrick *et al.* [27] found that the predominant domain of SIX1 that enables the interaction with EYA occurs in the SIX domain, near the n-terminus region, specifically on the SIX1 α-helix 1 consisting of amino acids 9 to 27 (*FTQEQVACVCEVLQQGGNL*).

The disruption of the interaction of the SIX1-EYA complex is of great interest for this research, and this complex has also been related to the development of diverse tumor types. Specifically, the SIX1-EYA2 protein interaction was shown to be crucial in the spread of breast cancer [28] and in the branchio-oto-renal syndrome (BOR) [27].

2.3 Biochemistry Terminology

Amino acids

Amino acids are organic molecules that are used in every cell to build proteins. More than 50 different amino acids have been identified, but only 20 of them are used in the making of naturally occurring proteins. Each amino acid has specific properties and characteristics and a single letter is used for its representation.

Co-purification

It is the physical separation or isolation of two or more proteins from a complex mixture of other proteins. The fact that two proteins co-purify means that they attract each other from among the other proteins in the mixture.

Homeodomain

A common structural motif that binds to DNA, found in many eukaryotic regulatory proteins. Proteins that have a homeodomain are involved in the transcriptional control of developmentally important genes and play a critical role in many cellular processes. [29, 30]. The homeodomain part of a protein is also known as the DNA binding region. The DNA binding regions mentioned in this thesis were obtained from the Universal Protein Resource (UniProt) [31].

Motif

A short subsequence of amino acids or a determined region in the protein structure that performs a specific function.

Peptide

A short chain of amino acids consisting of around 50 amino acids or fewer.

Protein expression

The production of proteins in living cells, performed through different techniques.

Protein sequences

Linear sequence of amino acids that describe the primary structure of a protein.

Protein naming

Identification of the natural proteins described in this thesis follows the naming convention of the Universal Protein Resource (UniProt) [31]. For example, when referring to the protein *Guanine nucleotide-binding protein G(i) subunit alpha-2* (GNAI2,

P04899), the first description is the name and in parenthesis is the *gene name* followed by the UniProt *id*.

Protein n-terminus region

Also known as the amino terminus, this is the left end of a peptide chain (closest to the start of the peptide). Peptide sequences are written from the n-terminus to the c-terminus, also known as carboxyl-terminus (closest to the end of the peptide).

Proteome

The complete set of proteins expressed in a determined cell or organism.

Wet-lab experiments

Wet-lab experimentation can be defined as the biochemical experiments using living cell cultures. This refers to experiments performed in a *wet laboratory*. These spaces are equipped for practical (non-theoretical) scientific research through diverse materials and drugs.

Chapter 3

Review: Methods for the Prediction of Protein-Protein Interactions and Protein Design

The importance of protein-protein interactions (PPIs) and a few techniques for their experimental validation were described in Chapter 2. As is mentioned in Chapter 8, performing a biochemistry experiment may involve a long and complicated process. Moreover, the possibility of performing validations for a massive amount of data (e.g., from the scanning of interaction networks) by means of pure biochemistry experiments remains challenging. Over the past few decades computational methods for the prediction of PPIs have been well studied [32, 33] with an extensive variety of approaches. These methods arose out of a necessity to increase the reach of biochemistry research: instead of being used as the primary method to obtain results, laboratory experiments have started to be used to corroborate and validate computationally generated data. Since the emergence of computational methods for predicting PPIs, a wide variety of approaches have been proposed. Similarly, the amount of publicly available protein data [34–36] generated through biochemistry experiments, computational experiments, or by the complementary use of both techniques, has dramatically increased.

The extensive study of PPIs has led to a variety of research paths, among which protein design has been the focus for a number of approaches with the implementation of computational methods. A precise description of the interactions between molecules, together with a powerful algorithm to enable the exploration of a vast number of possible solutions, are fundamental ingredients for protein design [37].

This chapter reviews some of the most relevant research as described in the literature on the tools related to the computational method applied in this thesis. The review is divided into two main sections. The first section, 3.1, on methods for the prediction of PPIs, describes the different categories of computational tools and reviews some of them. The second section, 3.2, on computational protein design, describes relevant studies of computational protein design and includes some recent applications in the field.

3.1 Computational Methods for the Prediction of PPIs

The various computational methods for the prediction of PPIs have been categorized depending on aspects such as the type of protein relationship investigated, and the type of protein data explored (i.e., protein structure, amino acid sequence) [38, 39]. Pitre *et al.* at Carleton University, classified the computational methods for the study of PPI prediction into five general categories [38]: (1) genomic methods, focused on the genes of proteins homologous or physically close in different genomes; (2) evolutionary relationship, intended to infer interactions based on the phylogenetic profile of a protein; (3) protein structure, aimed at assessing the compatibility of interacting regions given three-dimensional structure information; (4) domain-based, used to evaluate conserved domains among proteins for potential interacting domain protein pairs; and (5) primary protein structure, concentrated on the use of the primary sequence of amino acids to predict PPIs.

Different approaches have been presented for each of the general categories described earlier. To highlight only work relevant to this thesis, two of the categories are reviewed. First, methods that fall into the category of three-dimensional (3D) protein structure are described to give the reader a general idea of how they work and to have a point of comparison for the other category presented. Second, methods that fall under the category of primary protein structure are examined, as the work presented in this thesis is based on this category.

The task of comparing different approaches, even when they are in the same category, is quite complicated. There might be considerable differences in the classification features, in the data sets, the testing schemes, classification methods, and consequently in the resultant accuracy. Therefore each of the methods described in

this section is presented with the most relevant characteristics. It should be noted, however, that some of the characteristics are common to most of the methods. Due to the lack of available data on protein interactions, the authors of the different methods have their own strategies for building the non-interacting pairs and they use data sets of non-redundant proteins with more than 50 amino acids, generally using yeast or human protein data. Also, only a few methods are specialized in interactome (the totality of PPIs that happen in a cell) prediction. Those methods that are do not consider in a realistic manner the possibility of imbalanced data sets, which might be more than a 600:1 ratio of negative to positive interactions [38]. The most used testing methods are 3-fold, 5-fold and 10-fold cross validation. Lastly, with just a few of the methods, were results validated through wet-lab experimentation.

3.1.1 Methods Based on Protein Structure

A major objective in current biological research is to organize molecular interactions as networks to analyze and identify their parts and connections, and in particular to characterize the relationships of activation or inhibition among proteins [40]. As described in [41], structural modeling helps to gain a better understanding of the molecular basis of a PPI. According to the authors, the necessity of protein-protein docking algorithms emerged due to the fact that the structures of many protein complexes have not been experimentally validated, related to cost and experimental limitations. A high resolution structural model of a protein complex can provide the atomic details for a given PPI, helping with the design of therapeutic molecules aiming to either interact with a protein or to inhibit another PPI [41].

Yugandhar and Gromiha [42] collected experimental binding affinity data for a set of 135 protein-protein complexes and calculated sequence-based and structure-based features based on previous literature. They classified the complexes into different function-based classes and developed a method for predicting the binding affinity of protein-protein complexes by using a multiple regression technique.

Pierce *et al.* presented ZDOCK, a server to predict the structures of protein-protein complexes and symmetric multimers. The server is publicly available to produce structural models of these predictions via a user web interface [41]. The rigid-body docking program ZDOCK uses the fast Fourier transform algorithm for a global docking search on a 3D grid and a combination of shape complementarity, electrostatics and statistical potential terms.

MEGADOCK, presented by Ohue *et al.*, is a method for the prediction of PPIs using a rigid-body docking approach [43]. This approach, based on electrostatic forces and shape complementarity, accelerates the docking calculation process. MEGADOCK consists of two segments: one is a *docking calculation*, where an all-to-all docking calculation is performed; the other is a *PPI decision segment*, where the structural distributions for each protein pair are analyzed to finally decide whether it interacts. By using a technique called the real Pairwise Shape Complementarity (rPSC) score, and parallelizing the tool, the protein docking calculations were faster than in previous versions of the software. The rPSC score consists of three-dimensional voxels, where each voxel represents the area occupied by proteins or unoccupied space. In addition to the rPSC score, the authors also used a physicochemical score based on the electrostatic interactions of each amino acid residue. MEGADOCK was used to evaluate docking pose prediction and PPI screening performances from 44, 120 and 176 complexes from ZLAB Benchmark 2.0 [44] and 4.0 [45], with bound and unbound protein-complex data. The PPI searching and analysis was feasible, taking into account 3D structures at the interactome scale, by parallelizing with MPI (Message Passing Interface) and OpenMP (Open Multi-Processing) and running the proposed tool in an advanced computing environment with several thousands (192 to 4,608) of CPU cores.

Several other approaches based on protein structure have been proposed. Wass *et al.* applied standard docking programs to predict PPIs [46]. Petsalaki *et al.* used known protein-protein complexes to predict binding sites [47]. Finally, Baspinar *et al.* presented the PRISM (Protein Interactions by Structural Matching) web server and repository, intended to predict a structural model of the complex of two protein structures. This prediction method is based on other matching template interfaces, taking into account structural similarity and evolutionary conservation [48].

Machine Learning Techniques for Structure-based Methods

Murakami and Mizuguchi proposed a computational method for the prediction of PPIs by using characteristics derived from known homologous PPIs. This method was designed to predict the interaction between two proteins of unknown structure. Their aim was to improve the discrimination power of homology-based PPI prediction by applying a machine learning (ML) algorithm termed Averaged One-Dependence Estimators (AODE) [49]. The AODE algorithm was described as a variant of the

naïve Bayes classifier, and was trained using three features for classification: sequence similarities, statistical propensities and the sum of edge weights along the PPI network. The data set for training and testing considered a ratio of 400:1 negative to positive interactions. That is, 5,000 positive pairs and 2,000,000 negative pairs from a total of 2000 proteins. The authors hypothesized that a given protein pair would have a higher chance to interact if homologous proteins were close to each other in a known PPI network, even if the homologous protein pair was not known to interact directly. The AODE method trained on the three features was named PSOPIA (Prediction Server of Protein-protein InterActions) and it was freely available through a web portal allowing users to submit up to ten protein pairs for prediction.

Birlutiu *et al.* presented a Bayesian framework [50] that combines information related to proteins and the interactions between them along with information on the network topology. In this work the naïve Bayes classifier was used to express the likelihood of interaction based on the features for a given protein pair. Each protein pair was characterized by a feature vector of 27 dimensions based on Gene Oncology (GO) annotations, sequence similarity, co-occurrence in tissue, and domain interactions, among other features.

Other prediction methods that used ML techniques have been proposed, such as the one presented by Zhang *et al.* [51], that used 3D structural information and a naïve Bayes classifier to predict PPIs. The algorithm developed was comparable in accuracy to high-throughput experiments. Finally, Xu and Guan [52] presented a method for the detection of protein complexes based on biological process annotations. In this method, the authors used a k-means clustering algorithm to compute the similarity of biological functions in proteins.

3.1.2 Methods Based on Primary Protein Structure

Among the computational methods that have been developed to predict novel PPIs, sequence-based methods are the most universal as they are not dependent on additional information about the proteins [53] that most of the time is not available. Hu and Chan developed an algorithm to predict protein interactions termed variable-length associative sequential pattern discovery (VLASPD). According to the authors, sequence-based prediction methods only consider segments of a specific length to decide if two proteins interact. They presented VLASPD based on the idea that, if

segments of different length can be considered at the same time, the interactions between proteins can be predicted with a higher level of accuracy. This method aimed to determine if patterns of different length could be used for PPI prediction. Given a database of protein sequences, the first step consisted in identifying the frequent sequence segments (FSSs) of different length. Next different combinations with the presence and absence of these FSSs were used to form various associative sequential patterns (ASPs). The ASPs occurring more frequently were then identified as significant associative sequential patterns (SASPs). If one SASP was found in a pair of proteins, it was considered as evidence to support the existence of an interaction. The extent of this evidence was then computed for all SASPs to obtain weighted SASPs (wSASPs). The probability of the PPI was decided by evaluating the number and significance of the wSASPs found. The effectiveness of the method was tested using several sets of real data and VLASPD performed better than other classification methods [54].

Machine Learning Techniques for Sequence-based Methods

Martin *et al.* implemented a method for PPI prediction based on sequence information and experimental data, using a support vector machine (SVM) classifier [55]. To confront the problem of representing amino acid sequences of different length as vectors, a signature molecular descriptor was used. Such a signature would represent one amino acid and its neighbors. This proposal was benchmarked against previous methods and achieved similar performance. It had an accuracy of 80% when applied to *S. cerevisiae* and *H. pylori* species. It is worthy of mention that the authors of this study recognized that using their method with a realistic data set of unbalanced data would be problematic.

By using a learning algorithm based on SVM combined with a kernel function and a conjoint triad descriptor, Shen *et al.* developed a method for PPI prediction using only the information of protein sequences [56]. Different to other prediction methods that fall into this category, this method was used to predict three types of PPI networks: one-core networks, built by a core protein interacting with other proteins; multiple-core networks, consisting of several core proteins interacting with other proteins; and crossover networks, consisting of several networks built from the previous two types. In this approach, each protein sequence was represented by a

vector space consisting of features of amino acids. Then each pair of interacting proteins was represented by concatenating the corresponding vectors. The 20 different amino acids were clustered into several classes and the conjoint triad method represented any three amino acids as a unit, by considering the properties of each amino acid and its contiguous amino acids. For the data set preparation, known PPIs were collected from the Human Protein References Database (HPRD) (www.hprd.org) and the training set consisted of 16,243 protein pairs for each set, positive and negative.

Guo *et al.* proposed a sequence-based method for PPI prediction by using SVM combined with auto covariance (AC) [53]. Seven physicochemical properties of amino acids were used. Also, auto cross covariance (ACC) was used to build uniform matrices, avoiding the unequal length of vectors produced by protein pairs of different length. ACC produces two variables, AC (auto covariance) and CC (cross covariance), but only AC was used to avoid a large number of variants. In the previously described method proposed by Shen *et al.* [56], to represent any three contiguous amino acids as a unit, the properties of each amino acid and its neighboring amino acids were considered. In this method AC was used to cover the interactions between one amino acid and its 30 neighboring amino acids. For model construction, the positive data set contained PPIs that had been experimentally validated by two different methods. The negative data set was generated by randomly pairing proteins that appeared in the positive data set, and assuming that proteins located in different subcellular compartments do not interact. The final data set consisted of 11,886 yeast protein pairs, half from the positive set and half from the negative set: three-fifths of each set were used for training and two-fifths for testing.

Yu *et al.* highlighted the importance of elucidating PPIs with an unbalanced ratio for interacting and non-interacting pairs, considering that this ratio is highly unbalanced in nature [57]. Similarly to the previously described method proposed by Shen *et al.* [56], the authors used the triad frequency of amino acids to represent the protein sequences as a feature vector. Additionally, taking into account amino acid distributions, a probability-based mechanism was implemented to estimate the significance of triads. For the classification, unbalanced data sets were used with different ratios that went from 1:1 to 1:15, positive to negative interactions. This study showed how the accuracy of the predictor decreases when the positive-to-negative ratio is more unbalanced.

Zhang *et al.* presented the tool PPI-PKSVM (Protein–Protein Interaction - Pairwise Kernel Support Vector Machine), with two methods of amino acids feature extraction, that yielded similar prediction accuracy [58]. One method was DFPCA, that used the frequency of the distance (DF) between two successive amino acids to represent the sequence, grouping amino acids by four physicochemical properties. DF was combined with the statistical method Principal Component Analysis (PCA) to reduce the dimension of the vector. The second method was an amino acid index distribution (AAID) representing the protein sequences with the physicochemical value of each residue, statistical information, sequence-order information, and serializing these features in a combined feature vector. Besides the two kinds of feature extraction approaches, pairwise kernel function and SVM were used for the prediction of PPIs. The PPI data was collected from the PRISM (Protein Interactions by Structural Matching) [59] server and from the PDB (Protein Data Bank) database [60]. Validation was performed with 10-fold cross-validation and the highest prediction accuracies of DFPCA and AAID were 93.95% and 94% respectively.

You *et al.* proposed a multi-scale local descriptor (MLD) scheme to extract features from a protein-sequence. This method was able to capture multi-scale local information by varying the length of segments of the protein sequence [61]. The prediction task was performed with the Random Forest (RF) classifier. The learning capabilities of the RF model allowed the achievement of a sensitivity of 94.34% at a precision of 98.91% when applied to PPI yeast data. The main idea surrounding this work was that contiguous amino acid segments of different length played a critical role in determining protein interactions. For each contiguous region three types of descriptors were used: composition, to measure the number of amino acids of a particular property; transition, to measure amino acids of a particular property followed by amino acids of another property; and distribution, to measure the chain length in a location associated with amino acids with a particular property. The performance was better for this method when compared against other similar methods, such as the previously described method proposed by Guo *et al.* [53].

Zhou *et al.* presented a framework for the prediction of PPIs from protein sequences [62]. Different to methods that predict PPIs by using amino acid composition (ACC) and aiming to improve the overall accuracy, the method termed LELM (Low-rank approximation-kernel Extreme Learning Machine) besides using all the possible subsequences of length k (*k-mers*) considered sequence order information. The main

prediction tasks were divided into three steps. First, transforming each protein sequence into a matrix, with both ACC and sequence order information. Next, using the method termed LRA (low-rank approximation), every row vector was extracted to numerically represent each sequence. Finally, the probability for PPI was performed with a kernel-ELM predictor.

3.2 Computational Protein Design: Advances and Challenges

Along with work on the prediction of PPIs, there is another growing field in biochemistry, that of protein engineering. In contrast to the prediction of PPIs, where the main task is to predict the interaction between a pair of proteins given their sequences or any other information (e.g., 3D models), in protein engineering the objective is to design a synthetic protein. A special review described the variety of research presented during the inaugural *Protein Engineering Canada Conference* highlighting the importance of protein engineering [63]. As mentioned by the author Chica, protein engineering efforts still depend largely on the structure and functions of known proteins. The author also mentioned the importance of computational protein design methods, due to their ability to evaluate amino acid sequences on a scale impossible to achieve with conventional experiments.

Synthetic proteins can be defined as those proteins not found in nature, designed to perform a specific function in living organisms. These proteins are intended to modify or inhibit predetermined protein functions and have been the focus of a vast number of research projects, especially when they are complemented by methods for computational protein design. Extensive reviews about the computational design of proteins have defined the protein design problem as the inverse of the protein folding problem (e.g., [11, 12, 37]). The protein folding problem consists in predicting the 3D structure of a protein; in other words, how the protein folds into a specific shape. Protein design is considered the inverse folding problem, as the objective is to design a specific protein structure or amino acid sequence with specific characteristics. Since protein folding is a recurrent topic in protein design, computational tools for the modeling of protein structure and protein-protein docking software are usually discussed in these studies.

Some aims of computational protein design are understanding protein folding and

stability, enhancing the function of natural properties and developing novel proteins [64]. The ultimate goal in computational protein design is to find at least one sequence able to fold into a predefined folded structure [12], so that it interacts (docks) with determined proteins. A successful computational protein design is determined by three factors: accurate structure modeling, the stability of the designed protein, and how well the interactions are optimized with target molecules [65].

3.2.1 Redesign of Proteins and *De novo* Protein Design

Computational protein design can be divided in two general areas, redesign of proteins and *de novo* protein design [66]. The objective in the redesign of proteins is to modify or improve the function of natural proteins, whereas in *de novo* protein design the objective is to design novel proteins (not found in nature) with a specific function. The idea of the *de novo* design of proteins has been around for a few decades. Baltzer *et al.* emphasized the understanding of biomolecular structure and function as fundamental for the design of *de novo* proteins [67].

In a study focused on the redesign of metalloenzymes [68], the objective was to improve or design novel properties into existing proteins. On the other hand, *de novo* protein design was defined as an approach where the proteins were similar to native proteins but designed from scratch. The authors emphasized that, different to *de novo* protein design, the advantage of protein redesign was that the modified native proteins were more stable and adaptable to changes; this represents a challenge for *de novo* proteins.

Redesigned or re-engineered proteins have been shown to be more stable than the original native proteins. However, with *de novo* protein design it is possible to develop proteins with predetermined structures, properties or even functionalities, as was described by Samish *et al.* [69].

In a study about the advances in computational protein design [65], the authors reviewed computational approaches for both redesign and the design of novel proteins. As described in this study, efficiency and reliability are still objectives difficult to attain in the computational design of proteins. Factors such as the high complexity of real biological systems, the lack of information on known protein structures, and in addition the lack of communication on failed designs, were also highlighted.

3.2.2 Advances in *De novo* Protein Design

De novo protein design through the use of computational methods was explored more than 20 years ago in [70]. Bryson *et al.* described protein design as the technique to design a protein with a specific structure and function. Since then, important advances in the field have arisen for different applications. For example, computational protein design has been successfully used to identify novel molecules with medical applications to treat several diseases, as described in [37].

One of the big challenges in novel protein design is the lack of known three-dimensional structures for the great majority of natural proteins; it is difficult to predict protein structures when the structures of similar proteins are not available. The development of methods for the prediction of structure, sequence and function of proteins has made possible progress in the field of protein design [64]. Kang and Saven reviewed computational methods for protein design and found interesting approaches such as computationally guided mutations to stabilize proteins, or computationally designed variants of membrane proteins to facilitate studies of structure and function.

In [11], Khuory *et al.* reviewed computational approaches for *de novo* protein design. This overview of successful computational applications for protein design focused on specific problems. The described applications were organized by different targets, such as cancer, HIV, Alzheimer's disease and antibody therapeutics. The authors of this study also defined the protein design problem as closely related to the protein folding problem. Therefore, they emphasized the challenges for protein structure prediction and listed a few methods that aimed to solve this problem.

The *de novo* design or redesign of proteins with the ability to act as PPI inhibitors is one of the proposed applications of protein engineering. The development of inhibitory proteins could be the basis for therapeutic treatments. In [71] Villoutreix *et al.* studied the importance of drug-like PPI modulators. According to the authors, around the year 2000 academic and private laboratories started to research low molecular weight, drug-like compound modulators of PPIs. Since then, there have been further efforts to design PPI modulators and many databases as well as computational tools to assist in drug discovery have been developed [71]. Root *et al.* explored the idea of designing a small protein intended to bind a specific region of a protein, aiming to act as an inhibitor of the HIV-1 virus [72]. Years later, Saito *et al.* developed a motif programming method to generate artificial proteins with desired functions [73]. Some examples of PPI inhibitors that have reached clinical

trials were reviewed in [74]. Also, Stranges and Kuhlman [75] compared successful *de novo* protein interface designs. The authors investigated the reported designs with a popular molecular modeling program.

Computational protein design is often performed using existing software, such as protein-protein software docking tools, but other methods such as evolutionary algorithms (EA) have been used in several studies. *De novo* drug design tools applying EA techniques were reviewed in [76]. Vasundhara *et al.* described what they called computer-aided drug design (CADD) as a technique to speed up the process of drug design and discovery. In this study, the term *de novo* is used for the design of drug molecules (unique in nature) from scratch. The study describes EA as popular heuristic algorithms that are of help in finding an optimal solution, due to their probabilistic, stochastic and randomized nature. Among the variety of EA techniques, the genetic algorithm (GA) is referred to as the most popular. The review includes a comparison of different evolutionary techniques used for *de novo* drug design.

When speaking about synthetic biology, protein engineering or protein design, today it is almost impossible to avoid mentioning computational techniques. A combination of biochemistry and computer science provides a major opportunity to overcome the various challenges present in *de novo* protein design [77]. In what follows is described some relevant work related to protein design using computational methods.

3.2.3 Applications of Computational Protein Design

In [78] OptCDR (Optimal Complementarity Determining Regions) was proposed as a computational method for the *de novo* design of the binding portions of antibodies with a high specificity and affinity against a targeted antigen. This method first selects the antibody complementarity determining regions (CDRs) most likely to bind the antigen. Then, it performs thousands of iterations looking to improve the interaction energy between the CDRs and the antigen.

Hot spots are a small subset of residues located on the protein-protein interface and, according to the study by Guo *et al.*, they are the ones that contribute the majority of the binding free energy [79]. Fleishman *et al.* presented a hot spot-based computational method for protein design to target a conserved region on the stem of *influenza hemagglutinin* (HA) [80]. The authors used protein-protein docking software to: first, identify a favorable target surface; and next, perform experimental

manipulation on a selected set of protein structures looking to maximize the compatibility with the identified regions. After the computational design, experimental validation by affinity maturation was performed to identify weaknesses in the modeled proteins. Two of the designed proteins were shown to bind HAs with low nanomolar affinity. One of those proteins was found with inhibitor properties for HA. The crystal structure of the other protein in the complex with HA was revealed to have a nearly identical binding interface to that in the designed model. The results of these experiments suggested that *de novo* computational design was feasible for the design of antiviral proteins.

Based on the previously mentioned hot spot-based method, Strauch *et al.* developed a methodology to design pH-sensitive proteins. The use of computational methods was similarly based on hot spot residues and protein-protein docking. The designed protein was shown to be highly stable, resistant to high temperatures and able to highly express in bacteria [81].

Lastly, Voet *et al.* used protein-protein docking software to design symmetrical β-propeller proteins, a protein family with diverse functions, such as enzymatic activities and protein-protein interactions. To validate their approach, the authors created a sixfold symmetrical, β-propeller protein and used the *X-ray crystallography* technique to validate the structure. According to the authors, their approach would be applicable to other protein templates [82].

Chapter 4

InSiPS, the *In-Silico* Protein Synthesizer

4.1 Introduction

In Chapter 3 fundamental issues related to this research were described, namely, computational methods for (i) the prediction of protein-protein interactions (PPI) and (ii) for protein design. This chapter reviews the computational method used in the research: InSiPS, the *In-Silico* Protein Synthesizer. As described in [39], the term *in-silico* refers to techniques performed on a computer or via computer simulation. Also, as the name suggests, InSiPS is a computational method for protein design. The computational tasks performed by InSiPS are complemented by the use of PIPE, a protein-protein interaction prediction engine. The following sections will describe both the PIPE and InSiPS methods, as well outline information specific to each of them: the input data, the internal algorithm, the output data and, finally, how the resultant data is validated.

4.2 PIPE

PIPE, a protein-protein interaction prediction engine developed by Pitre *et al.* at Carleton University [1], is a computational sequence-based method for the prediction of PPIs. The fundamental idea behind this method is that protein pairs can be predicted to interact based on the co-occurrence of short polypeptide sequences found in a data set of known PPIs that have been scientifically validated. Of the five general groups of computational methods for the prediction of PPIs described in Section 3.1, PIPE falls into the category of primary protein structure, as it only needs the primary sequence information of proteins. Methods from other categories,

such as those described in Section 3.1.1, require detailed knowledge of proteins or information about homologous proteins; however, this information is still limited for the majority of proteins. One of the advantages of PIPE, then, is that it is possible to evaluate protein pairs without complex protein knowledge such as their secondary or three-dimensional structure.

The first version of PIPE was used to predict 100 positive interactions from yeast *S. cerevisiae* proteins. The computation time for these predictions was close to 1000 hours. A set of 100 negative interactions was similarly evaluated, which was expected to be reported as non-interacting proteins. The results indicated 61% for sensitivity and 89% for specificity, with an overall accuracy of 75%. The success rate was comparable to that obtained by *in vivo* experiments. Also, in a study about sequence-based prediction methods [83], PIPE was evaluated against three other computational methods within the same category: PIPE was shown to be better in terms of recall-precision and performance.

4.2.1 Input Files and Database

As a purely sequence-based method, PIPE uses only information from the primary protein structure: to predict only the protein sequences and a list of known PPIs are required. The two plain text files described below are required as input.

Plain Text Input Files

- **Protein sequences file.** This file contains the list of proteins, where each protein is represented by the identifier (id) followed by the amino acid sequence. The proteins that PIPE will evaluate for the predicted likelihood of interaction, as well as the proteins listed in the known protein-protein interactions file, must be in this file.

- **File of known protein-protein interactions.** This file contains the training data, consisting of the list of protein pairs that are known to interact. Each interacting pair is represented by the id of each protein. Aspects, such as the size and truthfulness of the data contained in this file, will have a direct impact on the PIPE predictions.

Preprocessed Input Data Files

Based on the information described before, a preliminary processing evaluates the input files and generates the data below.

- **PIPE database.** A database with one file for each individual protein in the protein sequences file.
- **Interaction list file.** A file to specify the interaction list for prediction; that is, which proteins will be evaluated against which other proteins. The default process generates a default interaction list for an all-to-all run. An all-to-all PIPE run means that every protein in the initial protein sequences file will be evaluated against the rest of the proteins in this file. Considering an all-to-all run for twenty thousand proteins, this implies evaluating almost 200,000,000 pairs. The interaction list file can be modified to evaluate specific protein pairs; for example, to only evaluate protein A against the rest of the proteins.
- **Known interactions graph file (G).** An interaction graph file (G) specifying through indexes the known interacting partners for each protein.
- **Parameters.** A file with calculated parameters such as the total number of proteins in the protein sequences file, the number of known pairs, the total number of pairs (combinations) to evaluate, the size of the database file, and the maximum number of neighbors (interacting partners) for a protein.

4.2.2 The Basic PIPE Algorithm

The PIPE algorithm uses a sliding-window approach to find co-occurring fragments of query proteins A and B on a list of known PPIs. The final goal is to predict the likelihood of interaction for proteins A and B. The four steps of the basic PIPE algorithm, illustrated in Figure 4.1, are described below.

Step 1

The input information for the algorithm consists of: (1) a list (L) of the known protein interactions; (2) the known protein interactions graph file (G) where each protein sequence in (L) is represented as a node; (3) the sequences of proteins A and B that will be evaluated to predict their likelihood of interaction.

Step 2

Protein A is divided into overlapping fragments of size w. For every fragment a_i on A, PIPE looks for similar fragments a'_j on A' of size w in L. The comparison of fragments is by using a substitution PAM120 matrix (Point Accepted Mutation matrix) to give a score according to the probability that each amino acid on a_i is replaced by the corresponding amino acid on a'_j (through an evolutionary process). If the total score S_{PAM} from the comparison of two fragments is larger than a given threshold, then the fragments are considered to be similar. For every similar fragment, the known interacting partners of A' are stored in a list of neighbors (R). The size w used to compare fragments is 20 amino acids.

Step 3

The previous step is repeated for every fragment of size w on B, looking for similar fragments on B' of size w in R, the list of neighbors. The comparison of fragments is also done through the PAM120 matrix to determine whether they are similar. Next, a result matrix (M) for proteins A and B is generated, where each row represents a fragment a_i on A and each column a fragment b_j on B. A score s on each cell of the result matrix indicates how many times the fragments a_i and b_j co-occurred in G, the interaction graph file indicating known protein pairs.

Step 4

The last step produces a 3D plot to visualize the result matrix (M) for proteins A and B. The rows and columns on the plot represent every fragment of A and B, and the elevation represents the co-occurrence score s. The maximum score in the result matrix is the final PIPE score that represents the predicted likelihood that proteins A and B interact.

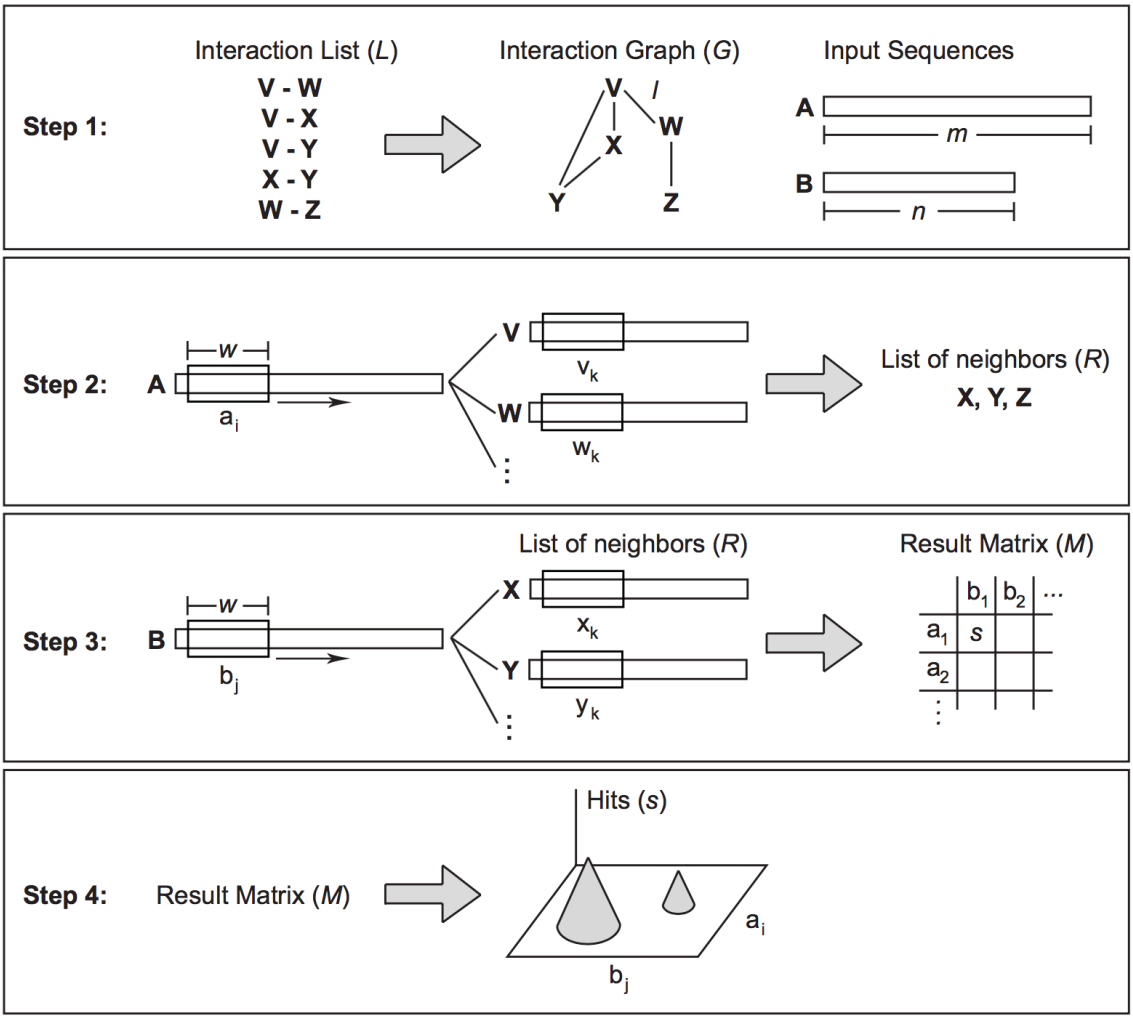


Figure 4.1: Illustration of the basic PIPE algorithm (from [1]).

4.2.3 The PIPE Output File

The PIPE output file, illustrated in Figure 4.2, contains the predicted interaction score — the PIPE score — for the protein pairs provided as input. Besides the output file with the predicted interaction score for the evaluated protein pairs, PIPE generates the result matrix (M) file for every protein pair, to indicate the co-occurrence score for every fragment on A and B found in interaction graph (G).

1		2	3	4			
protein_a	protein_b	PIPE_score	running_tim	site1a_start	site1a_end	site1b_start	site1b_end
YLR106C	YML029W	0.999999	3.75196195	4117	4151	791	812
YBL100W-B	YOR192C-B	0.9999987	3.08115196	0	142	0	142
YDR261W-B	YLR410W-B	0.9999987	3.01191187	0	142	0	142
YKR054C	YNL161W	0.99999867	1.64730811	2405	2435	203	260
YBR102C	YKR054C	0.99999866	1.43200207	669	692	2405	2436

Figure 4.2: Example of the information provided in the PIPE output file. Each line in the output file provides the following information: (1) the ids of the evaluated protein pair; (2) the PIPE score, which is the predicted likelihood that the pair of proteins interact; (3) the running time that it took to process the given pair; and (4) the predicted interaction sites for every protein pair, where the predicted site for each protein — a, b — has the first and last amino acid position (i.e., site1a_start, site1a_end and site1b_start, site1b_end).

4.2.4 Improvements to Further Versions of PIPE

Since the appearance of the first version [1], important improvements have been made to the PIPE algorithm aiming to improve primarily two aspects: the processing time, so as to enable production of a proteome-wide, all-to-all run; and the specificity value, to reduce the high rate of false positives (FP) produced in a proteome-wide evaluation. The subsequent version of PIPE represented a considerable improvement in computation time (16,000 times faster) and in the specificity value (around 99.95% and 0.05% FP). It was with this newer version that the first all-to-all prediction with a sequence-based method was performed with yeast data, identifying 29,589 high confidence predictions from nearly 20,000,000 possible pairs [84]. Later, PIPE-Sites put a major emphasis on the accurate prediction of binding sites with a high specificity. By analyzing the predicted interaction sites, the authors found highly re-occurring subsequences that may represent novel binding motifs [85].

MP-PIPE: The Parallel Implementation of PIPE

The most recent version of PIPE presented by Schoenrock *et al.* is termed MP-PIPE. This is a massively parallel implementation of PIPE, specialized in high performance predictions of PPIs [8]. This version includes improvements to the basic sequential algorithm: the amino acid representation was changed from character to binary; also the sliding-window process was modified to make use of incremental updates, improving the speed and the pre-computation of all possible protein fragments to compare. An important objective was to have a flexible and portable parallel system to be used in different parallel architectures. The two-level master/slave model of MP-PIPE consists of a single scheduler process in charge of the list of protein pairs to be processed. The scheduler distributes packets with a small number of protein pairs, so that each worker process executes the basic PIPE algorithm for the protein pairs received. Efficient use of memory and load balancing were crucial aspects for this implementation.

The fact that MP-PIPE is aimed at executing a large number of processes for the prediction of PPIs makes it a unique tool. It allowed the performance of the first ever scan of the entire interaction network of human proteins, with 22,513 proteins (253,406,328 pairs to examine) and 41,678 known protein pairs as input. This represented a computational challenge that took three months of full-time computation on a dedicated cluster with 50 nodes, 800 processor cores and 6,400 hardware-supported threads. MP-PIPE was able to predict 172,183 protein interactions at a specificity of 99.95% (0.05% false positive rate) from which 130,470 interactions would be novel (not known), setting new paths for biologists to investigate.

Further research with the parallel implementation of PIPE showed that co-occurring polypeptide subsequences for interacting protein partners appeared to be conserved across organisms [86]. Schoenrock *et al.* later worked on demonstrating the usefulness of the MP-PIPE predictions data through analysis and diverse experimental techniques [87].

4.2.5 Leave-One-Out Cross-Validation Tests

The process for the calculation of PIPE's accuracy was described in Schoenrock's doctoral thesis [88]. The following is a brief explanation of this validation. To measure the accuracy of PIPE predictions for a given organism, the known interacting protein

pairs are used for the positive set. Due to the lack of available data for non-interacting protein pairs, a random set of protein pairs (e.g., 100,000 for human) serves as the negative set. With these two sets of protein pairs a leave-one-out cross-validation is performed; that is, a pair is removed from the PIPE input database to see if PIPE predicts the positive or negative interaction. After doing this validation for the positive and negative sets, the results are put together and ordered by (PIPE) score. Then, for a given score cutoff, the sensitivity and specificity values are calculated from the number of true positives/negatives and false positives/negatives produced by PIPE at that operating point (score). To consider a PIPE prediction a high confidence interaction, the operating score is set as low as possible while still achieving a specificity (true negative rate) of 99.95%. This high specificity means an extremely low false positive (FP) rate of 0.05%, and hence a high confidence interaction.

Sensitivity and Specificity

The way the accuracy of PIPE is measured is through sensitivity and specificity. The measure of sensitivity is calculated as *[TP / (TP + FN)]* (*where TP is True Positive and FN False Negative*) and specificity is calculated as *[TN / (TN + FP)]* (*where TN is True Negative or non-interaction and FP False Positive*). In the first version of PIPE, from a positive set of 100 yeast PPIs, 61 interacting pairs (true positives) were successfully reported and hence 39 non-interacting pairs (false negatives); that is a sensitivity of 61%. For the negative set of 100 PPIs, PIPE successfully reported 89 non-interacting pairs (true negatives) and hence 11 interacting pairs (false positives); that is a specificity of 89%.

In a proteome-wide all-to-all run for yeast proteins with approximately 20,000,000 pairs to evaluate, around 33,000 would be assumed to interact taking into account the 600:1 ratio of negative to positive interactions proposed by Pitre *et al.* and as described in Section 3.1. With the above sensitivity and specificity, PIPE would correctly predict 20,130 true interactions (TP) and 17,770,630 true non-interactions (TN). However, it would also incorrectly predict 12,870 false non-interactions (FN) and 2,196,370 false interactions (FP). The high number of false positives would make the results worthless. A specificity of 99.95%, then would imply a low sensitivity, thus a lower number of detected true interactions (TP); but it would also increase the confidence in the reported true interactions, with a very low likelihood to be a false positive result.

4.3 InSiPS

The In-Silico Protein Synthesizer (InSiPS) is a massively parallel computational tool for protein design, developed by Schoenrock *et al.* at Carleton University [7]. InSiPS has the ability to design synthetic protein sequences (not found in nature) predicted to interact with a specific target protein, without affecting other non-target proteins. The non-target proteins are normally located in the same cellular component as the target protein and thus an interaction with the novel synthetic protein is not desired. This novel method has great potential for designing inhibitory proteins, with the objective of blocking a protein-protein interaction and modifying a specific cellular function as the final goal.

One of the main characteristics of InSiPS is that it is purely sequence based. As reviewed in Section 3.2, in contrast to InSiPS, other approaches for protein design require knowledge of secondary or 3D protein structure. This makes InSiPS a unique tool, as the non-availability of complex protein data is one of the major obstacles with computational methods for protein design. InSiPS was previously used to design novel proteins to target yeast *S. cerevisiae* proteins. Wet-lab experiments using living yeast *S. cerevisiae* cells demonstrated that the InSiPS-designed synthetic proteins did have an inhibitory effect on the corresponding proteins when exposed to them [7].

InSiPS was implemented on the IBM Blue Gene/Q cluster (BGQ). As a massively parallel computational tool, InSiPS requires high processing capacity to identify a synthetic protein able to interact with a target protein without affecting other proteins. It was shown that this BGQ cluster efficiently made use of all computational threads on each node, scaling almost linearly up to 1024 nodes, in performing the InSiPS computations.

4.3.1 InSiPS Algorithm Overview

The final goal of InSiPS is to design a novel synthetic protein sequence (also referred to as a new sequence) predicted to interact with a specific target protein and predicted not to interact with a defined list of non-target proteins. The InSiPS implementation consists of a two-level master-worker/all-workers parallel algorithm. The algorithm is a combination of a genetic algorithm (GA) with a protein-protein interaction prediction method that is based on mining a large database of known PPIs. One of the main tasks in the prediction of PPIs relies on the massively parallel implementation

of PIPE (also known as MP-PIPE). In the following are described the main characteristics of the master process and the worker processes respectively, as well as how the genetic algorithm works together with PIPE.

The Master Process

The master process is implemented using the Message Passing Interface (MPI) and it is responsible for all the genetic algorithm tasks. The InSiPS master process loads all the relevant data, the known protein-protein interactions file (L), the PIPE database, and the defined target and non-target proteins. It then broadcasts all this information to the worker processes. The master process is also responsible for maintaining load balancing across all the worker processes, by working in an on-demand fashion.

The Worker Processes

The InSiPS worker process algorithm implements an all-workers model and all the internal parallelization is implemented in OpenMP. The worker processes are responsible for the PPI prediction tasks. They do not load any data from disk; instead all data is sent over the network from the master process. Hence the worker processes receive the information to produce PIPE predictions (algorithm is described below). Once the assigned tasks have been completed, the worker process notifies the master process by sending a new work request and this loop continues until the master process lets the worker process know that there is no more work to do.

The Genetic Algorithm (GA)

The InSiPS algorithm is illustrated in Figure 4.3. The master process is responsible for the GA operations. It starts with a predetermined number of new protein sequences (e.g., 1000) of random amino acids. This is the initial set (first generation) of the population. Then the loop is started; every new protein sequence from the population is evaluated by a worker process. The worker processes receive the set of new sequences (one by one) to produce the predicted interaction scores against the target and non-target proteins through PIPE (algorithm described in Section 4.2.2). When all the predictions have been calculated for all the target and non-target proteins, the worker process notifies the master process by sending a new work request. Once all the new sequences have been evaluated (against target and non-targets),

the master process determines the fitness value for every sequence with the following fitness function:

fitness(seq) = {*1 - MAX[PIPE(seq,non-targets)]* } × *PIPE (seq, target)*

The fitness value of the new sequence is equal to 1 minus the maximum PIPE (interaction) score predicted for the new sequence against the non-targets, multiplied by the PIPE score predicted for the new sequence and the target. Thus, the fitness value indicates how likely the new sequence is to interact with the target and not to interact with the non-targets. By having the fitness value for all the new sequences of the population, the best candidate sequence (i.e., with the greatest fitness value) can be identified. Following this, the GA will create a next generation of protein sequences based on the current population, applying one of three operations: *copy*, leaving the exact same sequence; *mutate*, changing the sequence given a predetermined low probability for changing each amino acid; and *cross over*, joining a portion of a new sequence A with a portion of another new sequence B. Once the new population of new sequences is complete, the loop of the algorithm will start again with the prediction of interactions (through PIPE), and will continue to improve the fitness of the new sequences until the termination criteria are met.

The objective of the GA is to optimize a fitness value and to determine which of the candidate protein sequences is the best option, in terms of having the highest PIPE score for the target protein and the lowest maximum PIPE score for the non-target proteins. Therefore, the best candidate protein sequence is the one with the greatest chance to interact with the target, and the lowest predicted likelihood of interaction with the non-targets.

Genetic Algorithm Parameters

The following describes some important parameters for the InSiPS GA algorithm.

- **Sequence length.** The length (number of amino acids) of the new sequence that will be generated. All the sequences in the population will be of this length.
- **Population size.** The size of the pool of synthetic protein sequences. This value defines the number of new sequences that will be generated for each generation.

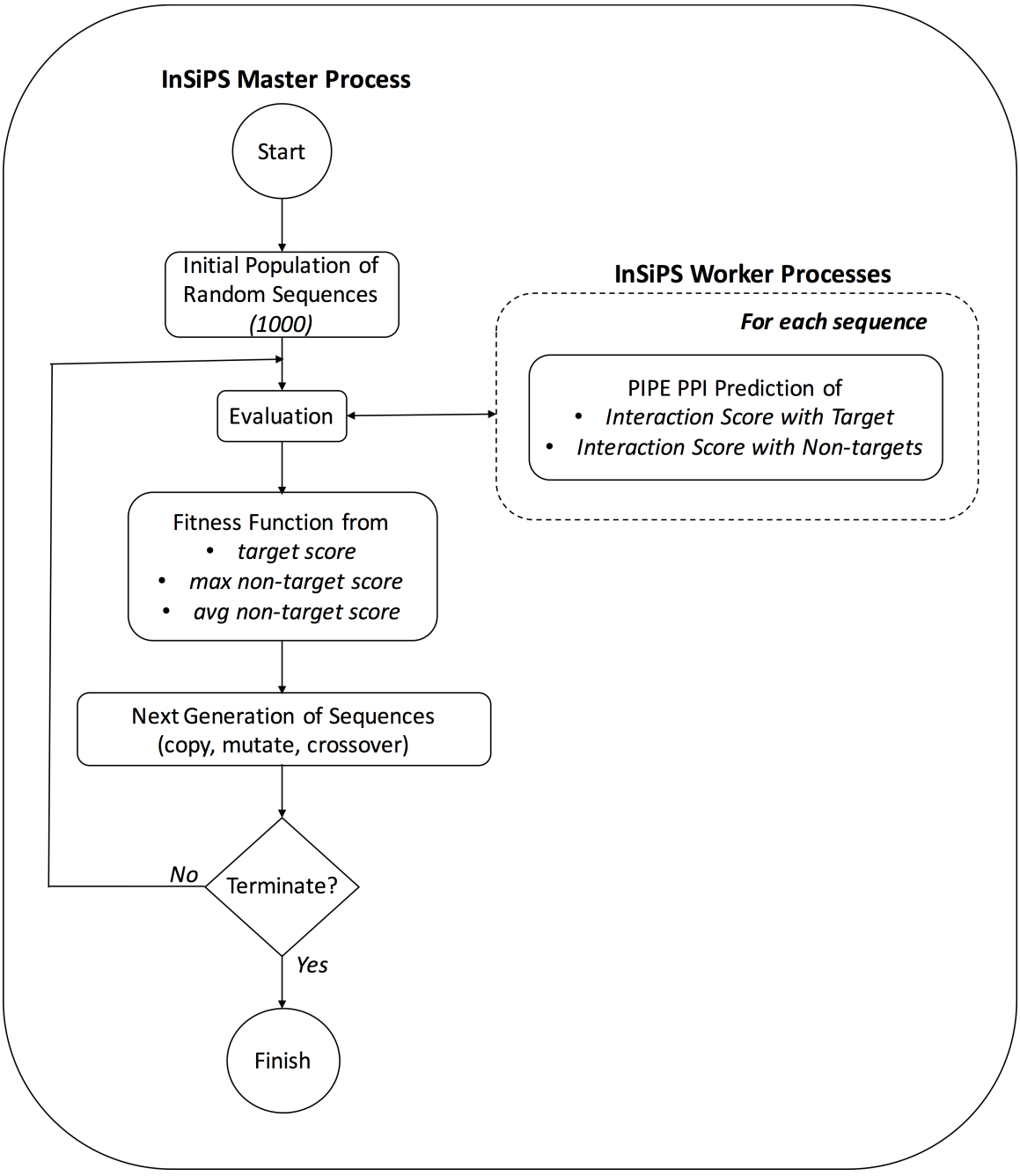


Figure 4.3: The InSiPS two level master-worker/all-workers parallel algorithm. Adapted from [7].

- **Probability for the standard GA operations.** The probabilities to perform each one of the GA three standard operations: *copy, mutate* and *cross-over*. Also, the probability (*mutate-aa*) to change each amino acid when performing the operation of *mutate*.

- **Number of generations (*minimum, maximum, unchanged count.*)** The *minimum* and *maximum* number of generations that InSiPS will run looking to improve the fitness value. Also, the minimum number of generations (*unchanged count*) that InSiPS will run when the fitness value is not improved (i.e., is not higher than in previous generations). The first criteria for the InSiPS GA algorithm is to meet the *minimum* number of generations; next, the *unchanged count* and *maximum* values. The criteria to terminate an InSiPS execution based on the number of generations can be described as follows: the number of generations is less than the *maximum* value, it has reached at least the *minimum* number of generations and there has been no improvement in the fitness value for a determined number of generations (*unchanged count*).

- **Fitness cutoff.** Maximum limit for the fitness value.

- **Random seed.** Flag (true/false) to indicate whether the initial population of new sequences will be made of random amino acids (false) or a random seed will be provided as input (true).

- **Number of computational threads.** Number of possible threads used by each worker process. This value will depend on how many threads are supported by the machine running the process.

4.3.2 InSiPS Input Files and Database

InSiPS relies on PIPE to perform the tasks of PPI prediction. Therefore, it uses the two plain text input files (protein sequences and known PPIs) and preprocessed input data described for PIPE in Section 4.2.1. Additionally, the information described below is required by InSiPS to specify the target and non-target proteins for which the new sequence will be designed. Note that the proteins in both files of target and non-targets must be part of the PIPE protein sequences file.

Plain Text Input Files

- **Target.** A text file with the id of the protein for which InSiPS will create a novel sequence predicted to interact with it.
- **Non-targets.** A text file with the list of all the protein ids for which the novel sequence created by InSiPS will be predicted not to interact. That is, the proteins for which the new sequence should have the lowest predicted interaction score.

Preprocessed Input Data

Based on the preprocessed PIPE database, a preliminary processing will generate the following file required as input by InSiPS.

- **Unified database.** A unified database file containing all the individual files from the preprocessed PIPE database. The input data are the PIPE database and the known interactions graph file (G).

4.3.3 InSiPS Output Files

As mentioned earlier, the final goal of protein synthesis with InSiPS is to find one new protein sequence with the highest fitness value. Besides the new protein sequence of amino acids with the highest predicted interaction score with the target and the lowest predicted interaction score with the non-targets, other results from the InSiPS calculations are provided in two files: the console file and the results file. The two files are described below and have information from only those generations where the highest fitness value obtained was higher than the one obtained in the former generation. Thus, the last result in each file will be from the last generation with the highest fitness value.

Console File

The console file contains the total *runtime*, the *hit generation* (number of the generation with the synthetic protein sequence of highest fitness value), and the final number of generations the algorithm produced to reach the best result. It also contains the values below for each generation of protein sequences, known as InSiPS scores; except

for the fitness value, the other score values come from the PIPE predictions (PIPE score).

- **Fitness.** A value between 0 and 1. The best synthetic protein sequence will be determined in terms of the highest fitness value (described in Section 4.3.1).
- **Target score.** This value indicates the PIPE score (predicted interaction score) between the new sequence and the target protein. It represents the predicted likelihood that the synthetic protein interacts with the target protein.
- **Maximum non-target score.** This is the maximum PIPE score between the new sequence and the non-target proteins. It indicates the maximum predicted likelihood that the synthetic protein interacts with any of the non-target proteins.
- **Average non-target score.** This is the average of the PIPE score between the new sequence and the non-target proteins. It indicates the average of the predicted likelihood of interactions between the synthetic protein and the non-target proteins.

Results File

The information that this file contains for each generation of new sequences is as follows. First, the new sequences of the generation are displayed; the length and the number of new sequences depend on the predefined parameters. For each new sequence of the generation, the predicted PIPE (interaction) score with the target protein is displayed and the new protein sequence with the highest PIPE score is highlighted. Second, for the new sequence of the generation with the highest target score, the predicted interaction score (PIPE score) against the target and non-target proteins is presented. The highest predicted PIPE score is for the target protein and the rest of the predicted interactions are identified as *off-targets* (predicted interaction with the non-targets). The information in this part of the file was described in Section 4.2.3 for the PIPE output file; thus for every evaluated protein pair, protein A will be the new sequence, and protein B any protein from the target and non-target list.

4.3.4 Validation of the Results

As stated previously, the final result is the new protein sequence with the highest fitness value, calculated from the PIPE score (the predicted interaction score) against the target and non-target proteins. To validate the accuracy of the final results, any individual PIPE score (against target or non-target proteins) is taken into account. Thus, one can get an idea of the values for sensitivity (true positives) and specificity (true negatives) that would be produced at any operating point (score). The way the accuracy of the final result is measured is the same process as described in Section 4.2.5, Leave-One-Out Cross-Validation Tests. Similarly to the PIPE validation for a proteome-wide evaluation, it is expected to have an extremely low number of false positives; therefore, a specificity above 99.95% would indicate a high confidence interaction.

Chapter 5

Problem Statement

The focus of this research was to use a sequence-based computational tool of high-performance to design synthetic proteins. This thesis work formed part of current collaborative research on Duchenne muscular dystrophy (DMD). A possible treatment for DMD (described in Section 2.2.1) consists in injecting patients with quantities of healthy muscle satellite cells (essential for muscle growth and regeneration) grown in tissue culture.

Unfortunately, as mentioned in Section 2.2.2, producing a large number of satellite cells while preserving their regenerative properties is currently a major challenge for medical science. When investigators attempt to grow them in tissue culture, they tend to convert (differentiate) into muscle cells prematurely.

If satellite cells could be grown in tissue culture, tightly controlling for the premature differentiation, they could be considered for use as a therapy for DMD. The disruption of a protein-protein interaction (PPI) (detailed in Section 2.2.3) has been identified as a potential target to avoid the premature differentiation of muscle satellite cells.

The aim of this proposed research is to design a synthetic protein sequence (of characters representing amino acids) able to disrupt the PPI proposed to regulate the premature differentiation of muscle satellite cells. The synthetic protein would act as an inhibitory protein by competitively binding to either one of the two proteins involved in the PPI. Computational experiments with InSiPS, the **In**-**Si***lico* **P***rotein* **S***ynthesizer* (described in Section 4.3), will be carried out on the IBM Blue Gene/Q supercomputer. Specifically, human protein data will be used to generate protein sequences predicted to interact with either one of the target proteins without interacting with the rest of the proteins expressed in the human proteome.

Chapter 6

InSiPS Experiments with Human Proteins Data

6.1 Introduction

Chapter 5 detailed the general and specific objectives for this research study and Chapter 4 provided implementation details of InSiPS, the sequence-based protein synthesizer that was used to find a solution for the problem this research study sought to solve. This chapter summarizes the methods that led to the design of synthetic proteins that could potentially disrupt the SIX-EYA interaction. Certain synthetic protein sequences (also known as new sequences) generated in this phase were selected for the data analysis of results detailed in Chapter 7 that preceded the wet-lab experimental validation phase described in Chapter 8.

As noted, the aim of the experimental phase with human proteins described in this chapter was to use InSiPS to design several new sequences to tackle the problem described in Chapter 5. The specific objectives when performing the computational experiments with InSiPS are listed below.

- **Generation of several candidate (synthetic) protein sequences to target either SIX1 or EYA2 proteins.** As described in Section 2.2.3, the interaction between these proteins is of special interest for this study. Having a synthetic protein able to interact with either SIX1 or EYA2 could potentially disrupt their interaction.

- **The non-target list should be the whole human proteome.** Unlike this study, previous InSiPS studies with yeast proteins had considered a set of only

1,701 target and non-targets [7]. In the human proteome more than 20,000 proteins are expressed. In this study, that would be the set of non-targets. That is, the new sequence should be able to interact with the target protein and not interact with the rest of the proteins expressed in the human proteome.

- **The length of the new sequences should be between 25 and 35 amino acids.** A size shorter than 25 amino acids would be better for wet-lab experiments because it has been shown that such peptides are more stable and long lasting. However, the sliding-window approach used by InSiPS (described in Section 4.2.2) that looks for similar fragments is of size 20. With a length of 25 or 35 amino acids, at least six or 16 overlapping fragments of the new sequence could be searched for in the list of known PPIs.

- **Performing of preliminary testing with human proteins.** This research study represented the first time InSiPS was used with human proteins. At the time the experiments started the only data available had been collected approximately five years earlier. Although this data was not considered for the formal experiments with InSiPS, it was used to determine how feasible it would be to use InSiPS with human proteins.

- **The new sequence should have a high fitness value.** The new sequence should have a high predicted interaction score with the target and a very low maximum predicted interaction score with the list of non-targets — which implies the final fitness value should be as high as possible.

The above-described objectives were the foremost considerations when performing the experiments with InSiPS. However, since it was the first time InSiPS was applied to a research study involving human data, other aspects were also taken into account to have a better understanding of the results. Therefore, not only SIX1 and EYA2 proteins, but also other proteins from the SIX and EYA families were used as targets. Also, subsets of the list of non-targets were used before deciding upon the final set. Additionally, new sequences of more than 35 amino acids (i.e., 50, 75, 100, 150) were generated to see how the fitness value and runtime varied.

The highlights of the experimental phase using InSiPS to design synthetic protein sequences with a high likelihood to disrupt the SIX-EYA interaction will be reviewed in the following sections. Section 6.2 describes the human proteins data used in the

running of the experiments, including the old data and the newly collected data. Section 6.3 describes the different targets and the lists of non-target proteins as well as the configuration of parameters for the Genetic Algorithm (GA) used for the experiments. Section 6.4 describes the technical characteristics of the computer clusters where the experiments with InSiPS were performed. Section 6.5 summarizes the preliminary testing done before proceeding to the formal experiments. Section 6.6 details the formal execution of the experiments from which the output data is analyzed in later chapters. Section 6.7 details results about specific new sequences and how they were validated. Lastly, Section 6.8 includes a summary of the last round of experiments that targeted specific regions of the SIX1 protein sequence.

6.2 Human Proteins Data Set

This section describes the data used for the protein sequences file and the known protein-protein interactions (PPIs) file. As mentioned in Section 4.3, InSiPS was used to design yeast *S. cerevisiae* proteins and their efficacy was demonstrated through wet-lab experiments. However, InSiPS had not been used with human proteins before. Hence preliminary trials were necessary to demonstrate that it would work with human data given the larger size of the data set and the greater complexity of human protein sequences in comparison to yeast data. At the time the preliminary executions with InSiPS started the only available human data was a data set collected approximately five years earlier. A project to have more recent and reliable data from human was later started. The characteristics of the old and the recently collected data are described below.

6.2.1 Old Human Data Used for Preliminary Testing

The data set summarized in Table 6.1, and termed *old human data*, was collected approximately five years ago from different sources such as the BioGrid interaction database [35]. At the time this research project started it was the only data set available; therefore, it was considered in the preliminary computational experiments. The known interactions of this data set could not be considered of high confidence for two reasons: first, the data was considered outdated and new reported protein interactions could help to achieve a better outcome; also, due to the implications of the project, a stricter process of extraction for experimentally validated protein

interactions was needed.

Description	Value
Total proteins	20,271
Known protein pairs	78,318
Proteins with at least one binding partner	12,751
Proteins with no known interactions	7,520
Maximum number of known interactions/protein	8,278
Average of known interactions/protein	7.66
Average of known interactions/protein with at least one binding partner	12.17
Longest protein length	8,797

Table 6.1: Summary of statistics for the old human data. This proteins data was collected approximately five years ago and was used for preliminary computational experiments.

6.2.2 New Human Data Used for Experiments

The collection of the most recent human proteins data used for computational experiments was performed by Dick *et al.* [89] at Carleton University. For the process of data collection, first the list of reported protein-protein interactions (PPIs) for all organisms was downloaded from the BioGrid repository [35]. Next, only PPIs with an identifier for human were extracted. Then, from this subset, two different data sets termed *permissive* and *conservative* were filtered. For the *permissive* data set, physical and genetic reported interactions were considered; the final data set consisted of 154,514 known PPIs. For the *conservative* data set only physical interactions were considered; also, only interactions with multiple lines of evidence were retained; therefore, every reported interaction was confirmed by two or more independent research groups; the final data set consisted of 13,938 known PPIs. Lastly, all the protein sequences from the obtained known PPIs were extracted from UniProt [31] with the format required by InSiPS.

As the *conservative* data set provided a higher likelihood of certainty that each

reported PPI was true, this was the data used for the experiments with InSiPS. The termed *new human data* is summarized in Table 6.2.

Description	Value
Total proteins	23,493
Known protein pairs	13,969
Proteins with at least one binding partner	5,205
Proteins with no known interactions	18,288
Maximum number of known interactions/protein	197
Average of known interactions/protein	1.16
Average of known interactions/protein with at least one binding partner	5.25
Longest protein length	8,797

Table 6.2: Summary of statistics for the new human data. This is the most recently collected [89] human proteins data used for computational experiments.

Specifications of the New Human Data

The following are two important changes applied to the final new human data set of 13,969 known protein pairs summarized in Table 6.2 and used for the experiments with InSiPS.

First, the *conservative* data set of known protein interactions had only two unique known interactions for EYA and SIX proteins. In order to increase the chance that a synthetic protein sequence generated by InSiPS would interact with a target from either the EYA or SIX family of proteins, 36 more EYA- and SIX-related known protein pairs were added from the *permissive* data set. In total 38 unique known interactions involved either EYA or SIX proteins in the final data set. The total of known binding partners for each SIX and EYA protein is summarized in Table 6.3. Note that 38 is the total number of unique interactions; however, the total number of interactions in this table is 43, as ten interactions (five repeated pairs) occur between proteins from the EYA and SIX families.

Second, the original data set consisted of 23,495 proteins (instead of the listed

23,493) as it included two long proteins sequences: Mucin-16 (MUC16, Q8WXI7) with 22,152 amino acids and no known interactions; and Titin (TTN, Q8WZ42) with 34,350 amino acids and five known interactions. These two proteins and their known interactions were removed from the final set of protein sequences because they significantly impaired the computation; this left 23,493 sequences. Thus, the longest protein in the final data set consisted of 8,797 amino acids.

To sum up, the final number of known protein interactions in the final new human data set is calculated as follows: 13,938 initial known PPIs plus 36 unique protein pairs (SIX- and EYA-related), minus five known pairs related to the longest protein (TTN, Q8WZ42) which equals 13,969.

Protein (gene name)	Interactions
EYA1	5
H2AFX, SIX1, FZR1, GSK3B, FBXW7	
EYA2	7
DMRTB1, RBPMS, SIX4, GNAI2, GNAZ, SIX1, MMS19	
EYA3	5
SKI, H2AFX, TCEAL1, ZDHHC17, SIX5	
EYA4	1
SIX1	
SIX1	12
CCDC85B, EYA2, FZR1, VTN, EYA4, MYH3, MYH7, MYH4, EYA1, MDFI, SKI, AES	
SIX2	1
APP	
SIX3	5
TLE1, AES, NR4A3, MTA1, PAX6	
SIX4	2
LRRC46, EYA2	
SIX5	2
ATXN1, EYA3	
SIX6	3
TLE1, AES, GTF2A1L	

Table 6.3: Known interactions for EYA and SIX proteins in the new human data.

6.3 Input Data for Experiments

The data used for the formal execution of experiments with InSiPS was the new human data described in the previous Section, 6.2. This is the data used for this section and for all subsequent sections unless otherwise indicated.

The input data required by InSiPS was detailed in Section 4.3.2. In brief, two sets of two plain text files are needed for the actual InSiPS execution. The first set is the list of protein sequences and the list of known protein-protein interactions. The second set of files is the list of target and non-target proteins. The contents of the first set was summarized in Table 6.2; hence, the baseline data for the computational experiments described in this chapter is the list of protein sequences with a total of 23,493 proteins and the list of known protein-protein interactions with a total of 13,969 interacting pairs; this includes the known binding partners for the SIX and EYA proteins summarized in Table 6.3. The contents of the second set of files is described below in Section 6.3.2 and 6.3.3. Multiple runs of InSiPS were performed trying to find a new sequence that would interact with the target with a high fitness value. Also, different combinations of target and non-target proteins were considered as input.

6.3.1 Clustering the Highly Homologous Protein Sequences

In order to reduce the size of the sets of proteins, the CD-HIT (Cluster Database at High Identity with Tolerance) tool [90] was used to cluster highly homologous protein sequences. This would lead to a smaller set of proteins where the homologous sequences would be represented by one sequence. Depending on the pre-configured value for sequence identity, CD-HIT determines the uniqueness of protein sequences before grouping the sequences at a given percentage. As documented in the web portal [91], when clustering a set of proteins, two parameters are important: the percentage of sequence identity at which the sequences will be clustered; and the word size, to indicate the size of overlapping *k-mers* (substrings of length k) that will be compared for similarity. When clustering at a 50% of sequence identity, the word size is three; at a 60% of sequence identity, the word size is four; at a 90% of sequence identity, the word size is five.

Proteins Used as Targets

Proteins from the same family are so characterized because they have a common ancestor. Among the proteins in the SIX and EYA protein families are certain ones with some degree of homology (defined in Section 2.1.2). When SIX proteins were clustered with CD-HIT at a sequence identity of 60%, SIX3 and SIX6 were found to be similar with an identity of 81.3%; and SIX1 and SIX2 had similarity with an identity of 77.82%. Likewise, when EYA proteins were clustered, EYA1 and EYA4 were found to be similar at a sequence identity of 71.45%.

Proteins Used as Non-targets

When the experimental phase with InSiPS started, it was thought that using human proteins and the whole set of proteins expressed in the proteome as the list of non-targets would present a computational challenge. For previous tests with yeast a set of 1,701 non-target proteins was used while for this research a set of more than 20,000 non-target proteins would be required. Taking this into account, for the initial testing, smaller sets of non-target proteins as described below in 6.3.3 were used.

6.3.2 Target Proteins

The major objective for InSiPS is to identify a novel sequence that is predicted to interact with a determined target protein. The target protein is identified in a text file by the id of the protein; there is only one target per InSiPS execution.

As was described in Section 6.1, one of the specific objectives of the experiments with InSiPS was to design a new sequence to interact with either the protein SIX1 or EYA2. Therefore, these two proteins were considered as targets in different runs; other proteins from the SIX and EYA families were also used as targets. The idea behind these different targets (one per execution) was to identify if other targets besides SIX1 or EYA2 could produce a new sequence with a high fitness value. This would mean a target for a potential new sequence able to disrupt the SIX-EYA interaction.

The distinct targets used for the different InSiPS executions can be summarized as follows:

- Proteins from the SIX (SIX1 to SIX6) and EYA (EYA1 to EYA4) families.

- Proteins made up of specific sites of SIX1 (antiSix1_9-43, antiSix1_1-37, antiSix1_1-128, antiSix1_9-118). Further detail regarding these proteins is given in Section 6.8.1.

6.3.3 Non-target Proteins

Another specific objective for the experiments with InSiPS described in Section 6.1 was to use the whole set of proteins (excepting the target) expressed in the human proteome as non-targets. Consequently, the baseline set for the non-targets was the 23,493 proteins described in 6.2.2. Since this research study represented the first time that InSiPS was used with human proteins, the initial experiments included tasks to reduce as much as possible the quantity of input data. To reduce the set of non-target proteins, the CD-HIT tool was used. The three data sets described below were used in different computational experiments as the list of non-target proteins. Note that the target is never part of the list of non-target proteins.

- *Dataset1*. New human data with a total of 23,493 protein sequences.
- *Dataset2*. New human data clustered at a 90% of sequence identity. The result of this set was 21,055 protein sequences.
- *Dataset3*. New human data clustered at a 60% of sequence identity. The result of this set was 17,891 protein sequences.

For experiments with *dataset2*, *dataset3* and the majority of experiments with *dataset1*, the target and the family proteins of the target were excluded. That is, when any SIX or EYA protein was used as a target, all the proteins from the same family, SIX or EYA, were removed from the list of non-targets. As previously described, SIX and EYA family proteins have a certain degree of similarity. Considering the sliding-window approach, which is the fundamental idea behind PIPE (described in Section 4.2.2), it would be impossible to design a synthetic protein sequence 100% specific if some of the non-target proteins are highly homologous to the target. Therefore the intention of removing the whole family of proteins (either EYA or SIX) was to make it less difficult for the algorithm to design a new sequence with a high affinity for the target.

For other experiments with *dataset1* the family proteins of the target were included in the set of non-targets. That is, only the target protein was excluded from the non-targets. For example, if SIX1 was the target, the non-target list contained 23,492 proteins including the SIX family proteins from SIX2 to SIX6. These tests were executed with the intention of seeing how the fitness value changed and how feasible it was for the algorithm to design a protein sequence able to interact with the target without affecting the defined non-targets. Therefore, the two different methods for creating the list of non-targets for *dataset1* can be identified as follows:

- *Dataset1a.* Excluding the target and all the family proteins of the target from the non-target list.
- *Dataset1b.* Excluding only the target protein from the non-target list.

6.3.4 Genetic Algorithm Parameters

The genetic algorithm (GA) parameters were defined in Section 4.3.1. The parameter values that were used as input for the different experiments are listed in Table 6.4. Values such as the probability for the standard GA operations were defined based on the InSiPS parameter tuning described in Schoenrock's Doctoral thesis [88].

Different Input Values for Sequence Length

The values for the sequence length varied from 21 to 150 amino acids. The idea behind this approach was to see how the fitness value and the computation time varied for different sequence lengths. However, due to the sliding-window size of 20 amino acids (described in Section 4.2.2) and the specific objective of producing new sequences of lengths 25 and 35 explained in Section 6.1, these were the most used values for sequence length. More InSiPS runs were performed for a sequence length of 35 to assure that more fragments (16 fragments) of the sequence would be evaluated for comparison when looking for similar fragments in the known PPIs list. Also, InSiPS runs for a sequence length of 25 (six fragments to compare) were performed looking for new sequences of a short length. A sequence length lower than 25 amino acids would imply fewer fragments to compare, consequently lowering the chances to find similar fragments in the list of PPIs. Hence the sequence length of 25 was the preferred lower limit.

Parameter	Values
Sequence length	21 - 150
Population size	1000
Probability GA operations	
copy	0.10
mutate	0.40
mutate-aa	0.05
cross-over	0.50
Number of generations	
Minimum	250
Maximum	1,000,000
Unchanged count	50
Fitness cutoff	MAX
Random seed	False

Table 6.4: Values used for the genetic algorithm (GA) parameters. The parameter *mutate-aa* is the probability to change each amino acid when performing the operation of *mutate*.

6.4 Computer Clusters Available for Experiments

This section describes the two computer clusters used for the InSiPS experiments. The Carleton Bioinformatics data lab cluster (also known as Data Lab cluster) was used for preliminary testing and for preprocessing of the data used for the formal execution of experiments. The IBM Blue Gene/Q (BGQ) cluster was used for some preliminary testing and for the formal execution of experiments.

6.4.1 Carleton Bioinformatics Data Lab Cluster

The Data Lab cluster consists of 19 nodes available for bioinformatics investigations at Carleton University. Each node has 32 GB of RAM and one quad-core Intel Core i7 3.40 GHz processor, supporting 2 threads per core (i.e., 8 hardware-supported threads every 4 cores). Although the InSiPS implementation requires a much higher compute capacity, this cluster was used for preliminary testing to ensure that the

algorithm was able to work with human data. Only 14 of the 19 nodes were used due to tests from other research projects being performed using the cluster at the time these experiments were carried out.

6.4.2 IBM Blue Gene/Q Cluster

The BGQ cluster is a Southern Ontario Smart Computing Innovation Platform (SOSCIP) BlueGene/Q supercomputer. It is a 3rd generation Blue Gene IBM supercomputer with 4096 nodes and 16 GB of RAM per node. Each node has 16 cores of 1.6 GHz, supporting 4 threads per core. The BGQ system consists of the front-end and the compute nodes. One cannot log in directly to the compute nodes; instead all the work is sent from the front-end through the development node *bgqdev-fen1*, which is a Power7 machine running Linux. Programs are compiled on this development node and then submitted as jobs to the queue using a batch scheduler, called the *LoadLeveler*. The only way to run a program on the BGQ is to submit a job through the development node. On the development node it is possible to see and manage files, compile a program with the special BGQ compilers, write and submit the job scripts, and check the job status. The maximum execution time for each job submitted is 24 hours and there is no limit on the number of jobs. However, the start time for a submitted job depends on, among other variables, the number of nodes requested, the maximum execution time limit (*wall clock limit*) and the current availability of resources (i.e., how many jobs are in the queue). Therefore, a job submitted requesting that it be executed in 2,048 nodes with a *wall clock limit* of 24 hours would take significantly longer to start than a job submitted for 128 nodes with a *wall clock limit* of 12 hours.

6.5 Preliminary Testing

Because this was the first time InSiPS was executed with human proteins, some preliminary tests were performed, first on the Data Lab cluster and then on the BGQ cluster. The input data for this testing was the old human data (described in Section 6.2.1) as this was the only data available when the experiments were started. The results of this preliminary testing allowed seeing how the final computation time and fitness value changed, depending on the combination of certain input values such as the number of non-targets, the sequence length, the population size and the number

of computational threads and nodes used.

6.5.1 Preliminary Testing on the Data Lab Cluster

Using a random sample of 1000 non-target proteins and 100 as the maximum limit of generations, different values for sequence length and population size were tested on 14 nodes of the Data Lab cluster. Table 6.5 displays the results. These parameter settings allowed assessing how execution time and fitness value varied as these variables changed. As can be observed, set 1 with the longest sequence length (250) was the one with the lowest fitness value; besides sequence length, the small population size influenced this result. In the second set, the fitness value improved after shortening the expected sequence length. For the third and fourth sets the population size was changed to 500 as this would be a value closer to what would be needed for a real test; the fitness value was more consistent; however, the execution time started to become an issue as tests with very low fitness values lasted up to 26 hours. The observed memory consumption on the Data Lab cluster during this testing was around 6 GB using 8 threads and around 9 GB for 16 threads. After the trial InSiPS executions and taking into consideration that the BGQ cluster nodes would have 16 GB of memory, it was determined that testing with human data on the BGQ could be possible.

Set	Sequence length	Population Size	Fitness	Runtime (hrs.)
1	**250**	**25**	0.057664	1.0
2	**50**	**25**	0.191410	0.8
3	**100**	**500**	0.198044	25.7
4	**50**	**500**	0.199077	14.8

Table 6.5: Results of InSiPS preliminary tests on the Data Lab cluster. A random sample of 1000 non-target proteins with 100 as the maximum limit of generations with different values for sequence length and population size was tested to observe effect on fitness value and runtime.

6.5.2 Preliminary Testing on the BGQ Cluster

Once it was validated that InSiPS was able to run with human proteins on the Data Lab cluster, further preliminary testing was performed on the BGQ. This was to see

how execution time and fitness value improved with a high performance computer cluster. Using a random sample of 1000 non-target proteins, with 50 as the new sequence length, and 500 as the size of the population, five different sets were tested on the BGQ cluster. As can be seen in the results shown in Table 6.6, set 4 took more processing time than set 5. Although the reduction in runtime was expected because fewer nodes were used for set 5, the one hour difference did not reflect the increase in the number of nodes. It might also be noted that set 4 obtained the best fitness value. In contrast to the execution on the Data Lab cluster, on the BGQ cluster the fitness value was higher and the computation time was considerably lower with a population size of 500 new sequences.

Set	Generations	Nodes	Threads	Fitness	Runtime (hrs.)
1	**100**	**64**	**16**	0.199446	6.9
2	**100**	**128**	**32**	0.196848	2.4
3	**250**	**128**	**32**	0.205227	9.5
4	**250**	**256**	**32**	0.206458	5.9
5	**250**	**512**	**32**	0.202775	4.9

Table 6.6: Results of InSiPS preliminary run on the BGQ cluster. A random sample of 1000 non-target proteins with 50 as the sequence length and 500 as the population size was tested on the BGQ cluster to observe how runtime time and fitness value improved with a high performance computer cluster.

In this preliminary testing it was found that the maximum number of computational threads supported by the BGQ cluster to run InSiPS with human proteins was 32. With more than 32 threads an error occurred due to there not being enough memory to allocate PIPE matrices during execution.

6.6 Formal Execution of Experiments

This section describes the preprocessing of input data on the Data Lab cluster and the overall results from the subsequent formal execution of experiments on the BGQ cluster with different input values. Other important issues for this phase are described, such as how the output data was extracted for analysis, how the new protein sequences were named and how they were validated to be unique.

6.6.1 Data Preprocessing on the Data Lab Cluster

The preprocessing of the input data (described in Section 4.3.2) used for the actual InSiPS experiments was done on the Data Lab cluster through the two programs described below.

- **PIPE setup.** The first step was to preprocess with the PIPE Setup program the input data (described in Section 4.2.1) based on the input files with the protein sequences and the known protein pairs. This program is used as a first step before starting to work with either the PIPE prediction tool or with InSiPS. Its function is to ensure that the input files have the proper format and to validate that the proteins of the sequences file contain valid amino acids. This program generates the PIPE database files and the parameters file from which some data statistics were outlined earlier in Table 6.2.

- **Combine database**. The second step after executing the PIPE Setup program was to process the resultant data to generate a unified database (described in Section 4.3.2) with the Combine DB program.

6.6.2 Protein Naming

The name of each new protein sequence was defined as *anti[target_name]-[id]*. The prefix *anti* is because the aim of the new sequence to disrupt a natural behavior of a protein; the *target name* indicates the target protein the new sequence was designed to interact with; and the *id* is a unique identifier corresponding to the InSiPS run on the BGQ. Therefore, by looking at the protein name, it is possible to find all the information related to the run (log files, input and output files) in the computer cluster and also in the general report and in the reports specific to the predicted interactions for the new sequence (i.e., top 100 interactions) as described in Section 6.6.4.

6.6.3 Testing on the BGQ Cluster and Overall Results

In this section the highlights of the overall results obtained in the formal testing phase on the BGQ cluster are summarized. This phase consisted of computational experiments (InSiPS runs) with the input data described in Section 6.3. With this

input data several combinations of the described *targets*, *non-targets* and the *GA* parameters with different sequence length were used.

When submitting jobs to the BGQ, the *wall clock limit* was set to the maximum (24 hours) and the number of nodes used per run varied from 64, 128 and 256, with 179 nodes on average. Using more than 256 nodes represented a potential delay in the start of each submitted job, as the BGQ automatically prioritizes those runs requiring a shorter execution time and fewer resources. With InSiPS runs on 64 nodes the runtime was 21 hours, while on 256 the runtime was reduced to 6.7 hours. All of the experiments ran with 32 computational threads (for each node) as this was found to be the maximum supported threads for human data on the BGQ. A lower number of computational threads represented an increase in the execution time.

In this phase 180 synthetic protein sequences (one per run) were generated on the BGQ, with a total computation time of 2,047 hours for all the runs and a median time of 10.8 hours per run. The average of the hit generation was 144, with 306 and 1 as maximum and minimum values, and a standard deviation of 87. All the new sequences were considered candidate protein sequences, and the highest ranked new sequences were carefully analyzed (as described in Chapter 7) before proceeding to the experimental validation phase described in Chapter 8. The following describes the highlights of this round of experiments in terms of how the fitness and runtime changed depending on the sequence length, and the sets of non-targets and targets used.

It should be noted that the InSiPS algorithm is stochastic in nature as the first generation, also known as seed, for the genetic algorithm is made up of sequences with random amino acids and each subsequent generation tries to improve the fitness value from the previous generation. Consequently, great differences in terms of the obtained InSiPS scores, runtime, and hit generations might be found for tests with equal or similar input values.

Fitness and Runtime for Different Sequence Lengths

Table 6.7 illustrates how the fitness value and runtime changed depending on the length of the new sequence generated. All of the new sequences were generated on 256 nodes using the above described *dataset2* for the list of non-targets. The number of new sequences generated for each sequence length is indicated in parentheses. For the antiSix1 sequences, the longer the sequence, the lower the fitness value and also

the longer the runtime. For the antiEya2 sequences, an increase in the runtime can be observed for longer sequences; and the highest fitness value was obtained for sequences of 75 amino acids.

Sequence length	25	50	75	100	150
antiSix1 (runs)	(3)	(3)	(2)	(2)	(1)
Avg. Fitness	0.878892	0.871314	0.868917	0.682358	0.422390
Avg. Runtime (hours)	6.91	10.87	15.9	18.0	18.0
antiEya2 (runs)	(3)	(3)	(2)	(2)	(2)
Avg. Fitness	0.446906	0.549583	0.573912	0.541054	0.461188
Avg. Runtime (hours)	7.1	12.81	13.34	15.0	18.0

Table 6.7: InSiPS testing results for new sequences of different length. All the new sequences were generated on 256 nodes using *dataset2* for the non-targets. The number of new sequences (runs) generated for each sequence length is shown in parentheses.

Fitness Value for Different Sets of Non-targets

The average of the fitness values obtained for 39 new antiEya2 and antiSix1 sequences generated with the different sets of non-targets described in Section 6.3.3 is illustrated in Figure 6.1. The 39 sequences are distributed as follows: for *dataset1a*, three antiEya2 and 15 antiSix1 sequences; for *dataset1b*, three antiEya2 and six antiSix1 sequences; for *dataset2*, three antiEya2 and three antiSix1 sequences; for *dataset3*, three antiEya2 and three antiSix1 sequences. The highest fitness value for both antiSix1 and antiEya2 sequences was obtained with *dataset3*, which is the smallest set of non-targets. A slight increase in the fitness value can be observed for antiSix1 sequences generated with smaller sets of non-targets. The fitness value was lower for both antiSix1 and antiEya2 using *dataset1b* because this set included the target family proteins (excepting the target) in the list of non-targets. Hence, it was more difficult to generate a new sequence that would interact with the target (e.g. SIX1) without interacting with the non-targets (e.g. SIX2 to SIX6 and all the others). In this figure it can also be observed that antiSix1 sequences obtained a considerably

higher fitness in all cases.

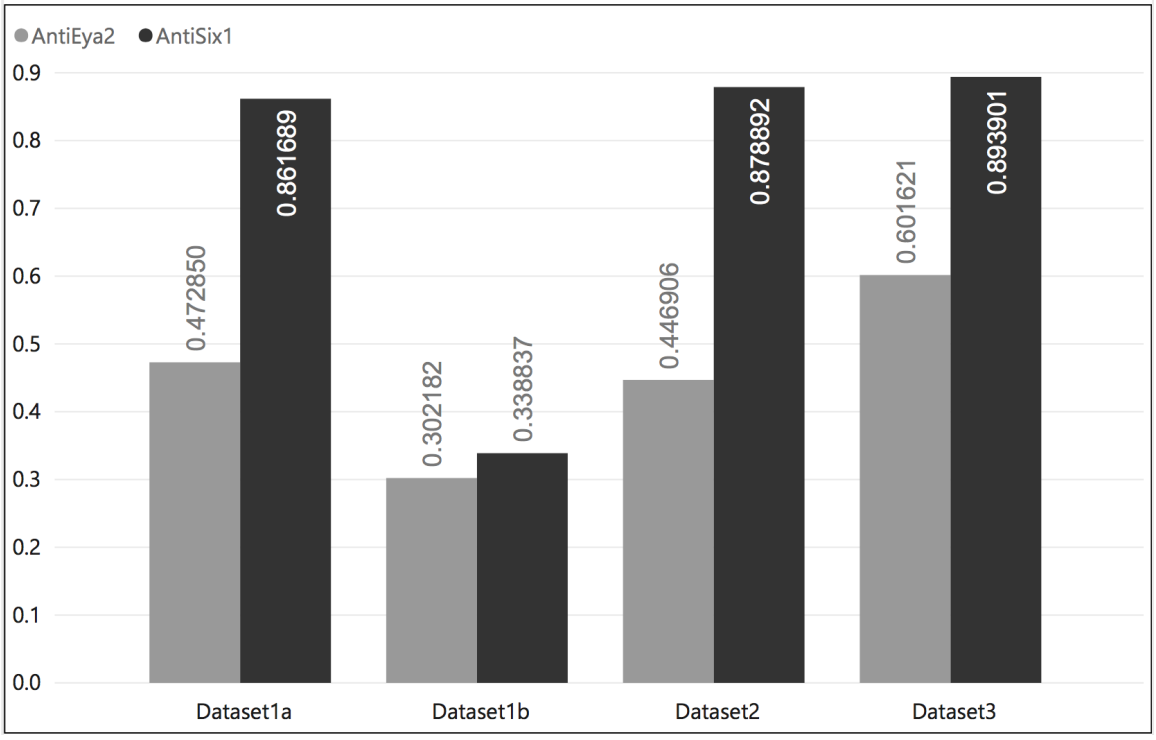


Figure 6.1: Average of fitness value for 39 new antiEya2 and antiSix1 sequences generated with different sets of non-targets. All of the new sequences are of length 25. *Dataset1* is the complete set of proteins (23,493); *dataset1a* represents a non-target set with all the human proteins except the target family proteins; *dataset1b* represents a non-target set with all the human proteins except the target protein. *Dataset2* represents a non-target set clustered at a 90% of sequence identity (i.e., 21,055 proteins). *Dataset3* represents a non-target set clustered at a 60% of sequence identity (i.e., 17,891 proteins).

Fitness of New Sequences for SIX and EYA Proteins as Targets

Another round of experiments consisted of testing how the fitness value changed for the different SIX and EYA proteins as targets. The average of fitness for EYA1 to EYA4 and SIX1 to SIX6 proteins as targets is summarized in Figure 6.2. The average fitness was calculated from at least three InSiPS runs for each of these proteins (EYA1-EYA4, SIX1-SIX6) as targets; the non-target set used was *dataset1b*, which excludes only the target protein; and the length of the new sequences was 35. As previously mentioned, for runs using this specific set of non-targets the fitness value was lower; however, it is interesting to observe that for the EYA2 and SIX1 proteins as targets

the fitness value was the second highest for each family. This indicates that EYA2 and SIX1 are easier to target than the majority of the other proteins in each family.

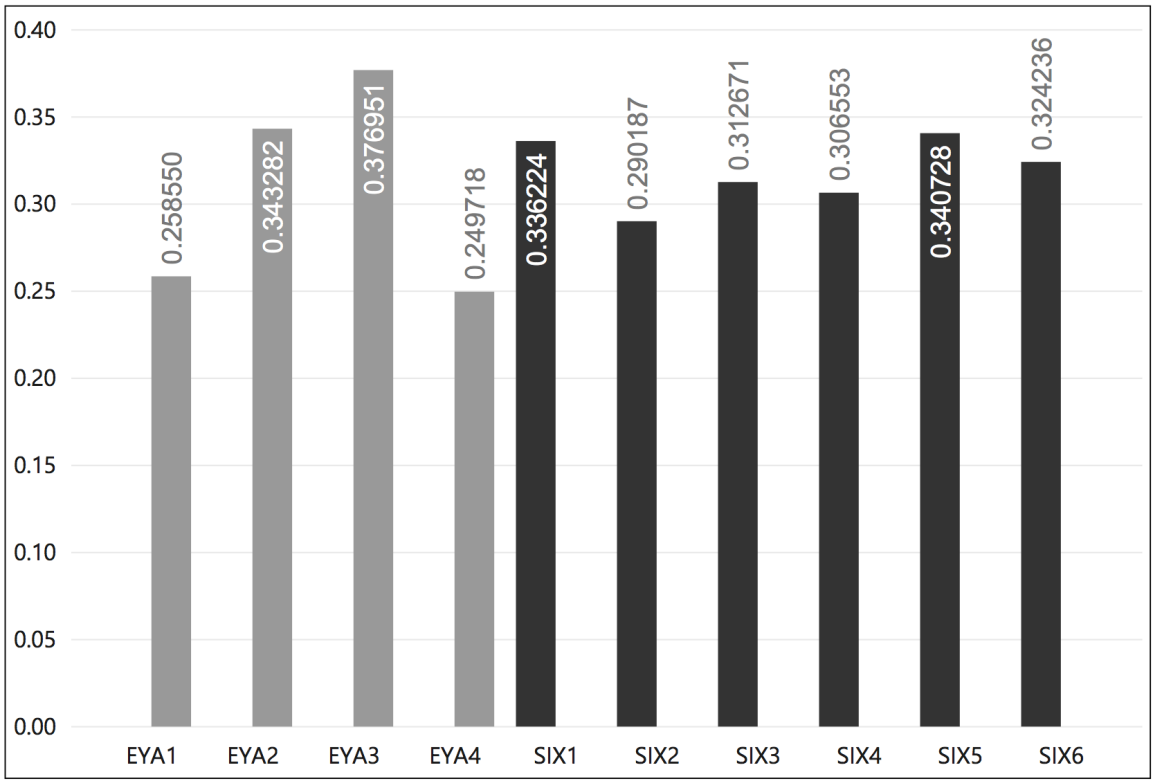


Figure 6.2: Average of fitness value for new anti-EYA and anti-SIX sequences generated with the same set of non-targets. All of the new sequences are of length 35 and were generated with *dataset1b*, which is a non-target set with all the human proteins, except the target protein.

6.6.4 Extraction of Data for Analysis

The output files (i.e., console file and results file) described in Section 4.3.3 contain all the information produced with every InSiPS execution. The most important result at the end of each execution was the amino acid sequence of the synthetic protein with the highest fitness value. However, more data from the above-mentioned files was considered to determine whether the new sequence would be a good candidate for experimental validation. In the following is described what information from the last generation (with the highest fitness value) of each run was used and how the data was extracted for further analysis.

General Report

The most important value from the output console file was the amino acid sequence with the highest fitness value. Besides the fitness value, other information from the run was recorded in a general report: the target score, the maximum non-target score and the average non-target score; also recorded were the total number of generations and the number of the generation in which the highest fitness value was achieved (*hit generation*); and finally, the runtime for each execution.

Reports of the Top 100 Predicted Interactions

From the output file all the predicted interactions between the new sequence and the target/non-target proteins were extracted. The data consisted of the protein identifiers and the PIPE scores (predicted interaction) between each pair of proteins (i.e., new sequence against target/non-targets) with the top three predicted binding sites. There were two main processes for data analysis and visualization. Firstly, a report was prepared (one per new sequence) for the data analysis as is presented in Chapter 7. In this report the amino acid sequences of target and off-targets (non-targets with some degree of predicted interaction) were added, as well as the subsequences from the top predicted interaction site for each protein pair; for each report only the top 100 predicted interactions were considered. Secondly, a report was prepared of the sensitivity and specificity values (Section 4.2.5 described the way the sensitivity and specificity were calculated) corresponding to each predicted PIPE score for the new sequence against target/non-targets. Then, ROC (receiver operating characteristics) curves were outlined showing the predictive discrimination performance for each of the top 100 predicted PPI. ROC curves are illustrated in Section 6.7.4.

6.6.5 Uniqueness of New Protein Sequences

Similar to the clustering with the CD-HIT tool to reduce the number of non-targets described in Section 6.3.1, before identifying the highest ranked protein sequences, all the new sequences were analyzed with CD-HIT. The aim of clustering was to validate the uniqueness of the new amino acid sequences, ensuring that the analysis of the predicted interaction against the target and off-targets was worthwhile. The set of new sequences was processed with CD-HIT looking to cluster similar sequences at a sequence identity from 50% up to 90%. In all cases the tool indicated that the

sequences were unique and hence no significant overlapping *k-mers* at different sizes (three to five) were found.

6.7 Results of Experiments Highlighting New Sequences

The results presented in this section were obtained in the formal execution of the experiments on the BGQ cluster described in Section 6.6. As mentioned in Section 6.1, new sequences designed to target the SIX1 and EYA2 proteins of lengths 25 and 35 amino acids were of special interest for further analysis and experimental validation. The results in the following sections highlight the new sequences fulfilling these conditions. Firstly, overall InSiPS scores (fitness, target and maximum non-target) for antiSix1 and antiEya2 sequences are presented; secondly, individual InSiPS scores for specific anti-SIX and anti-EYA new sequences are detailed; and finally the validation of results by means of the sensitivity and specificity statistical measures is described.

Analysis of the homology and interaction sites of the new protein sequences presented in this section is given in Chapter 7.

6.7.1 InSiPS Scores by Target (EYA2, SIX1) and Sequence Length (25, 35)

The InSiPS scores for 49 new protein sequences of length 25 and 35 designed to target either EYA2 or SIX1 proteins are illustrated in Figure 6.3. The list of non-targets is the previously mentioned *dataset1a* (Section 6.3.3), which excludes all the target (EYA or SIX) family proteins. From the 49 new sequences, 19 target EYA2 (antiEya2), among which are 16 antiEya2 of length 35, and 3 antiEya2 of length 25. The other 30 of the 49 new sequences target SIX1 (antiSix1), among which are 15 antiSix1 of length 35, and 15 antiSix1 of length 25. As can be observed in the figure, the fitness value for the antiSix1 sequences of length 25 and 35 is higher than for the antiEya2 sequences; also, the average target score is higher (around 0.85) and the average maximum non-target score lower (around 0.14) for antiSix1 sequences than for antiEya2 sequences. It can be observed as well that there is a wider range of fitness values for antiEya2 sequences than for antiSix1 sequences: the standard

deviation (SD) of the fitness for antiEya2 sequences is 0.0459 and 0.0887 for lengths 35 and 25, respectively; for the antiSix1 sequences the SD of the fitness is 0.0019 and 0.0008 for lengths 35 and 25. The low SD of the fitness for the antiSix1 sequences is because there was little variation in the target and maximum non-target scores, respectively. For some new sequences these scores were very similar; however, as was previously explained (Section 6.6.5), the new sequences were found to be unique. To sum up, the InSiPS algorithm was able to find new sequences with a higher fitness for SIX1 as target; that is, a higher target score and a lower maximum non-target score, in comparison to the new sequences designed to target EYA2.

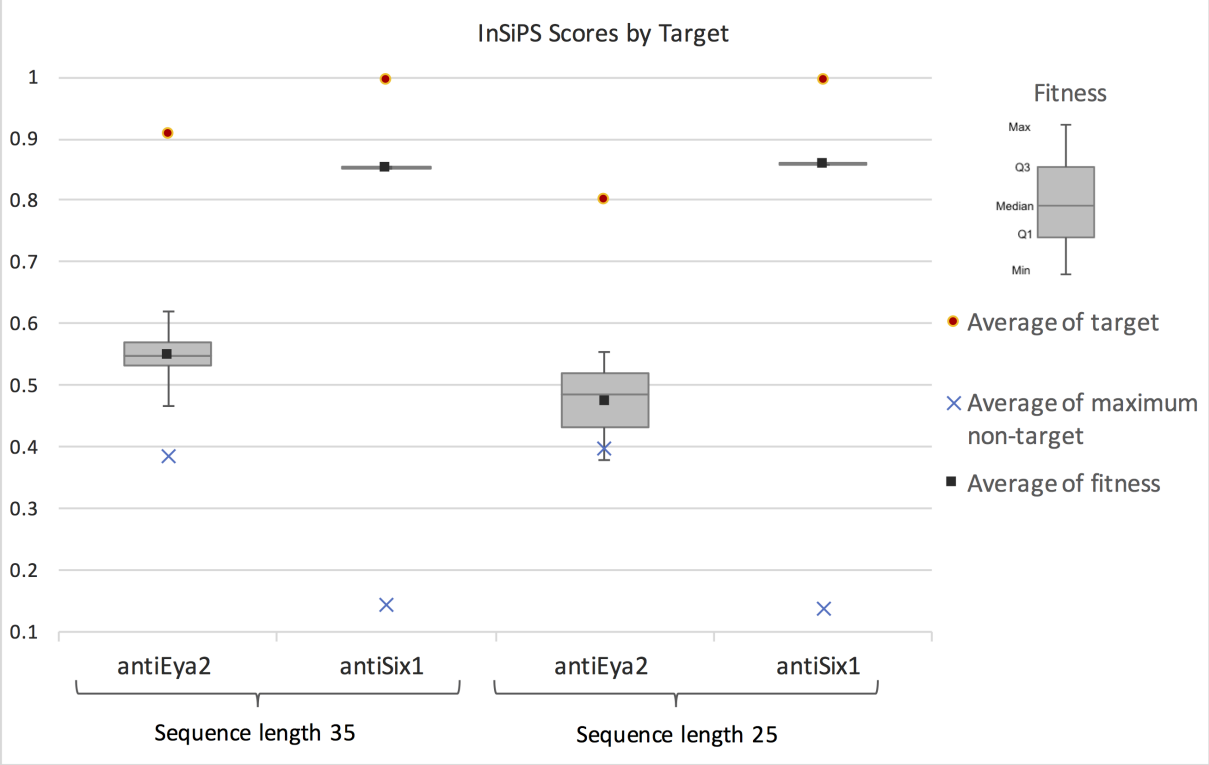


Figure 6.3: InSiPS scores for new antiEya2 and antiSix1 sequences of length 35 and 25. In total 49 new sequences of length 35 and 25 generated with *dataset1a* are evaluated. The boxplots indicate the distribution of the fitness values. Additional markers indicate the average of: fitness (black square), target (red dot) and maximum non-target (blue cross) scores.

6.7.2 New Anti-SIX Sequences and their Interactions

The new sequences designed to target the different SIX proteins for which output data is further analyzed later in this thesis (Chapter 7) are displayed in Table 6.8. As can be observed, the fitness value decreases considerably for those new sequences derived using *dataset1b* (described in Section 6.3.3) for the list of non-targets. The fact that the SIX family proteins (except the target) were included in the non-targets affected how InSiPS chose the candidate sequence with the highest fitness value from each generation (in Section 6.3.1 the homology of SIX proteins is described).

New Sequence	Length	Non-targets	Fitness	Target	Max non-target
antiSix1-76186	25	dataset1a	0.859335	0.995597	0.136865
antiSix1-76289	25	dataset1a	0.857706	0.993711	0.136865
antiSix1-76307	35	dataset1a	0.851081	0.996226	0.145695
antiSix1-76405	35	dataset1a	0.854813	0.994811	0.140728
antiSix1-76407	25	dataset1a	0.859335	0.995597	0.136865
antiSix1-76409	25	dataset1a	0.857706	0.993711	0.136865
antiSix1-76629	35	dataset1a	0.854813	0.994811	0.140728
antiSix1-77087	35	dataset1b	0.336455	0.643160	0.476872
antiSix4-77411	35	dataset1b	0.350961	0.696932	0.496421
antiSix4-77446	35	dataset1b	0.295193	0.561516	0.474292
antiSix6-77448	35	dataset1b	0.388486	0.765419	0.492453
antiSix4-88115	35	dataset1a	0.846570	0.998688	0.152318
antiSix4-88202	35	dataset1a	0.842024	0.999180	0.157285

Table 6.8: List of InSiPS scores for new anti-SIX sequences of length 25 and 35. New sequences designed to target the different SIX proteins for which output data is further analyzed in Chapter 7.

The top ten PIPE interaction scores for antiSix1-76405 against the rest of the proteins (target and non-targets) are shown in Figure 6.4. The InSiPS scores (fitness,

target, maximum non-target) were listed in Table 6.8; the average of the predicted interactions against all the non-targets for this new sequence was 0.000678. The highest predicted interaction score is against the target protein (SIX1): 0.994811. How the new antiSix1 protein was predicted to interact with the SIX family proteins is also plotted; however, it is noted that the SIX proteins were excluded from the list of non-targets.

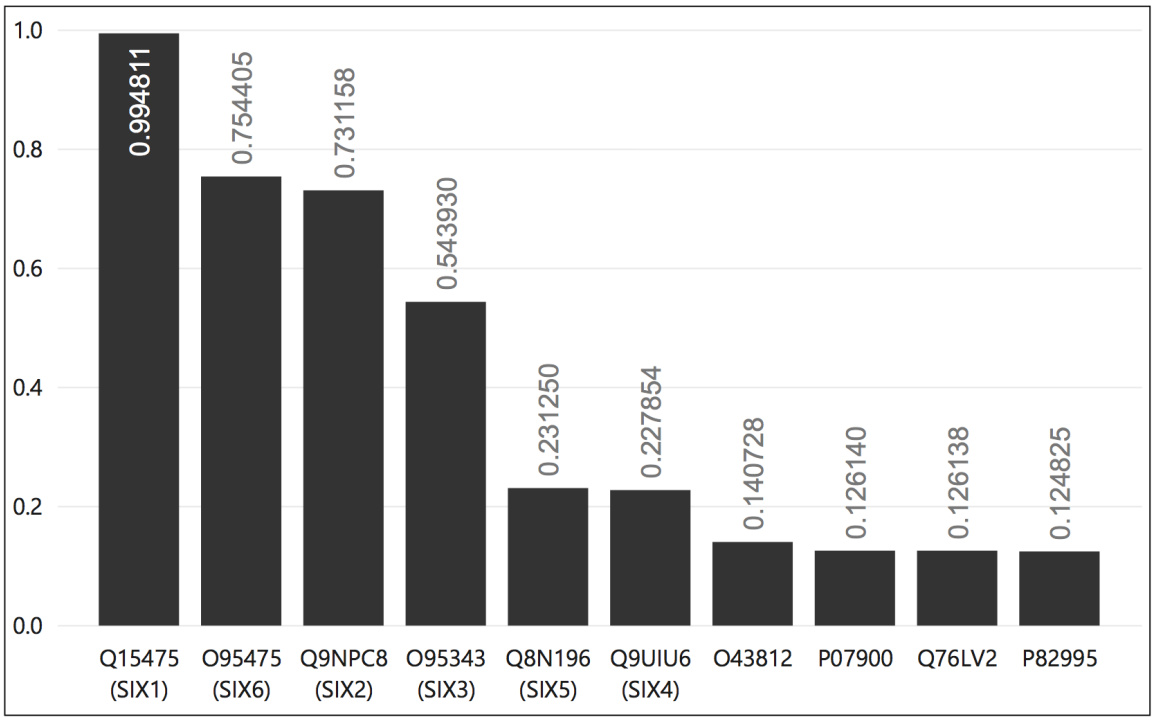


Figure 6.4: Top ten PIPE interaction scores for antiSix1-76405 against the rest of the proteins. The InSiPS scores (fitness, target, maximum non-target) were listed in Table 6.8. It is noted that the SIX proteins were excluded from the non-target list — *dataset1a* — but are plotted for the purpose of seeing the predicted interaction scores against proteins from the same family.

The top ten PIPE interaction scores for antiSix1-77087 are shown in Figure 6.5. The average for the predicted interactions against all the non-targets for this new sequence was 0.003556. It should be noted here that the interaction score for the target was lower because *dataset1b* was used for the list of non-targets; that is, it included the proteins from SIX2 to SIX6. Interestingly, proteins SIX4 and SIX5 were not in the top ten predicted interactions, but appear in respectively 28th and 26th positions.

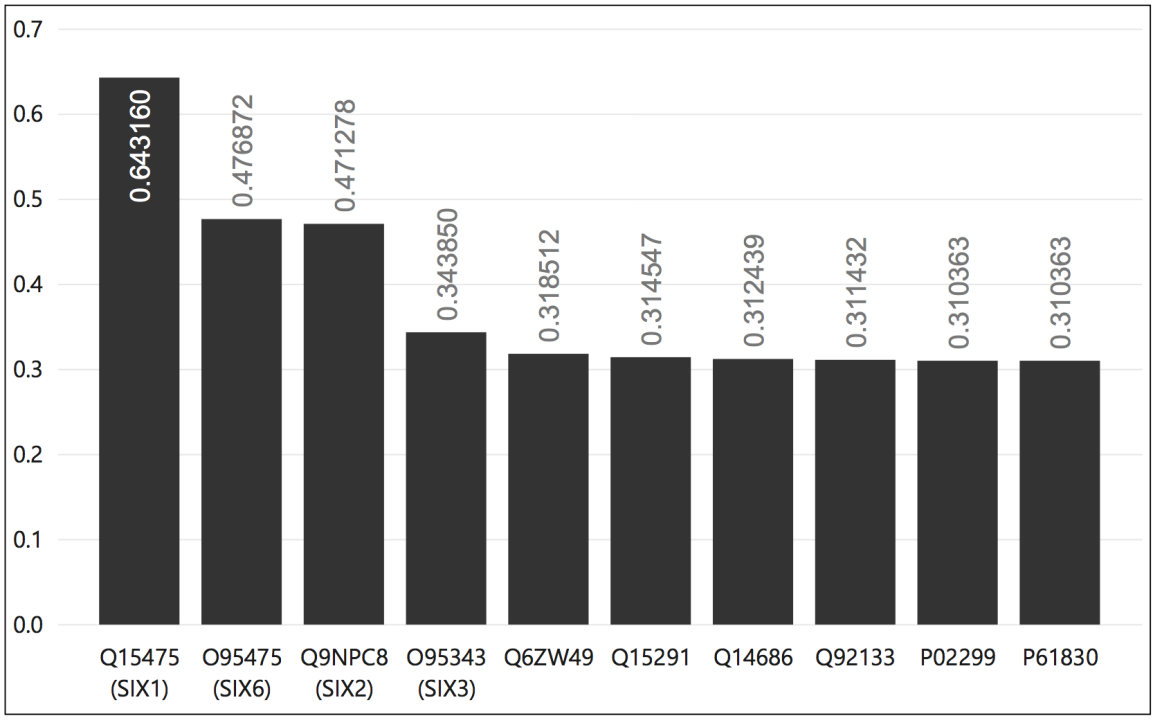


Figure 6.5: Top ten PIPE interaction scores for antiSix1-77087 against the rest of the proteins. The InSiPS scores (fitness, target, maximum non-target) were listed in Table 6.8. For the list of non-targets *dataset1b* was used. Therefore, the PIPE interaction score for the target (SIX1) was lower than that obtained for other antiSix1 sequences generated with *dataset1a* as the non-target list.

Lastly, the top ten PIPE interaction scores for antiSix4-88202 are shown in Figure 6.6. The average for the predicted interactions against all the non-targets for this new sequence was 0.000842. The interaction scores for this protein sequence were similar to the previous antiSix1-76405 sequence and the first predicted off-target (excluding SIX proteins) was the same protein (O43812) for both new sequences.

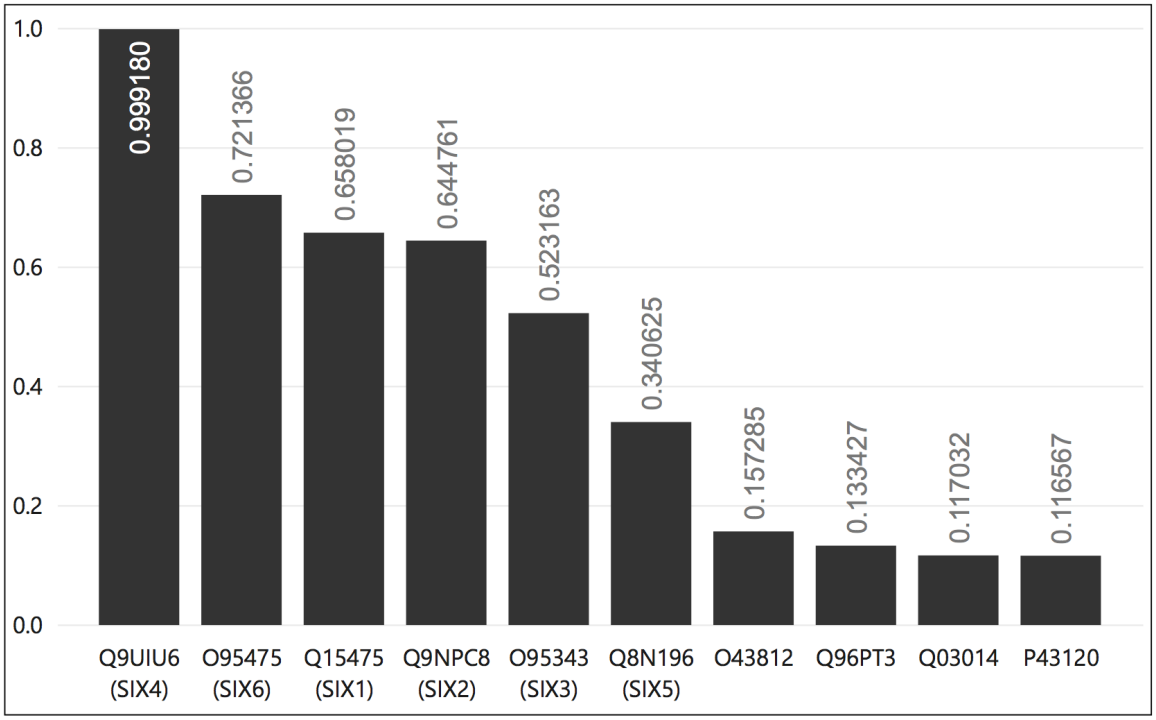


Figure 6.6: Top ten PIPE interaction scores for antiSix4-88202 against the rest of the proteins. The InSiPS scores (fitness, target, maximum non-target) were listed in Table 6.8. For the list of non-targets *dataset1a* was used.

6.7.3 New Anti-EYA Sequences and their Interactions

Similarly to the above-described anti-SIX sequences, the new sequences designed to target different EYA proteins and analyzed in Chapter 7 are displayed in Table 6.9. In contrast to anti-SIX sequences derived using *dataset1a* for the non-target list, the fitness values for the new anti-EYA sequences were lower. The target score was above 0.9 which is similar to that for the anti-SIX new sequences; however, the maximum non-target score was higher, thus affecting the fitness value.

New Sequence	Length	Non-targets	Fitness	Target	Max non-target
antiEya2-76843	35	dataset1a	0.560789	0.998796	0.438535
antiEya2-88118	35	dataset1a	0.619388	0.936296	0.338470
antiEya2-88203	35	dataset1a	0.611231	0.977722	0.374842

Table 6.9: List of InSiPS scores for new anti-EYA sequences of length 35. New sequences designed to target EYA2 for which output data is further analyzed in Chapter 7.

The top ten PIPE interaction scores for antiEya2-88118 against the rest of the proteins (target and non-targets) are shown in Figure 6.7. The InSiPS scores (fitness, target, maximum non-target) are listed in Table 6.9; the average for the predicted interactions against all the non-targets for this new sequence was 0.005614. The highest predicted interaction score is against the target protein (EYA2). To show how the new antiEya2 protein was predicted to interact with the EYA family proteins, these are also plotted; however, note they were excluded from the list of non-targets.

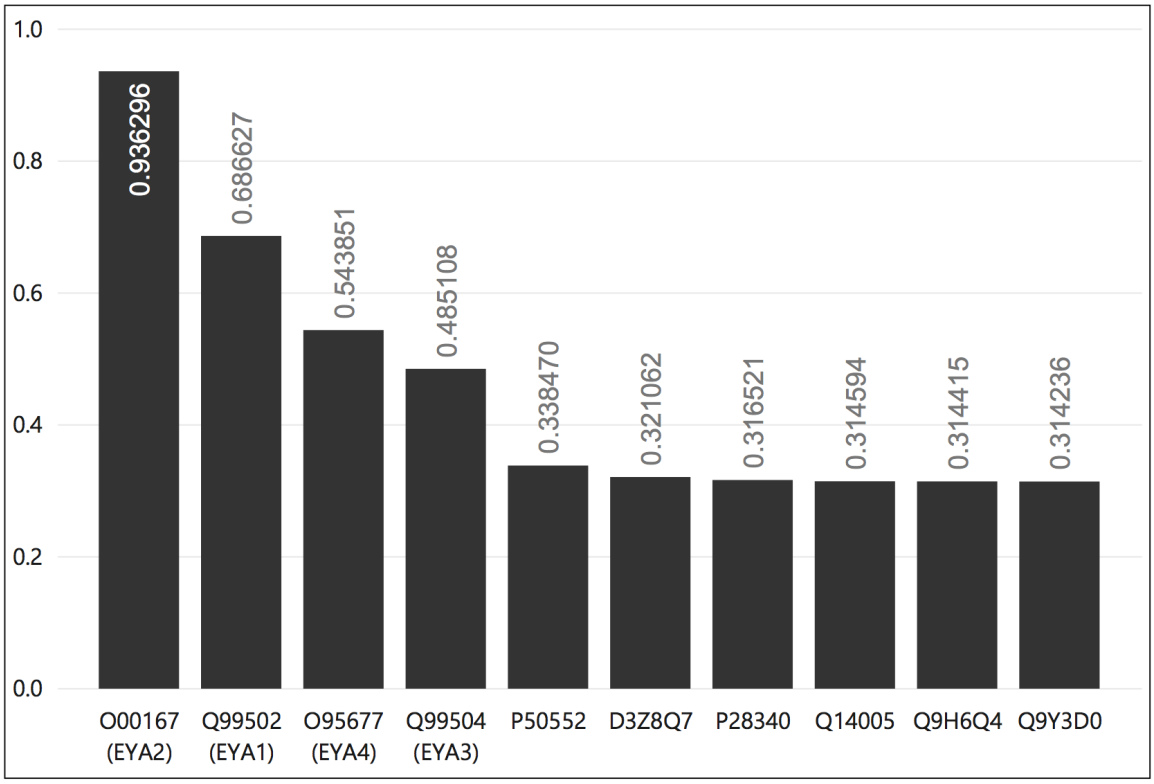


Figure 6.7: Top ten PIPE interaction scores for antiEya2-88118 against the rest of the proteins. The InSiPS scores (fitness, target, maximum non-target) were listed in Table 6.9

6.7.4 Validation of Results

The process for the validation of results of the new sequences generated in this phase of InSiPS experiments with human proteins was described in Section 4.3.4. For a given new sequence, the sensitivity and specificity values were calculated for the predicted interaction (PIPE) scores against target and non-targets. First, the leave-one-out cross-validation (LOOCV) was performed with the new human data, for which the positive set was the list of known protein pairs and the negative set was a random set of 100,000 protein pairs. Next, the results were ordered by PIPE score. Finally, for a given PIPE score (cutoff), the values for sensitivity and specificity were defined (see Section 4.2.5 for more details). The specificity to consider a predicted interaction as of high confidence of interaction was 99.95%. A sample of the values obtained from the LOOCV is given in Table 6.10. As can be observed in the table, a PIPE score equal to or above the cutoff of *0.69916594* would produce a specificity above 99.95%

and hence a high confidence of interaction. For example, a new sequence with a target score of 0.936296 would obtain a specificity of 99.98%.

Sensitivity	Specificity	Cutoff (*PIPE Score*)
5.88%	99.95%	0.69916594
4.48%	99.97%	0.81554061
3.29%	99.98%	0.90007275
1.23%	99.99%	0.99916500

Table 6.10: Sample of values from leave-one-out cross validation (LOOCV) with human data. Sensitivity and specificity for different PIPE scores (cutoffs). PIPE scores equal to or above the cutoff 0.69916594 produce a specificity above 99.95% and hence a high confidence of interaction.

To have a clearer idea of the accuracy of the results, the data obtained from the LOOCV was plotted as a receiver operating characteristic (ROC) curve. These plots have the TP (true positive) rate (specificity) on the y-axis and the FP (false positive) rate (1 - specificity) on the x-axis. The predicted (PIPE) interaction scores for a given new sequence against the target and 100 off-targets were plotted on the ROC curve to see how the TP rate and FP rate changed for each predicted interaction score.

The top ten predicted interactions for antiSix1-76405 were previously described (see Figure 6.4). For this new sequence the top 100 predicted interactions against target/off-targets is illustrated on the ROC curve in Figure 6.8, where the interaction scores represented by the yellow (target) and blue (off-targets) triangles go from highest to lowest starting with the yellow triangle for the target; thus, the higher the interaction score, the lower the FP rate. The predicted interaction score for this new sequence against the target was 0.994811 and therefore the calculated specificity value was above 99.98%; resulting in an extremely low FP rate of 0.02%. The cost of having a high specificity and consequently a low FP rate implies a low sensitivity, but these are desirable results keeping in mind the 600:1 ratio of negative to positive interactions explained in Section 4.2.5. As the interaction score decreases the FP rate increases, as indicated by the off-target triangles in the figure. It should be noted that the predicted interactions that follow the top interaction (with the target) are the other SIX proteins, SIX2-SIX6, which are the five off-target triangles next to the

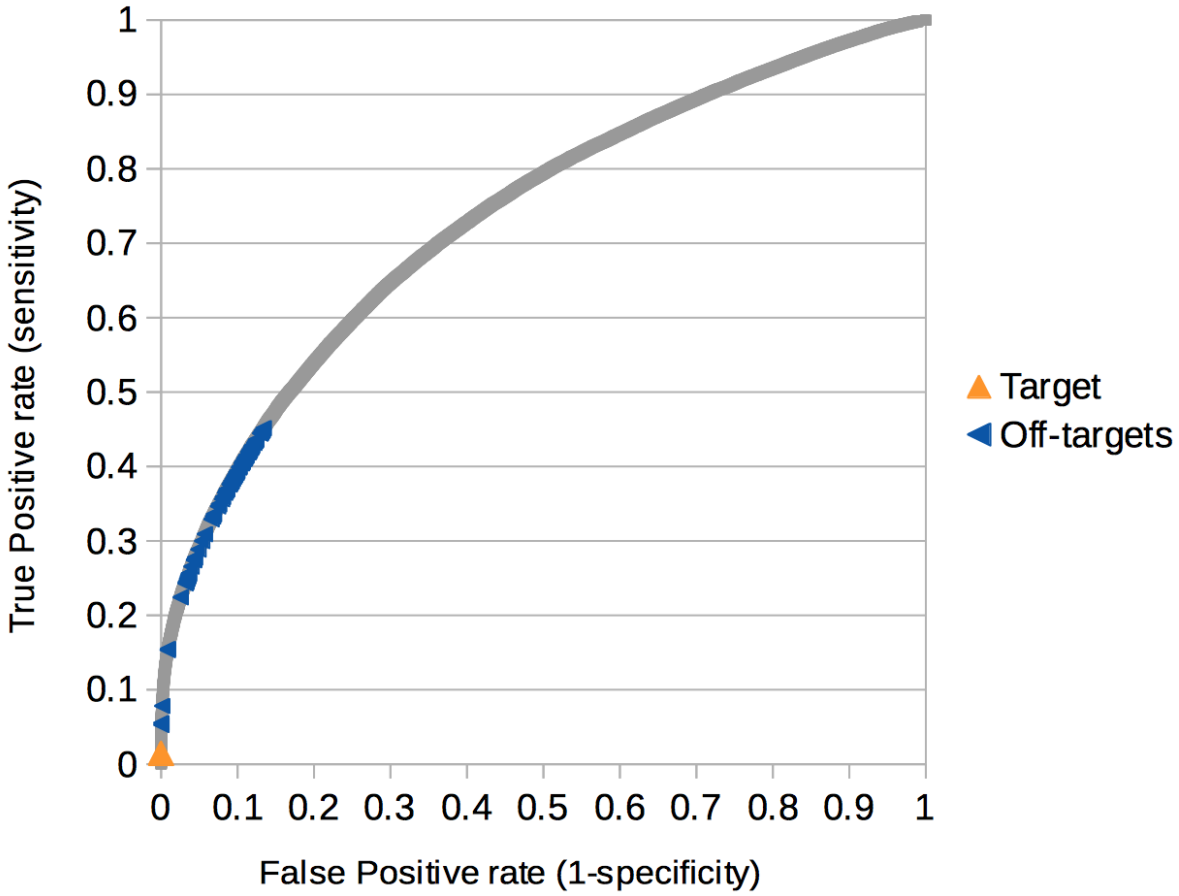


Figure 6.8: AntiSix1-76405 — PIPE score for top 100 interactions on the ROC curve. The predicted interaction score for this new sequence against the target SIX1 (yellow triangle) was 0.994811 and therefore the calculated specificity value was above 99.98%; resulting in an extremely low false positive (FP) rate of 0.02%.

target.

The top ten predicted interactions for antiEya2-88118 were described in Figure 6.7. The predicted top 100 interactions for this new sequence are shown in Figure 6.9. The PIPE score with the target (EYA2) was 0.936296 and therefore the calculated specificity was 99.98% and the FP rate 0.02%. The average of the predicted interactions for this new sequence and the non-targets was 0.33847 and this is reflected on the ROC curve. The calculated FP rate for the predicted off-targets that follow the top interaction (with the target) is lower (which means a higher specificity) in comparison to the above described results obtained for antiSix1-76405.

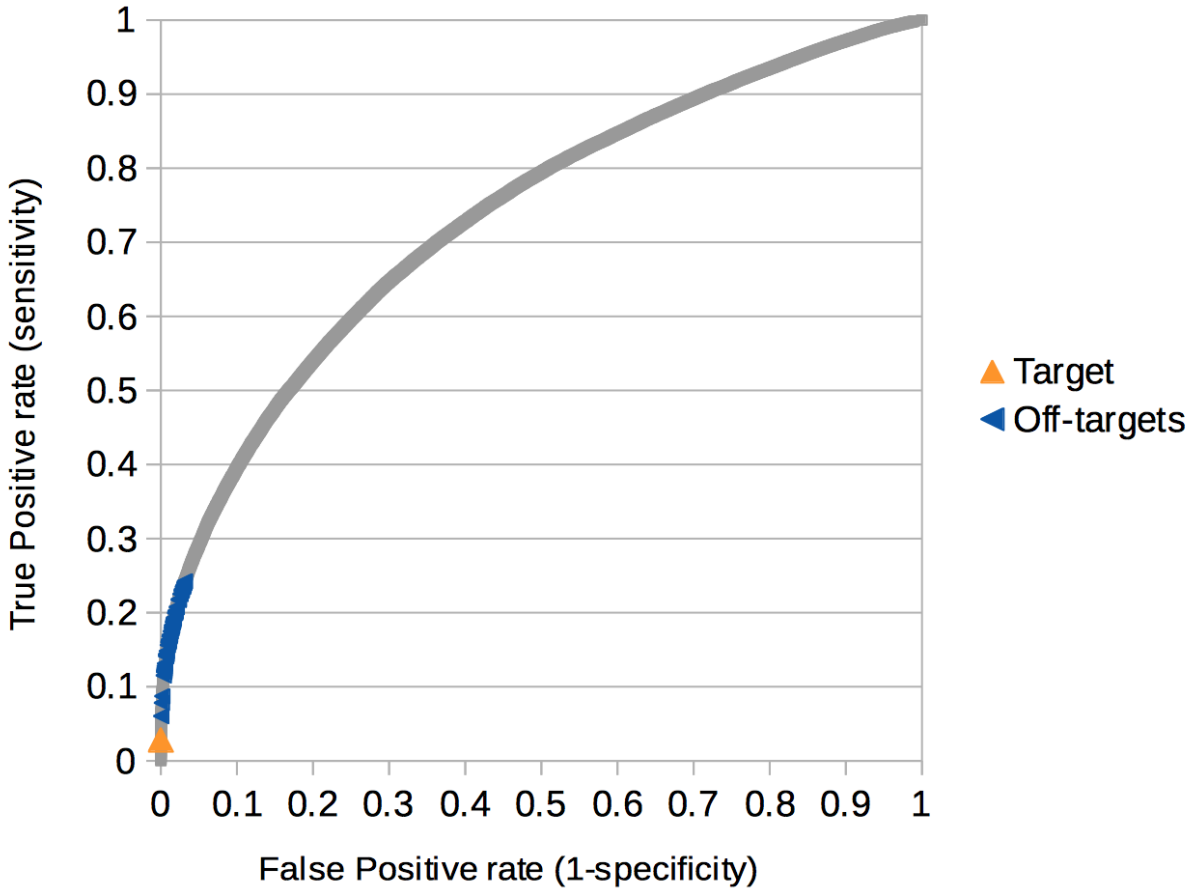


Figure 6.9: AntiEya2-88118 — PIPE score for top 100 interactions on the ROC curve. The predicted interaction score with the target EYA2 (yellow triangle) was 0.936296 and therefore the calculated specificity was 99.98% and the false positive (FP) rate 0.02%.

6.8 Prediction of Interaction Sites

PIPE-Sites proved to accurately predict protein binding sites between interacting proteins [85]. This functionality forms part of the InSiPS core algorithm, through PIPE, used for the prediction of PPIs between the new sequences against target and non-target proteins. However, further validation of the accuracy of the predicted interaction sites was still needed. The PIPE-Sites algorithm can give a good indication of the expected location of the interaction sites on a given pair of proteins. Furthermore, the predicted interaction sites could vary considerably from new protein sequence to new protein sequence aimed to interact with the same target protein.

Let us remember that each new protein sequence is unique. As mentioned in Section 4.3.1, the InSiPS genetic algorithm starts with a set of new protein sequences made of random amino acids. Then it continues to improve the pool of sequences using a fitness function, so that the predicted interaction score with the target is high and the predicted interaction score with the non-targets is low. Hence, the predicted binding sites between different new protein sequences (generated to interact with the same target protein) and specific human proteins might be quite different.

As previously mentioned in Section 2.2.3, the EYA-binding domain of SIX1 is near the n-terminus region on SIX1. Therefore, an antiSix1 protein with a higher likelihood to disrupt the SIX-EYA interaction would be one binding near the n-terminus region on SIX1 and not the central part where the homeodomain is. For most of the new protein sequences the predicted interaction site with the target was imprecise, as the predicted site was the whole target sequence. In spite of this, the predicted interaction sites for some off-target proteins were analyzed, for those off-targets known to be similar (in behavior or amino acid properties) to the target protein. The predicted interaction site for most of the off-targets was in the homeodomain region, which was an indicator that the target protein would be hit by the new sequence in the homeodomain region as well.

Therefore, another round of experiments with InSiPS was carried out trying to target the EYA-binding domain on SIX1. InSiPS takes as input the whole protein sequence given as target without considering a specific binding site. The workaround solution was to consider subsequences of the SIX1 protein as if they were whole protein sequences. Hence, more InSiPS runs with the new targets made from SIX1 subsequences were performed on the BGQ cluster.

The new antiSix1 proteins designed to bind specific regions of SIX1 and the produced InSiPS scores are detailed in the following section. The analysis of the predicted interaction sites for some of the new antiSix1 sequences with their target (SIX1) and family proteins of the target (SIX proteins) is described in Section 7.5.

6.8.1 New Sequences to Target EYA Binding Domain on SIX1

Based on the interaction site of interest on SIX1 described in 2.2.3, the following two regions of the SIX1 protein sequence were considered for the InSiPS experiments:

- **SIX1_9-118**. A region from amino acids 9 to 118 which is a binding site right before the homeodomain region: *FT QEQVACVCEV LQQGGNLERL GRFLWSLPAC DHLHKNESVL KAKAVVAFHR GNFRELYKIL ESHQFSPHNH PKLQQLWLKA HYVEAEKLRG RPLGAVGKYR VRRKFPLP*.

- **SIX1_9-27**. A region from amino acids 9 to 27 which is a narrower binding site on just the alpha helix 1 that mediates the interaction with EYA2: *FTQEQVACVCEVLQQGGNL*.

Considering the above described regions on SIX1, four subsequences of SIX1 were used as new targets as if they were whole protein sequences. The protein sequences used as targets are listed in Table 6.11. The protein sequence antiSix1_9-118 considers the region *SIX1_9-118* and antiSix1_1-128 the same region plus a window of a few more amino acids. The protein sequences antiSix1_1-37 and antiSix1_9-43 consider the region *SIX1_9-27* plus a few more amino acids to complete the length of 35. Note that a wider range of amino acids were added to a given subsequence from the SIX1 sequence aiming to give the PIPE algorithm (described in Section 4.2.2) more fragments to look at for similarities in the list of known PPIs.

Region	New Target Protein
SIX1_9-118	antiSix1_9-118
	antiSix1_1-128
SIX1_9-27	antiSix1_1-37
	antiSix1_9-43

Table 6.11: New target proteins made from SIX1 subsequences. The protein sequence antiSix1_9-118 considers the region *SIX1_9-118* (amino acid 9-118) and antiSix1_1-128 the same region plus a window of a few more amino acids. The protein sequences antiSix1_1-37 and antiSix1_9-43 consider the region *SIX1_9-27* (amino acid 9-27) plus a few more amino acids to complete the length of 35.

Execution on the BGQ Cluster

For this round of experiments ten new protein sequences were generated with InSiPS for each one of the four new targets as above defined (40 new sequences in total).

The input data used was the described in Section 6.2.2. The non-targets set used was *dataset1a*; that is, all the SIX family of proteins were excluded from the list of non-targets (i.e., 23,487 non-targets). The input GA parameters were those defined in Section 6.3.4 with 35 as the sequence length value. Like the previous new protein sequences generated with InSiPS, these new protein sequences were shown to be unique when evaluated with CD-HIT at 60% and 90% sequence identity.

The minimum number of generations was set to 250; however, the data distribution of the *hit generation* (described in Section 4.3) in Figure 6.10 shows that for new sequences designed to target antiSix1_1-128 the variation went from 43 as the minimum up to 303 as the maximum hit generation, with a median of 119; for the target antiSix1_9-118 the hit generation went from 39 up to 242 with a median of 85; for the target antiSix1_1-37 the hit generation went from 17 up to 212 with a median of 44; and for antiSix1_9-43 from 14 up to 249, with a median of 26. Although the hit generation highly depends on the initial generation of new sequences generated from random amino acids, it is particularly interesting to see how the variation in the distribution of hit generations for antiSix1_1-37 was narrower than for the other targets. This might indicate that for this specific subsequence of SIX1 it was easier to find similar fragments in the list of known protein-protein interactions.

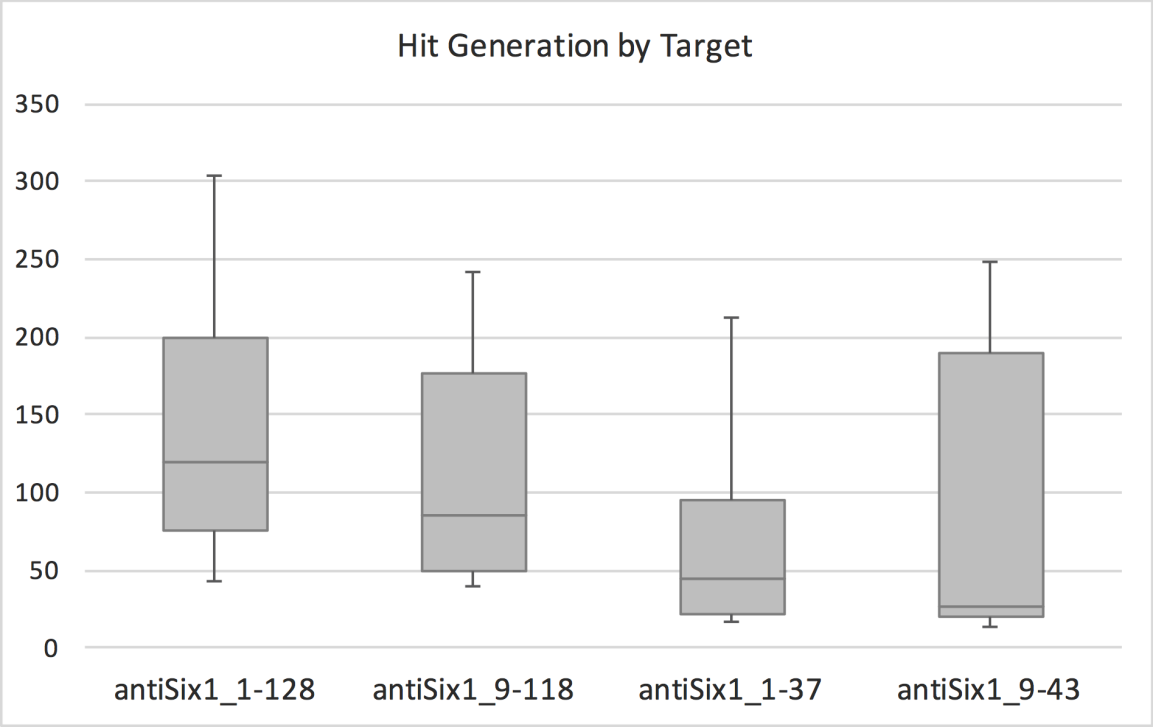


Figure 6.10: Distribution of *hit generations* for specific regions on SIX1 used as targets.

In regards to the runtime, as shown in Figure 6.11, the minimum time was around 11 hours for each of the four targets; the median between 11.4 and 12.3 hours; and the maximum time varied from 14.2 up to 17.2 hours. When the highest fitness value was found in generations 14, 44 or 119 (*hit generations*) the InSiPS algorithm still had to complete the predefined minimum number of generations of 250 and the *unchanged count* (generations with no improvement on the fitness values) of 50; therefore the total number of generations for these cases was 250. In the case of a *hit generation* of 242, the total number of generations would be 292 (242 plus 50 generations with no improvement on the fitness value). Therefore, the runtime would be proportional to the total number of generations and not to the hit generation.

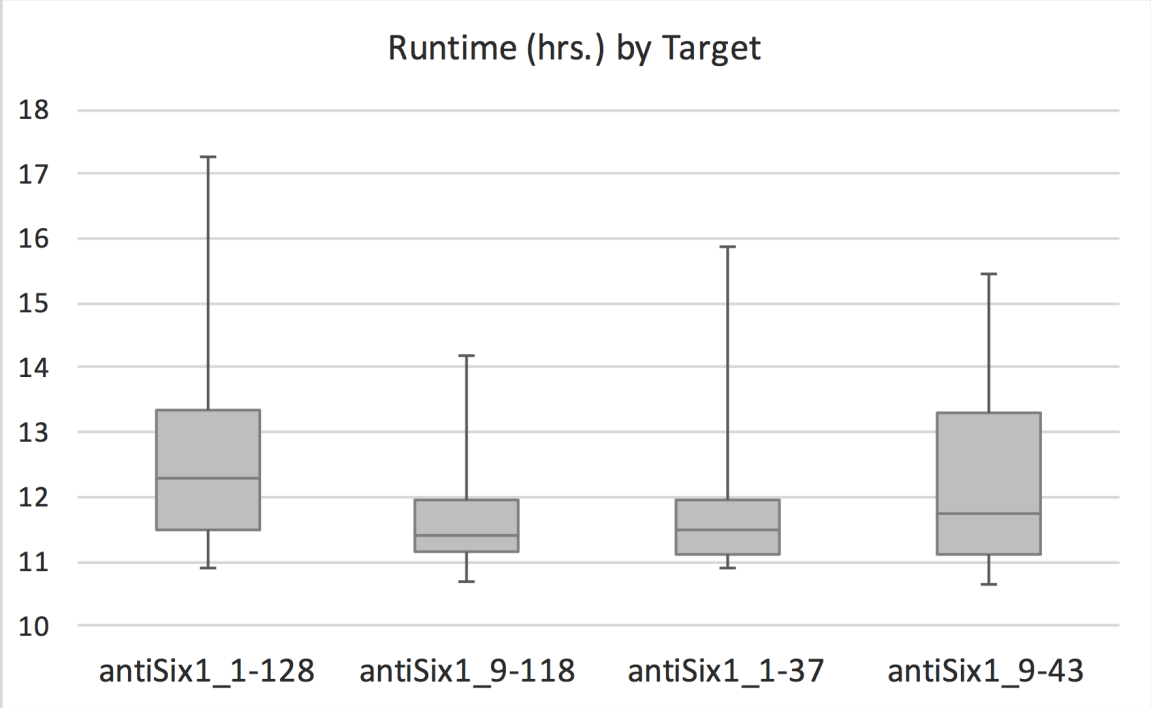


Figure 6.11: Distribution of runtime for specific regions on SIX1 used as targets.

General Results

The predicted interaction scores for the new sequences designed to target different regions on SIX1 and analyzed in Chapter 7 are displayed in Table 6.12.

In Figure 6.12 the average InSiPS scores obtained for each set of ten new sequences designed to target antiSix1_1-128, antiSix1_9-118, antiSix1_1-37, and antiSix1_9-43 are displayed. The average fitness value was similar to that produced for whole SIX1 sequences as targets, above 0.82. That is, the average predicted interaction score with the target protein was as high as the predicted scores for whole SIX1 sequences as targets, above 0.97; also, the average of the maximum predicted interaction score with the off-target proteins was as low as the predicted scores when having a whole SIX1 sequence as target, below 0.17.

The average of the target scores for the new sequences (above 0.97) described in this section therefore would produce a specificity above 99.98%. As previously described, according to the validation with this new human data, any value above a PIPE score of 0.69 would achieve a specificity of 99.95% and could be labeled as a high confidence of interaction. The analysis of the interaction sites for these new

sequences is described in Section 7.5.2.

New Sequence	Length	Non-targets	Fitness	Target	Max non-target
antiSix1_1-128-88401	35	dataset1a	0.846467	0.990826	0.145695
antiSix1_1-37-88406	35	dataset1a	0.832643	0.986111	0.155629
antiSix1_1-37-88692	35	dataset1a	0.833460	0.986111	0.154801
antiSix1_1-37-88693	35	dataset1a	0.832643	0.986111	0.155629
antiSix1_1-37-88695	35	dataset1a	0.832643	0.986111	0.155629
antiSix1_9-118-88767	35	dataset1a	0.846288	0.98489	0.140728
antiSix1_9-43-88775	35	dataset1a	0.831178	0.984375	0.155629
antiSix1_9-118-88803	35	dataset1a	0.664608	0.813874	0.183402
antiSix1_9-118-88805	35	dataset1a	0.849481	0.997253	0.148179
antiSix1_9-43-88814	35	dataset1a	0.831178	0.984375	0.155629

Table 6.12: List of InSiPS scores for antiSix1 sequences designed to bind specific regions on SIX1. New sequences designed to target the EYA binding domain on SIX1 for which output data is further analyzed in Chapter 7.

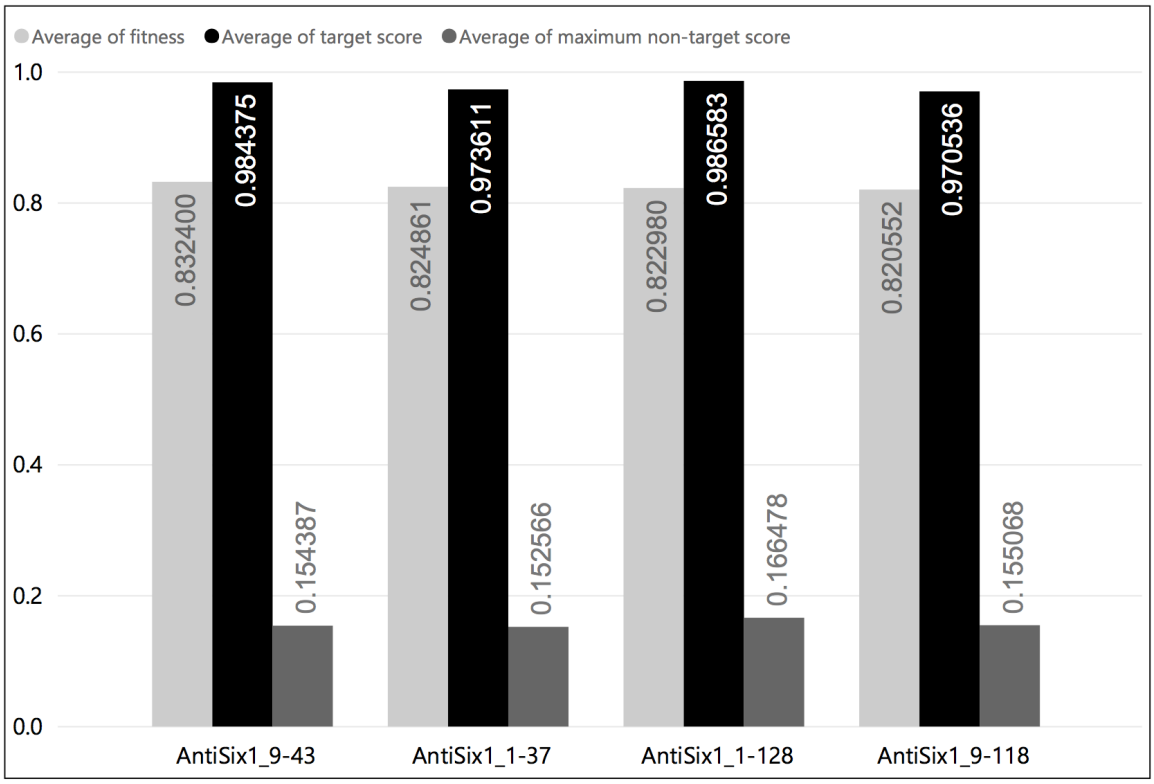


Figure 6.12: Average of InSiPS scores obtained for antiSix1 sequences designed to target the EYA binding domain on SIX1. In total 40 new sequences were generated with InSiPS: ten new protein sequences for each one of the four new targets (antiSix1_9-43, antiSix1_1-37, antiSix1_128 and antiSix1_9-118).

Chapter 7

Data Analysis of Results

7.1 Introduction

This chapter summarizes the analysis of the InSiPS results for some significant new protein sequences. As was mentioned in Chapter 2, further to previous research on SIX and EYA protein interactions, the SIX1 and EYA2 proteins were of greater interest to be candidate target proteins for InSiPS runs. Furthermore, the computational results indicated that new sequences designed to interact with either the SIX1 or EYA2 proteins as targets obtained good results in terms of the PIPE predicted interaction score, also known as the PIPE score. Additionally, other new sequences designed to interact with the SIX4 and SIX6 proteins were predicted to have a high likelihood of interaction with the target.

This analysis preceded and was determinant for the subsequent phase of experimental validation and helped to determine the new protein sequences considered for experimental validation. Presented first is a short report of the preliminary analysis carried out before the start of this research project. Also included is an analysis for a proposed new protein sequence, looking at some of its off-target proteins. The main sections of this chapter discuss the sequence homology analysis for some relevant new sequences, and also the analysis of the predicted interaction sites.

7.2 Preliminary Analysis

When this research began, InSiPS had not been used with human proteins; hence a preliminary analysis was first performed. This preliminary work consisted of a PIPE run to predict and evaluate the interactions for SIX family proteins. At the time

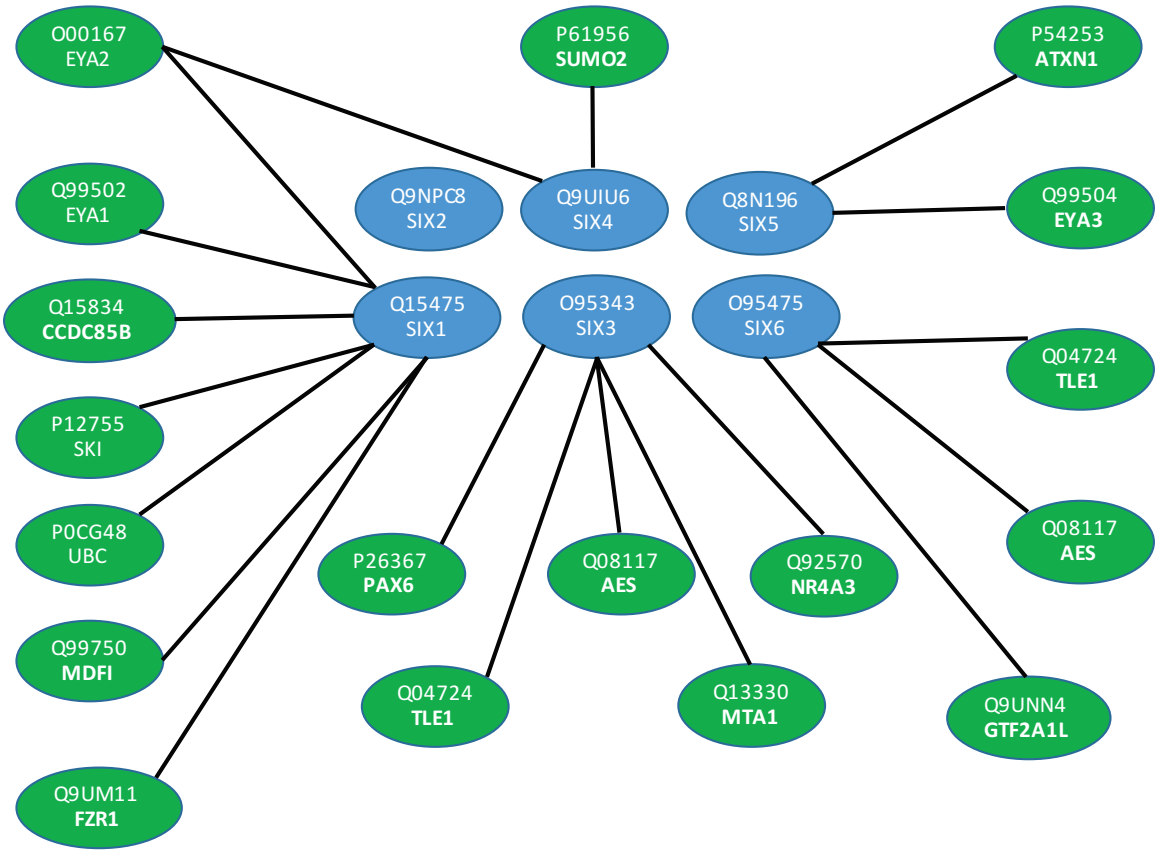


Figure 7.1: Old Human data — Known interaction pairs for SIX family proteins. Blue nodes represent the SIX family proteins and green nodes their known interacting partners.

of this analysis, the collection of a new validated data set of human proteins and interactions had not been started. Hence, the data set used was the old human data set, described in Section 6.2.1. This old data contained eighteen known interaction pairs for the SIX family proteins, shown in Figure 7.1.

The PIPE run predicted twenty-seven novel interaction pairs for SIX family proteins (Figure 7.2) at a specificity of 99.95%, which was considered the standard operating point, as explained in Section 4.2.5.

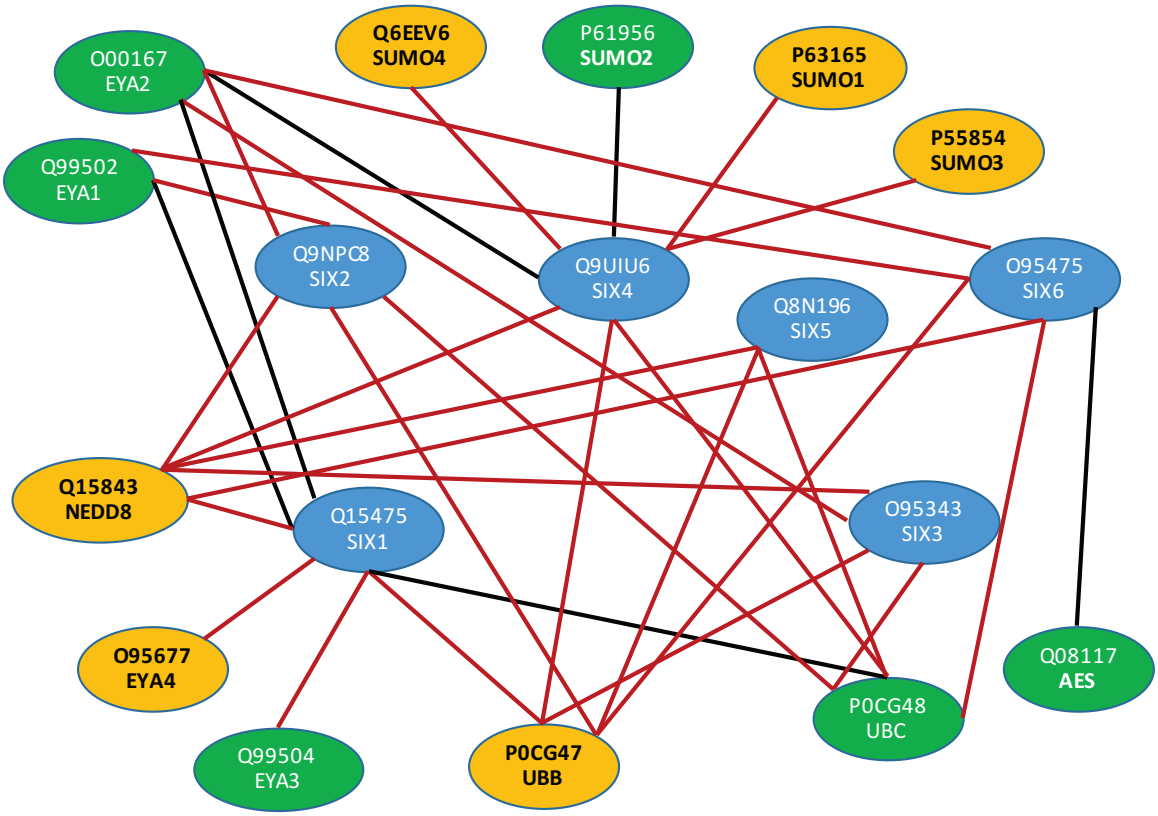


Figure 7.2: Old Human data — Predicted interaction pairs for SIX family proteins. Blue nodes represent SIX family proteins, green nodes with black edges represent predicted interactions that were previously known and green nodes with red edges represent novel predicted interactions. Yellow nodes represent novel predicted interactions of proteins not present in the known interactions list.

Proteins Predicted to Interact with SIX1

At a specificity of 99.95%, *Ubiquitin* (UBB and UBC) and *Ubiquitin-like* proteins (NEDD and SUMO) were found at the very top of the interaction pairs predicted by the PIPE run. Furthermore, EYA4 and EYA3 proteins were predicted to interact with SIX1 at this same specificity. EYA4 was not on the known protein interactions list, but it has been reported in the BioPlex Network [36]. Some proteins whose predicted interaction score was slightly below the PIPE threshold were also of interest as these are potential false negatives. For example, proteins of the families *Transducin-like enhancer proteins* (TLE1/2/3/4), *Amino-terminal enhancer of split* (AES), *Ski oncogene* (SKI) and *Ski-like protein* (SKIL) were found at the top of the list of off-target interactions.

Other proteins involved in the DNA damage response such as *cellular tumor antigen p53* (TP53) and *Breast cancer type 1 susceptibility protein* (BRCA1) [92, 93], were also found reasonably close to the top of the list of non-predicted interactions. Coincidentally, biologists who prepared this review had studied the possibility of a connection between these proteins and the SIX family of proteins. Finally, *Histone acetyltransferase p300* (EP300), a protein that the reviewers had been studying as an interacting partner of SIX1, was also found on the list of PPI predictions for SIX1.

Taking into account that a new human data set would be used for the InSiPS runs in the research, and anticipating that this fact would increase the likelihood of getting better results, this preliminary analysis was crucial in determining to start this project.

7.3 Sequence Homology

Sequence alignment helps to analyze the degree to which two proteins are similar and, by extension, how similar their interactions or functions may be. Sequence homology analysis was essential to determine whether the new sequences would be useful as synthetic proteins that would interact with a specific target. Hence, the BLASTP tool (Standard Protein — Basic Local Alignment Search Tool) [94] was used to validate how similar these new protein sequences were to experimentally validated SIX- or EYA-binding proteins. This section outlines the results for some of the sequence-homology analysis done for both new anti-SIX and anti-EYA sequences.

It is worth mentioning that, before starting this analysis, to avoid duplicates, it

was confirmed that the newly generated protein sequences were unique. The *CD-HIT* tool described in Section 6.3.1 was used to assure that the new sequences to be analyzed would represent different candidates.

After analyzing the generated new anti-SIX protein sequences, it was found that some of them were very similar to proteins known to bind SIX family proteins. Likewise, some anti-EYA sequences were found to be similar to proteins known to bind EYA family proteins. It should be noted that these known interactions influenced the new protein sequences generated, as most of the known protein interactions were already in the input data. The high PIPE score between the new protein sequence and the target was due to occurrences of similar fragments in the list of known protein interactions used in the InSiPS run. The InSiPS scores for the new protein sequences analyzed in this section were described in Section 6.7.

7.3.1 New Anti-SIX Protein Sequences

AntiSix1 and Myosin-8 (MYH8)

Some antiSix1 proteins generated by InSiPS resembled *Myosin heavy chain* proteins (MYH). In particular, two antiSix1 sequences resembled *Myosin-8* (MYH8, P13535). The two new sequences antiSix1-76405 and antiSix1-76629 are unique, but both sequences aligned to the same spot on MYH8. As illustrated in Figure 7.3, antiSix1-76405 mapped to amino acids 1431-1463 and antiSix1-76629 mapped to amino acids 1436-1470. This can be seen as analogous to convergent evolution, as the two sequences of random amino acids, that started independently, evolved until reaching a point where they align to the same spot on the same protein sequence. MYH3, MYH4 and MYH7 proteins have been reported as SIX1-binding proteins [95] and were part of the known interactions; however MYH8 was not part of this input data.

AntiSix1 and Vitronectin (VTN)

Two other antiSix1 proteins matched *Vitronectin* (VTN, P04004) at different places on the protein (Figure 7.4). AntiSix1-77087 matched to amino acids 41-63, and antiSix1-76307 matched to amino acids 322-354. VTN has been reported as a SIX1-binding protein in yeast two-hybrid assays [96] and it was also part of the known interactions list; it paired only with SIX1.

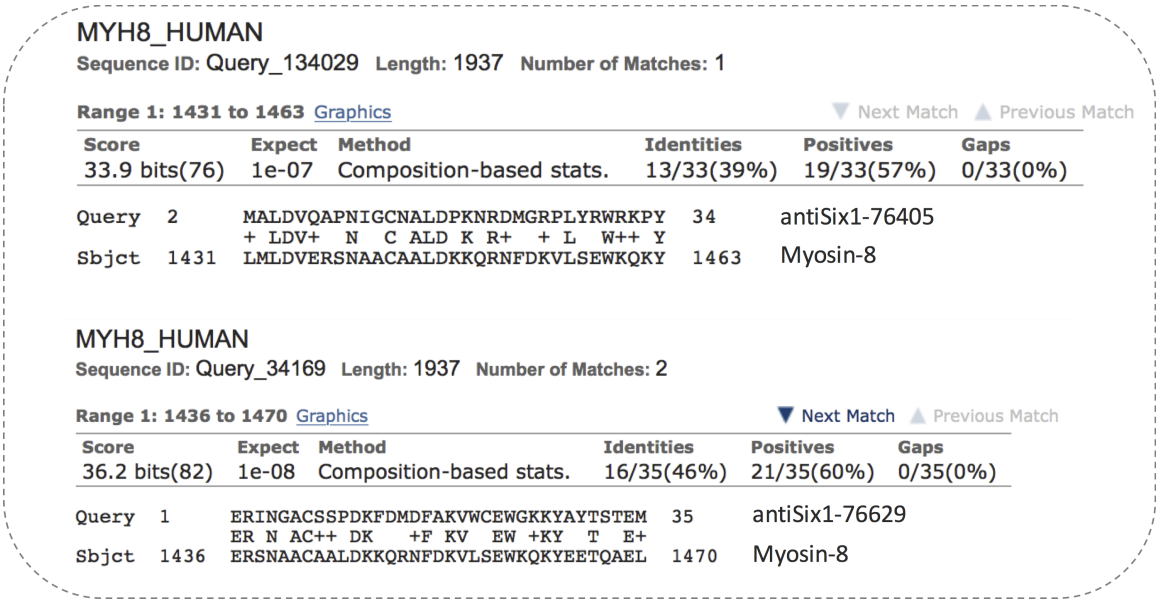


Figure 7.3: Alignment between antiSix1 proteins and Myosin-8.

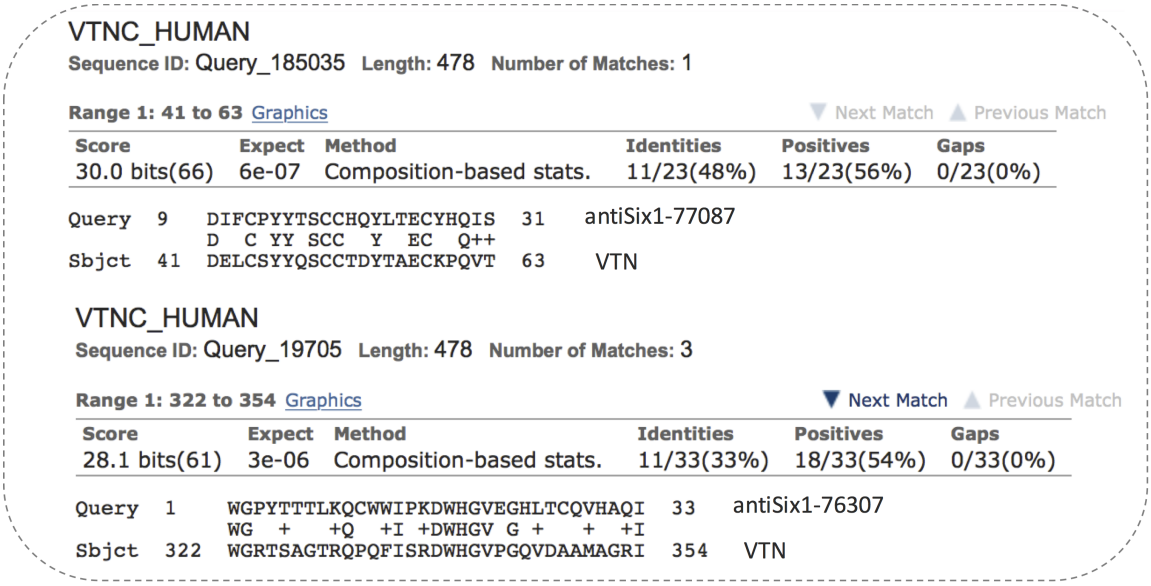


Figure 7.4: Alignment between antiSix1 proteins and Vitronectin.

AntiSix4 and Leucine-rich repeat-containing protein 46 (LRRC46)

Some antiSix4 sequences were found to have a certain similarity to a known SIX4-binding protein, *Leucine-rich repeat-containing protein 46* (LRRC46, Q96FV0), which had previously been reported to pull-down with SIX4 [36]. LRRC46 only occurs once in the known interaction list (with SIX4) and has a length of 321 amino acids. The new sequences, antiSix4-77411 and antiSix4-88115 (Figure 7.5), have different sequences, but they mapped to almost identical parts of LRRC46. AntiSix4-77411 mapped from amino acids 266-292, and antiSix4-88115 from amino acids 265-291.

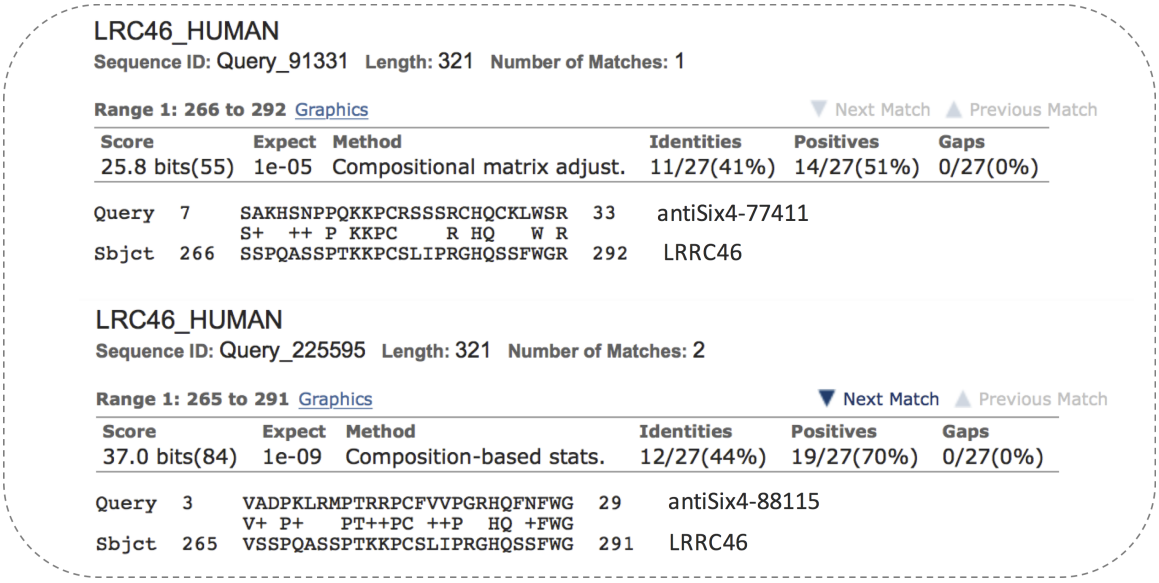


Figure 7.5: Alignment between antiSix4 proteins and Leucine-rich repeat-containing protein 46.

The new sequence antiSix4-88202 (Figure 7.6) was found to be very similar to a region of LRRC46 (amino acids 187-213), close to the region where the previous new sequences mapped.

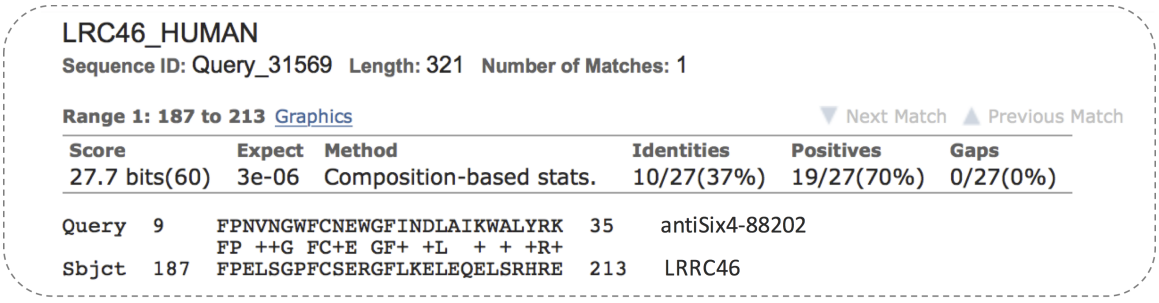


Figure 7.6: Alignment between antiSix4-88202 and Leucine-rich repeat-containing protein 46.

AntiSix6 and TFIIA-alpha and beta-like factor (GTF2A1L)

One antiSix6 protein was similar to a region in *TFIIA-alpha and beta-like factor* (GTF2A1L, Q9UNN4). AntiSix6-77448 mapped to amino acids 449-466 on GTF2A1L, as shown in Figure 7.7. GTF2A1L has been reported to interact with SIX6 [97]. This interaction was part of the list of known protein interactions, where there is only one occurrence for GTF2A1L.

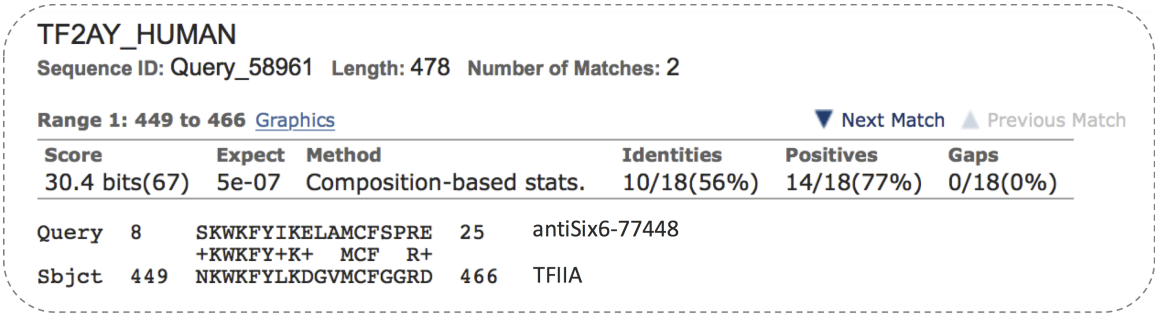


Figure 7.7: Alignment between antiSix6-77448 protein and TFIIA-alpha and beta-like factor.

7.3.2 New Anti-EYA Protein Sequences

AntiEya2 and Guanine nucleotide-binding protein G(i) subunit alpha-2 (GNAI2)

As described before, anti-SIX proteins had similarity with known SIX-binding proteins. Likewise, anti-EYA proteins had similarity with known EYA-binding proteins. As shown in Figure 7.8, antiEya2-88118 was similar to *Guanine nucleotide-binding*

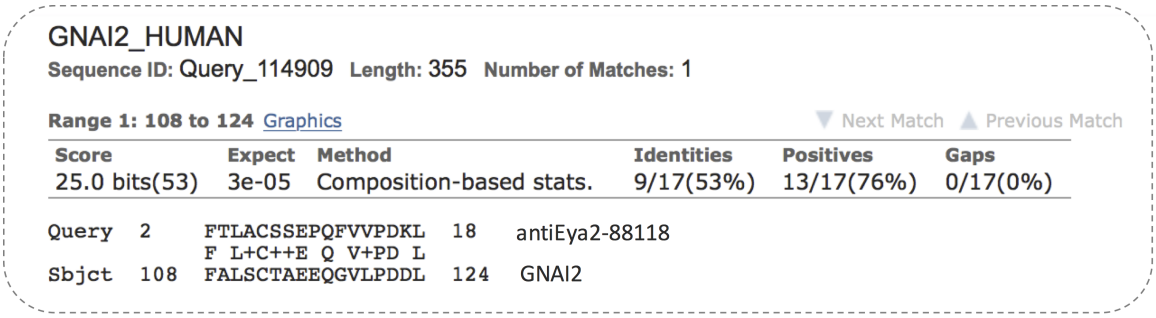


Figure 7.8: Alignment between antiEya2-88118 and Guanine nucleotide-binding protein G(i) subunit alpha-2.

protein G(i) subunit alpha-2 (GNAI2, P04899). GNAI2 has been reported as an EYA2-binding protein in BioGrid by two-hybrid and reconstituted complex experiments [98]. GNAI2 is also part of the known interaction list, with nine occurrences, of which one is with the EYA2 protein.

AntiEya2 and Fizzy-related protein homolog (FZRI)

AntiEya2-88203, as displayed in Figure 7.9, was also found to be similar to the known EYA1-binding protein *Fizzy-related protein homolog* (FZR1, Q9UM11), and reported in BioGrid [99]. FZR1 is in the known interactions list, with 23 occurrences, of which one is an interaction with EYA1 and another with SIX1.

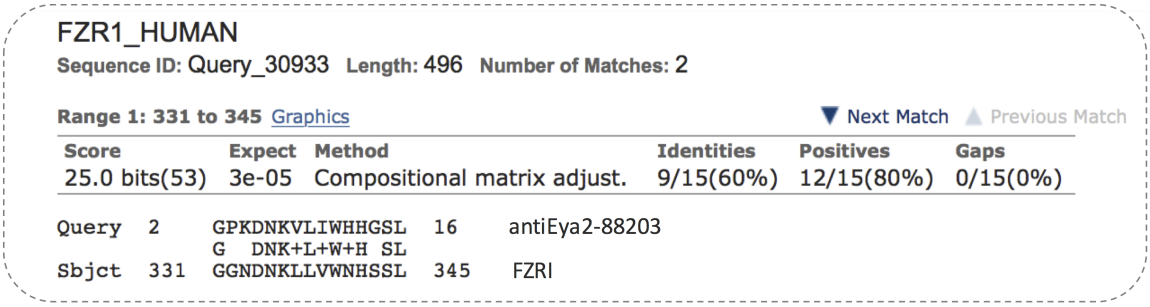


Figure 7.9: Alignment between antiEya2-88203 and Fizzy-related protein homolog.

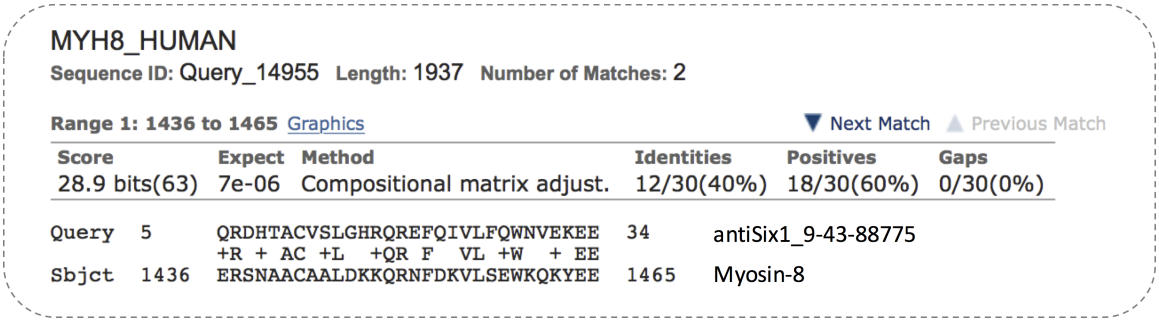
7.3.3 New AntiSix1 Protein Sequences for Specific Sites

The homology between known SIX-binding proteins and antiSix1 sequences was detailed in Section 7.3.1. The following describes the homology analysis for new antiSix1

sequences designed to interact with specific sites on SIX1. As described in Section 6.8.1, the name of the proteins indicates the amino acids they were designed to target. For example, antiSix1_9-43-88775 was designed to bind SIX1 from amino acids 9-43.

AntiSix1 for Specific Sites and Myosin-8 (MYH8)

As shown in Figure 7.10, the new antiSix1_9-43-88775 aligned to MYH8 on a region (amino acids 1436-1465) very similar to the region where the previously described antiSix1-76405 and antiSix1-76629 sequences (Figure 7.3) mapped to MYH8 (amino acids 1431-1463 and 1436-1470).


MYH8_HUMAN
Sequence ID: Query_14955 Length: 1937 Number of Matches: 2

Range 1: 1436 to 1465 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
28.9 bits(63)	7e-06	Compositional matrix adjust.	12/30(40%)	18/30(60%)	0/30(0%)

```
Query  5     QRDHTACVSLGHRQREFQIVLFQWNVEKEE  34    antiSix1_9-43-88775
             +R + AC +L  +QR F  VL +W  + EE
Sbjct  1436  ERSNAACAALDKKQRNFDKVLSEWKQKYEE  1465  Myosin-8
```

Figure 7.10: Alignment between antiSix1_9-43-88775 and Myosin-8.

Likewise, antiSix1_1-37-88692 and antiSix1_1-37-88693 mapped to MYH8 from amino acids 822-845 and 821-850 (Figure 7.11), again at a very similar spot. For these latter two proteins the predicted interaction site is different to the previous example, but the idea of seeing this as analogous to convergent evolution can also be identified on this analysis.

AntiSix1 for Specific Sites and Vitronectin (VTN)

As previously described (see Figure 7.4), antiSix1-77087 and antiSix1-76307 matched VTN, a known SIX1-binding protein, at different regions (amino acids 41-63 and 322-354). Also, at a different region (amino acids 192-212), antiSix1_1-37-88695 mapped to VTN, as described in Figure 7.12. However, it is worthy of note that, as shown in Figure 7.13, antiSix1_1-37-88406 and antiSix1_9-43-88814 mapped to VTN in a very similar region (amino acids 293-317 and 289 and 314).

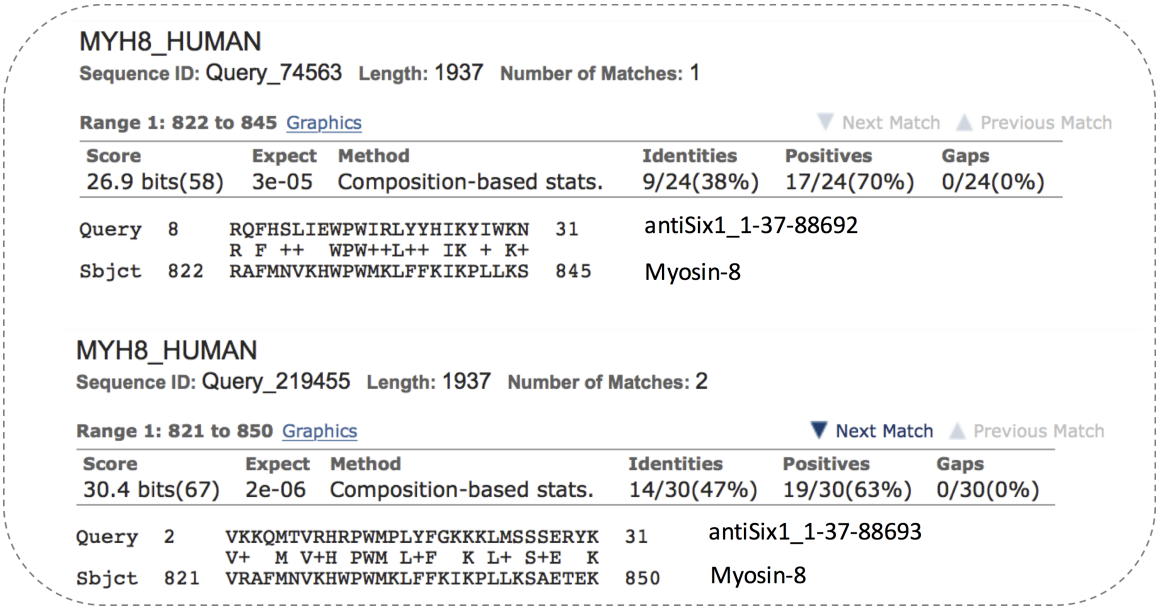


Figure 7.11: Alignment between antiSix1 for specific regions on SIX1 and Myosin-8.

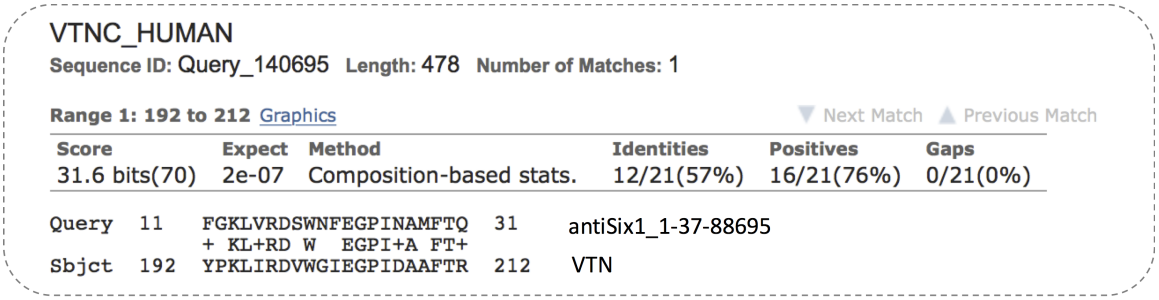


Figure 7.12: Alignment between antiSix1_1-37-88695 and Vitronectin.

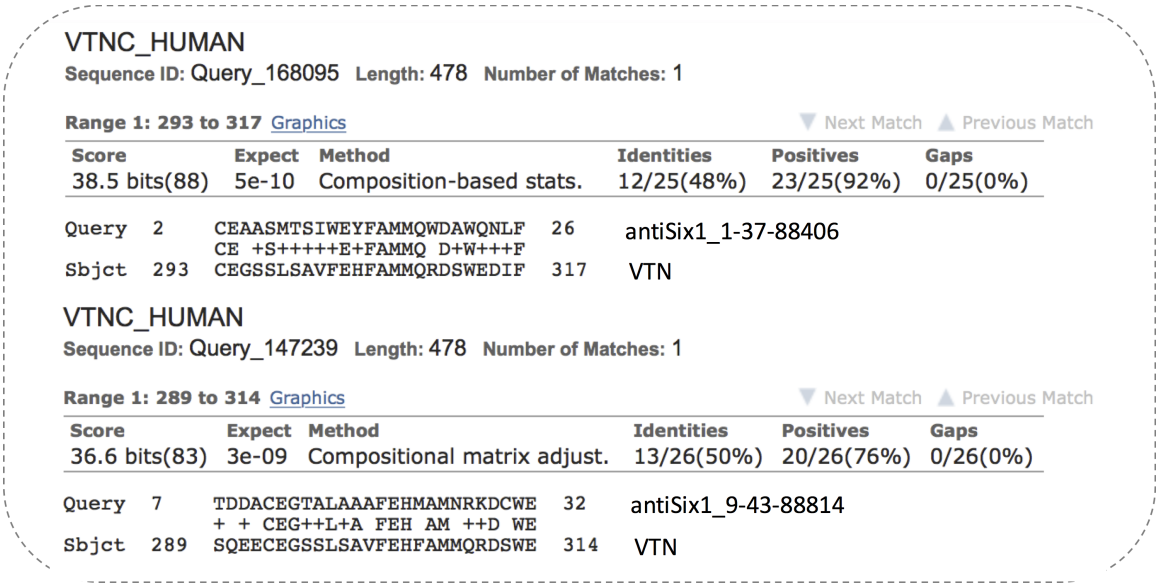


Figure 7.13: Alignment between antiSix1 for specific regions on SIX1 and Vitronectin.

7.4 Off-targets of New Anti-EYA Protein Sequences

As explained in Section 4.3.3, the off-targets are those proteins with which the new protein sequence is predicted to have some degree of interaction. Besides the analysis of how similar the new protein sequences were to suspected SIX- or EYA-binding proteins, it was considered of importance to analyze the homology of new sequences with the off-targets proteins.

For antiEya2-76843, one of the top off-target proteins was the protein *Leucine-rich repeat-containing protein 46* (LRRC46, Q96FV0). This protein has been reported to pull-down with SIX4 in the BioPlex Network [36]. Another off-target for antiEya2-76843 was *MyoD family inhibitor* (MDFI, Q99750), which is reported as a SIX1-binding protein, as determined by *Two-hybrid* and *Affinity Capture-Western* experiments, in [100]. Furthermore, another off-target on the list was *Ski oncogene* protein (SKI, P12755), which is also reported as a SIX1-binding protein. However, it is unknown whether the interaction is direct or not [101].

This analysis suggests that the new protein sequences designed to interact with EYA2 as a target are predicted to have some degree of interaction with proteins that bind SIX proteins. This only adds confidence to our newly designed proteins as one of

the properties of EYA is to bind SIX [27,28]. The results indicated that the off-targets of antiEya2-76843 share this property with EYA proteins. It has to be emphasized that the three known interactions involving the off-target proteins mentioned in this section are also reported in the known protein interactions list used to run InSiPS. LRRC46 is reported in the known interaction list only once, interacting with SIX4. MDFI is reported in 46 pairs and one of those interactions is with SIX1. SKI is reported 10 times, from which one interaction is with SIX1. The InSiPS scores for antiEya2-76843 are listed in Section 6.7.3.

7.5 Predicted Interaction Sites

This section describes the analysis performed to identify the predicted interaction sites for new antiSix1 sequences. Due to the importance of knowing the binding site to increase the chance of a successful experimental validation, this output data was evaluated before proceeding with the actual wet-lab biochemistry experiments. This part of the analysis was only done for protein sequences designed to target SIX1, because of time constraints to start the actual experiments. SIX1 was also identified as the best candidate to target due to previous homology analysis and the high PIPE scores with the target proteins.

The EYA-binding domain on SIX1 is near the n-terminus region (see Section 2.2.3 for more details). Therefore, an antiSix1 protein sequence predicted to bind SIX1 near the n-terminus region would represent a synthetic protein with a higher likelihood to disrupt the SIX-EYA interaction. The InSiPS scores for the new protein sequences analyzed in this section were described in Section 6.7.

7.5.1 Analysis for SIX Proteins as Target

For most of the new protein sequences designed to target SIX1, the predicted binding site with the target protein was not specific. For example, for SIX1 as a target, the interaction site was from amino acids 1-284, which is the whole SIX1 protein sequence. Likewise, for SIX4 as a target, the predicted interaction site was the whole length of the SIX4 protein, that is from amino acids 1-781. A high fitness value indicates that the new sequences would have a high likelihood of interaction with the target protein and a low likelihood of interaction with the non-targets. Keeping in mind the high fitness value predicted by InSiPS, the new protein sequences were still

considered trustworthy. However, because the predicted interaction of the target was not specific, and taking into account the high degree of similarity among proteins from the same family, the predicted interaction sites with other proteins from the target family (SIX) were analyzed.

As can be observed in Figure 7.14, the predicted interaction sites for antiSix1-76407 with the target SIX1 was the whole SIX1 sequence. However, the interaction site with other proteins of the SIX family, such as for SIX3, SIX5 and SIX6, was located on a region before the homeodomain of the corresponding SIX protein. In general, most of the new sequences designed to target SIX1 and SIX4 proteins were predicted to interact with SIX3 proteins in a region before where the SIX3 homeodomain is located, as is shown in Figure 7.15.

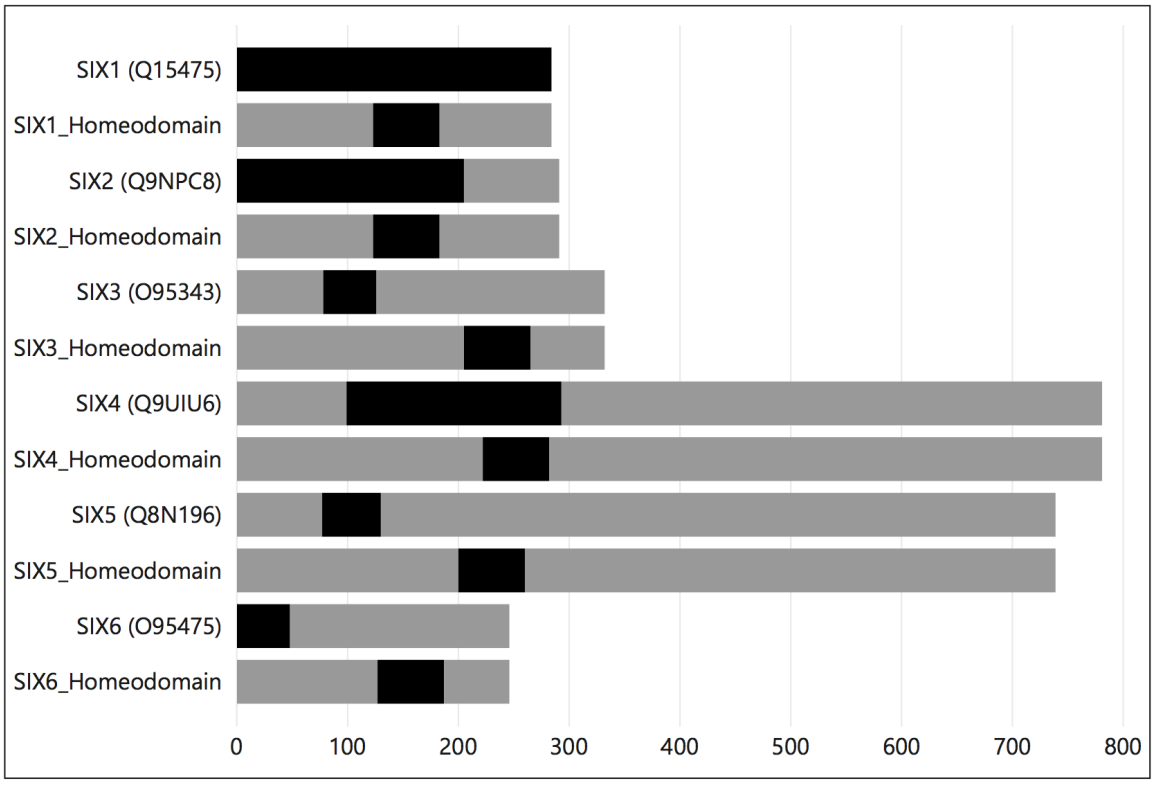


Figure 7.14: Predicted interaction sites for antiSix1-76407 against SIX proteins. Each SIX protein sequence is presented two times: first, the protein sequence — e.g., SIX1 (Q15475) — where the shadowed area represents the predicted interaction site with antiSix1-76407; second, the protein sequence — e.g., SIX1_Homeodomain — where the shadowed area represents the homeodomain region of the SIX protein.

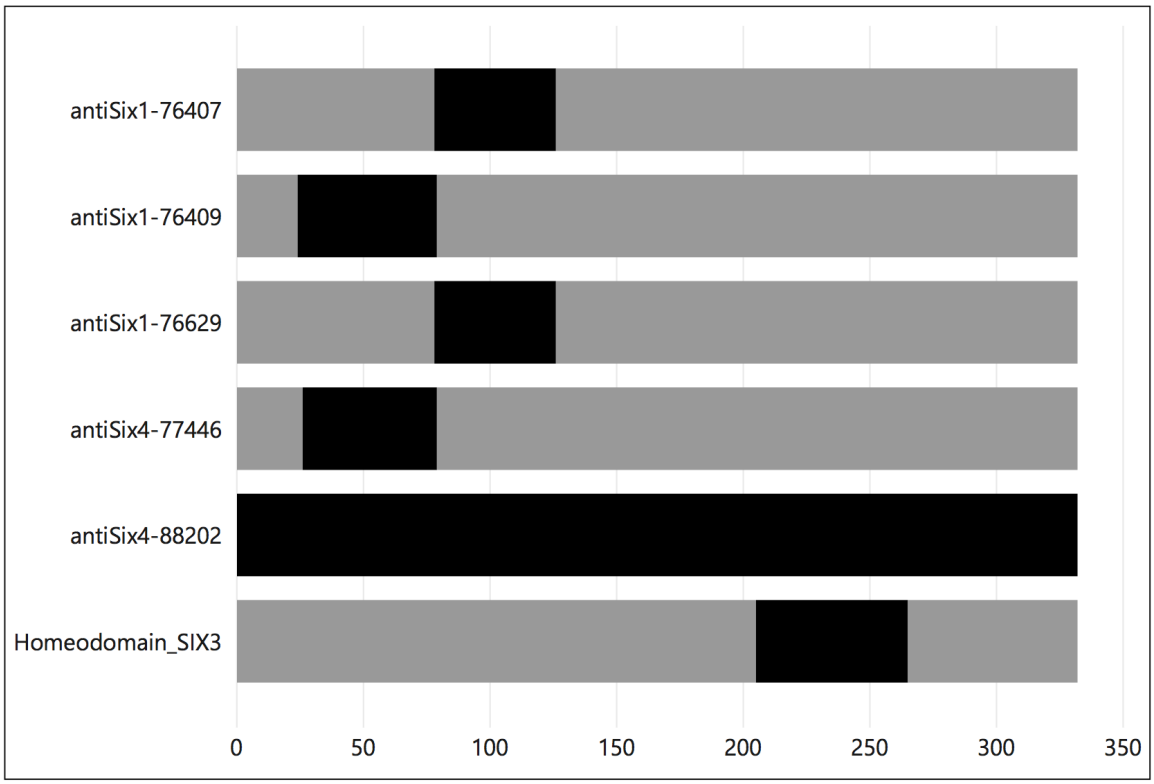


Figure 7.15: Predicted interaction sites between antiSix1 and antiSix4 sequences against SIX3. For each antiSix1 and antiSix4 sequence the bar represents the SIX3 protein sequence and the shadowed area represents the predicted interaction site for this protein sequence. The last bar represents the SIX3 protein sequence and the shadowed area shows where the homeodomain region is.

To obtain a better prediction of the interaction site with the target, the more accurate or narrower interaction sites from the top 100 predicted interactions were analyzed for each new protein sequence. Most of the new protein sequences, such as antiSix1-76186 and antiSix1-76289, were predicted to bind the target (SIX1), from amino acids 1 to 284, which is the whole SIX1 sequence of amino acids. For these two new sequences, the PIPE score was high for the target and low for the non-target proteins, which was considered a reliable candidate sequence. It was therefore considered worth analyzing other predicted interaction sites (from off-targets) different from the whole length of the target proteins.

Analysis of Other Predicted Interaction Sites with different Off-targets

AntiSix1-76186 and antiSix1-76289 sequences had some degree of predicted interaction with the off-target *T-cell leukemia homeobox protein 2* (TLX2, O43763) at different binding sites. TLX2 is a homeodomain transcription factor. As shown in Table 7.1, antiSix1-76186 was predicted to bind TLX2 from amino acids 81-107, while antiSix1-76289 was predicted to bind the same protein from amino acids 185-223 (Table 7.2). The homeodomain of TLX2 is at amino acids 157-216 and the new protein sequence antiSix1-76186 seemed to bind outside the homeodomain. This suggests that this new protein would have a higher likelihood to bind the target outside the homeodomain part. Moreover, for the same two antiSix1 sequences, the same phenomena was identified with another off-target protein, *Homeobox protein CDX-2* (CDX-2, Q99626) and its homeodomain from amino acids 186-245. AntiSix1-76186 was predicted to bind CDX-2 outside the homeodomain which was from amino acids 251 to 276 (Table 7.1). Similar to the previous analysis with the TLX2 protein, antiSix1-76289 was predicted to bind in the homeodomain area of CDX-2 (Table 7.2) from amino acids from 213-252.

Protein A	Protein B	Site1B	Homeodomain
antiSix1-76186	TLX2	81-107	No
antiSix1-76186	CDX-2	251-276	No

Table 7.1: Predicted interaction sites for antiSix1-76186 against off-target proteins. Protein A represents the new protein sequence and Protein B the off-target; Site1B indicates the predicted interaction site on Protein B; the last column indicates whether Site1B is in the homeodomain region on Protein B.

Protein A	Protein B	Site1B	Homeodomain
antiSix1-76289	TLX2	185-223	Yes
antiSix1-76289	CDX-2	213-252	Yes

Table 7.2: Predicted interaction sites for antiSix1-76289 against off-target proteins. Protein A represents the new protein sequence and Protein B the off-target; Site1B indicates the predicted interaction site on Protein B; the last column indicates whether Site1B is in the homeodomain region on Protein B.

The PIPE score of the new sequences with TLX2 and CDX-2 was very low. However, it was considered worth analyzing the interaction sites for these proteins, to have a better idea of the binding site for the new protein sequences and the target. Based on this analysis, for these two specific antiSix1 sequences, only antiSix1-76186 seems to have some degree of interaction with TLX2 and CDX-2 in a region outside the homeodomain. This might be an indication that, for this new protein sequence, the interaction with the target SIX1 would occur outside the homeodomain as well and hence have a higher likelihood of disrupting the SIX-EYA interaction.

To sum up this part of the analysis, if the new protein sequence was predicted to bind the homeodomain of the target or those of the off-targets, it suggested that the new protein would be less likely to disrupt the SIX-EYA interaction. As was explained earlier for these two proteins, the majority of the top 100 predicted interactions for the new protein sequences designed to interact with SIX1 were proteins that have a homeodomain. This would suggest that the identified peptides would bind to the DNA-binding region of SIX1, also known as the homeodomain. Finally, any off-target interaction of antiSIX1 protein sequences towards homeodomain proteins would suggest that the binding site would be on the homeodomain of SIX1 which,

while potentially interesting, was not likely to be useful for the purpose of disrupting the SIX-EYA interaction.

7.5.2 Analysis for EYA Binding Domain on SIX1 as Target

For new protein sequences designed to target SIX1, the predicted interaction site was the whole SIX1 sequence. Similarly, for proteins generated to target specific sites — the EYA binding domain — on SIX1 (detailed in Section 6.8.1), the predicted interaction site was the whole subsequence of SIX1. Also, while evaluating these new sequences against the whole SIX1 sequence, the predicted interaction site was again the whole sequence of SIX1. In regards to the interaction with SIX2, a few of the new protein sequences were predicted to interact with SIX2 on a site after its homeodomain. The majority of the new sequences were predicted to interact with SIX2 in a region covering almost all the SIX2 sequence, including its homeodomain. For SIX3, the majority of the new protein sequences were predicted to interact on a region before the SIX3 homeodomain, and some others on a region that included its homeodomain. For SIX4, most of the predicted interaction sites were on a region including its homeodomain, and just a few predicted sites were before or after the homeodomain. For SIX5, the majority of the predicted interaction sites were before the SIX5 homeodomain, with just a few new protein sequences predicted to interact after its homeodomain. For SIX6, the majority of the predicted interaction sites with the new protein sequences were before the SIX6 homeodomain, and only some of them were predicted to interact on a region that included its homeodomain.

From the new protein sequences described in Section 6.8.1, only antiSix1_9-118-88803 was predicted to interact with the target SIX1 at a site other than the whole SIX1 sequence, and the region was after the homeodomain on SIX1 (Figure 7.16). Similarly, the predicted interaction site for SIX6 was located in a region after the SIX6 homeodomain. For SIX2 and SIX3, the predicted interaction site was the whole protein sequence. For SIX4 the predicted interaction site started on a region before the SIX4 homeodomain, but it also included the homeodomain region. For SIX5, the predicted interaction site was located before the homeodomain region.

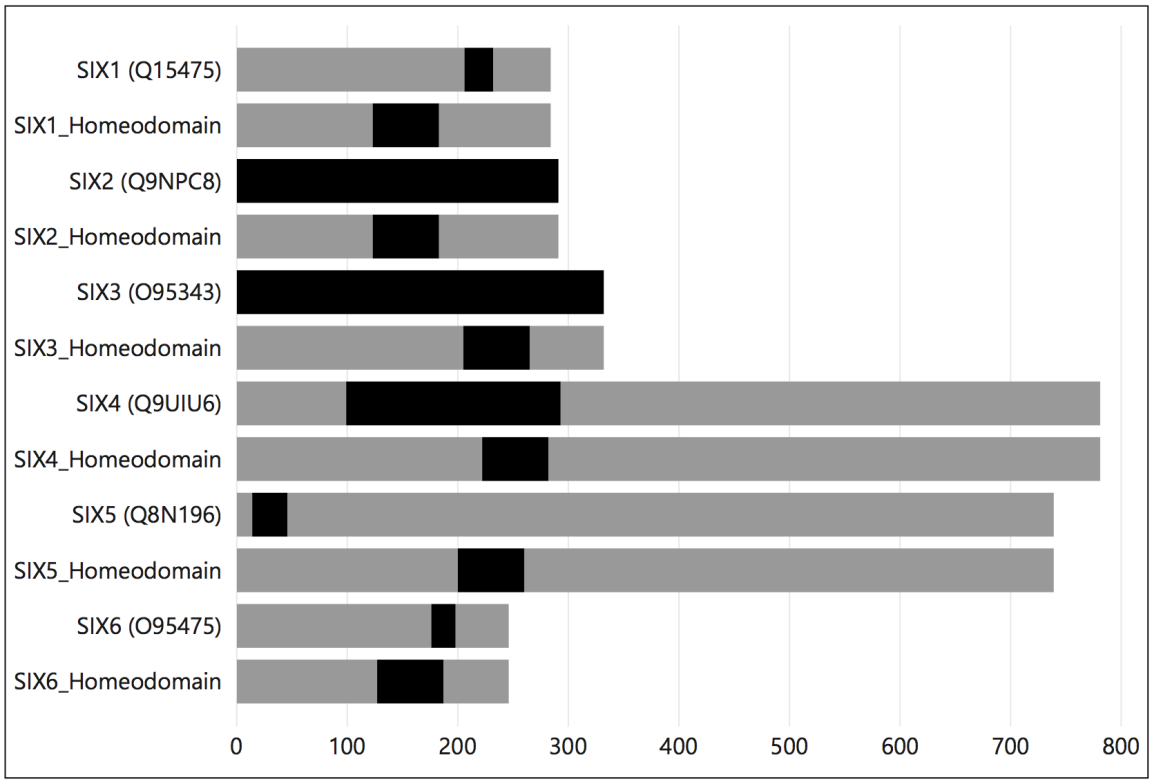


Figure 7.16: Predicted interaction sites for antiSix1_9-118-88803 against SIX proteins. Each SIX protein sequence is presented two times: first, the protein sequence — e.g., SIX1 (Q15475) — where the shadowed area represents the predicted interaction site with antiSix1_9-118-88803; second, the protein sequence — e.g., SIX1_Homeodomain — where the shadowed area represents the homeodomain region of the SIX protein.

In general, only imprecise predictions of interaction sites for the new protein sequences designed to bind a specific region on SIX1 were obtained. Therefore, similarly to what was done for the previously mentioned antiSix1 sequences, the predicted interaction sites were evaluated against the whole SIX family proteins.

Additionally, the predicted interaction sites for several specific antiSix1 sequences, aimed to target different sites of SIX1, against the off-target SIX6, are displayed in Figure 7.17. It can be observed that for three of the new protein sequences, the predicted interaction site for SIX6 was on a region before its homeodomain. For two other new protein sequences, the predicted interaction site was on a region before the SIX6 homeodomain but it also included the homeodomain part. Lastly, there was another new protein sequence, predicted to bind SIX6 on a region close to its homeodomain, but more on the right side, or c-terminus region.

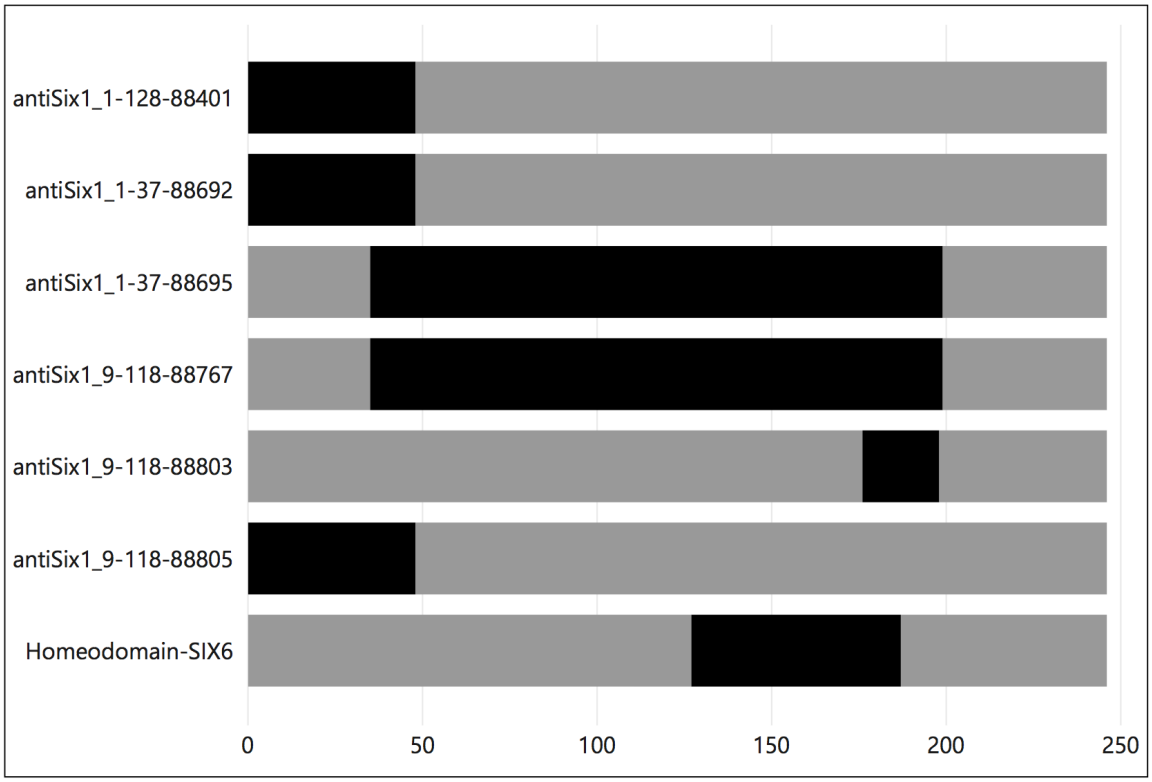


Figure 7.17: Predicted interaction sites between specific antiSix1 sequences against SIX6. For each specific antiSix1 sequence — designed to target specific regions on SIX1 — the bar represents the SIX6 protein sequence and the shadowed area represents the predicted interaction site for this protein sequence. The last bar represents the SIX6 protein sequence and the shadowed area shows where the homeodomain region is.

Chapter 8

Wet-Lab Experimental Validation

8.1 Introduction

As described in Chapter 3, computational PPI prediction tools are capable of processing a large amount of data, reducing considerably the required time, effort and costs involved in biochemistry experiments. Nevertheless, the produced data may not be enough to draw conclusions that might determine a full picture of a biochemistry research project. Rather, computational tools greatly simplify the experimental validation phase. Therefore, in bioinformatics research work, the wet-lab experimental validation is an important phase to corroborate the degree of confidence of the computationally generated data. Then, an accurate solution or approach to the initial problem that motivated the research can be designed. However, this type of validation might be a complicated process, in terms of time, required expertise, and the high costs involved in the whole process.

Regarding the wet-lab validation process, it might take several weeks or even months to design, prepare and complete a single biochemical assay, depending on the complexity and unforeseen problems that may arise. This often involves synthesizing specific biological molecules, ordering particular supplies, and performing genetic cloning prior to beginning an experiment. Also, the preparation of the materials and the actual execution of the experiments usually require a considerable amount of time. Furthermore, sometimes the outcome is not the desired one and it is necessary to try a different approach, or even to start the experiments over again, which might considerably increase the time required.

Secondly, a high level of expertise is always needed to perform these types of experiments. More often than not, the motivation for the research work is the search

for an innovative solution that requires a high level of background knowledge in the field. Also, for this phase it is mandatory to have the ability to: (1) validate the data resulting from computational tools, (2) perform different experimental techniques, and (3) design workaround solutions in case of eventual problems that require troubleshooting.

Thirdly, the more complicated the experiment, the higher the monetary cost. The total budget must cover the ordering of materials needed to do the actual experiment, and the remuneration of experts in the field and possible assistants. Even the possibility of having to pay for another round of experiments in case of not achieving the expected results, must be considered in the budget.

Dr. Alexandre Blais designed and led the experimental validation phase of the study. The information in this chapter is outside the scope of the computational study. However, in order to give a complete picture of the interdisciplinary work and the results obtained, this section summarizes how the data produced by InSiPS was used following the comprehensive analysis.

After the generation of data with InSiPS and data analysis to identify which new sequences would be the best candidates for actual experiments, the subsequent phase was logically experimental validation. What follows will describe the selected data, the experimental process and the biochemistry results. Finally, a summary of how this could be used for further research on this specific problem that aims to help in the approach to a treatment for muscular dystrophy will be presented.

Please note that for reasons of simplicity, in this chapter, the synthetic protein sequences generated by InSiPS will be also referred to as *peptides* (short chain of amino acids). The background information necessary for the understanding of the experimental details described in this section, is given in Chapter 2.

8.2 Selection of Synthetic Protein Sequences

In preparation for the actual experimental phase, the highest ranked synthetic protein sequences were carefully selected, based on several factors that will be outlined below. As stated in Section 4.3.1, the InSiPS fitness value defines in a general way the likelihood for a new protein sequence to bind to a specific target and not to bind to a list of non-targets. Planning for the actual experiments also required the analysis of the other scores produced by InSiPS (target, maximum non-target, and average

non-target scores) from which the fitness value is calculated.

The data analysis of results was described in Chapter 7. The sequence homology analysis helped clarify how similar new protein sequences were to other known binding proteins of the target (e.g. SIX1-binding proteins). How the new protein sequence was predicted to interact with off-target proteins, by means of reviewing the maximum non-target and the average non-target scores, and analyzing the predicted top 100 off-target proteins was also investigated. Also, the predicted interaction sites for the target, off-target, and target family proteins helped assess the likelihood that the new proteins would bind to the interaction site of interest.

Another consideration for selecting the best candidates (peptides) for the experiments was the analysis of the biochemical properties of the amino acids. Peptides with too many extreme or unnatural properties, for example, peptides including the amino acids *cysteines* or *prolines*, were avoided. The reason is that pairs of cysteine residues are able to react covalently to form disulfide bridges, which alters the 3D structure of the peptide. This increases the possibility of an unfavorable folding that would not allow a proper interaction with the protein of interest. Also, prolines are different from the other 19 amino acids because of their ring structure, which includes the n-terminus and the alpha-carbon of the amino acid. The ring in prolines can cause kinks in the peptide chain, which can destabilize the structure. Furthermore, peptides with a low level of hydrophobicity were selected because these would be able to dissolve in an aqueous solution. Lastly, moderately charged peptides, with more neutral amino acids, were preferred over other peptides.

The peptides that best met the described criteria are listed in Table 8.1. As can be observed, 14 peptides are of length 35; while four peptides are of length 25. As mentioned in Section 6.1, the ideal size for a designed peptide would be 25 or fewer amino acids, but 20 is the window size that InSiPS uses to evaluate similarity of fragments on sequences. Therefore, for a higher degree of confidence, the length considered in the computational experiments was 35. Thus, more peptides of length 35 were produced with the intention of increasing the number of fragments to compare in protein pairs. It should be noted that one of the peptides, ScrantiSix1_1-128-88401, is a scrambled version of the antiSix1_1-128-88401 protein with the same biophysical properties. In other words, the amino acids of this new protein sequence were put in a different order, with the objective of using this as negative control to increase the reliability of the results. Hence, ScrantiSix1_1-128-88401 would show no effect when

testing the interaction between the SIX1 protein and antiSix1 new proteins.

Name	Short Name	Sequence	Length
antiEya2-76189	aE2-76189	YHDWLNGLATLYVVMFCTELCPKWE	25
antiEya2-76843	aE2-76843	HVNSYFALITTDVWNWLRDRKNNEEHCSEAESQSD	35
antiEya2-77185	aE2-77185	NDLYQCCSNMWTPFHLYQRTAYYPH	25
antiEya2-77188	aE2-77188	FIVVGSTTRFSEFIYDVQTESMIKDDYWHCADKSV	35
antiEya2-88203	aE2-88203	QGPKDNKVLIWHHGSLGRPEWFLWTWPSFEIKRCQ	35
antiEya2-88737	aE2-88737	FSEGYTTVFLNAHAFLRGPRKSVHPIFKFPENEVD	35
antiSix1_1-128-88401	aS1_1-128-88401	QIDFIYCQLVWKNQPEQDLSVYSYANERWHIEDGY	35
antiSix1_1-37-88692	aS1_1-37-88692	QDNKVDERQFHSLIEWPWIRLYYHIKYIWKNQKQA	35
antiSix1_1-37-88695	aS1_1-37-88695	QQMDDSTWANFGKLVRDSWNFEGPINAMFTQWDHV	35
antiSix1_9-118-88767	aS1_9-118-88767	VVHSHKGTWHPNQYFIGHINYKTFMTRIFNIFCRQ	35
antiSix1_9-118-88803	aS1_9-118-88803	LKLICGVVIVKQTQTSHTTEMQGLCDVHAFKNGFK	35
antiSix1_9-118-88805	aS1_9-118-88805	KRNHAYTIPQSRKEQHFNRIVSPWKHGYQVHNTAW	35
antiSix1-76407	aS1_76407	WFDRDRFRSWMKKYFHCKPLGHCSY	25
antiSix1-76409	aS1_76409	RTNGNKHASVKLDDSRMHFAQSKFI	25
antiSix1-76629	aS1_76629	ERINGACSSPDKFDMDFAKVWCEWGKKYAYTSTEM	35
antiSix4-77446	aS4-77446	PVHESLGWWDNDERDHAWAQNGFPQYQCSFDTNVR	35
antiSix4-88202	aS4-88202	VVEHSQEEFPNVNGWFCNEWGFINDLAIKWALYRK	35
ScrantiSix1_1-128-88401	ScrS1-88401	SWDYEYNQVQLYSIQIHYDNRCWQFPKGEDAVLIE	35

Table 8.1: List of the peptides selected for experimental validation.

8.3 Experimental Validation Process

For the experimental validation, due both to time and budget constraints, only ten of the 18 anti-target peptides in Table 8.1 were cloned into expression vectors and expressed in mammalian cells. This included the scrambled sequence of amino acids in one of the peptides to be used as a negative control. This section describes how the *pull-down* (also known as affinity purification) experiments were carried out, using the *western blot* technique for the detection of proteins through their attached antibodies. Such experiments purify a target protein, in this case SIX1, and other proteins that are interacting with the target. A successful *pull-down* for our purposes would co-purify SIX1 with an antiSix1 protein, indicating that they are likely physically interacting in the cell. Besides the peptide with the scrambled sequence, a green fluorescent protein (GFP) was also used as a negative control, expected not to interact with the target. Finally, along with nine of the peptides, EYA3 was also used as a positive control, expected to interact with SIX1.

8.3.1 Preparation for the Experimental Procedure

After identifying the peptides to test in human cells, the next step was to clone them into a plasmid vector, which is a small DNA molecule used for protein expression. Because none of the synthetic proteins are found in nature, the new sequences had to be artificially synthesized. The first step in preparation for the experimental procedure was the ordering of reagents (chemical substance to produce a reaction) to clone the expression vectors to produce the peptides previously described in this chapter. The strategy was to obtain a single, large piece of DNA that coded for all the peptides, with the possibility to produce all the peptides as a single chain. Hence, IDT® synthesized the single long piece of DNA that included the sequences for the synthetic peptides to be tested in the lab. Additionally, primers, which are short strands of DNA, were designed to be used as a starting point for DNA synthesis, to amplify the expression of each synthetic protein. These primers were also synthesized by IDT.

Expressing the Peptides

The peptides were cloned in a mammalian cell expression vector and expressed with the VAP tag so that they could be purified and detected in our mammalian PPI study. Therefore, the approach was co-expressing in human cells VAP-tagged peptides along with myc-tagged SIX1, and observing if *pulling down* the peptides with VAP affinity would co-purify the SIX1 protein. The myc tag and Flag tag are antigens, which could be detected by antibodies and visualized using a *western blot*. After optimizing experimental conditions and after it was possible to clone expression plasmids for ten of the peptides, the *pull-down* by *western blot* experiment was performed.

8.4 Analysis and Results

When the ten peptides described in the previous section were expressed, five were detectable only in the *pull-down* eluates, meaning that they were poorly expressed and hence not sufficiently measurable to continue with the experiments. The other four peptides expressed to levels adequate to be visible in the input samples.

Of the four peptides that expressed to levels adequate for detection, AntiSix1_1-37-88692 and AntiSix1_9-118-88803 co-purified with the intended target SIX1, indicating an interaction. It is noteworthy that the two peptides that co-purified with SIX1 were

the ones designed to target the n-terminus region on SIX1. As described in Section 6.8, the predicted interaction sites were imprecise for the target protein, but those peptides generated to target SIX1 protein from amino acid 1-37, and from amino acid 9-118, were the proteins that yielded the best experiment results.

In Table 8.2 are described the InSiPS scores and specificity for the two peptides that co-purified with SIX1. As can be observed, the fitness for aS1_1-37-88692 (short name for AntiSix1_1-37-88692) is 0.83346, while for aS1_9-118-88803 (short name for AntiSix1_9-118-88803) it is 0.664608. The reason why the fitness is higher for one of the peptides is that it has a higher target score and a lower maximum non-target score. However, for both of the peptides, the specificity is above 99.97%. As explained in Section 6.7.4, a specificity above 99.95% represents a target score cutoff at which a positive interaction would be predicted with a high level of confidence for TN (true negatives) and thus an extremely small rate for FP (false positives). Hence, these results indicate that the predicted values can be reliable for experimental validation.

The described results provide strong initial evidence of binding between the peptides and SIX1, but more experimentation would be required to confirm a binary interaction and to elucidate the specific interaction site. Thus, these results indicate a high likelihood that the peptides generated with InSiPS could bind with the target they were designed to bind.

Short Name	Fitness	Target Score	Max non-target Score	Specificity
aS1_1-37-88692	0.833460	0.986111	0.154801	99.987%
aS1_9-118-88803	0.664608	0.813874	0.183402	99.974%

Table 8.2: InSiPS scores of peptides that co-purified with the target SIX1.

8.4.1 Results of the Pull-down Experiments by *Western Blot*

The results of the *pull-down* experiments by *western blot*, are displayed in Figure 8.1. Anti-myc (to detect SIX1) and anti-Flag (to detect the VAP-tagged proteins) were used as antibodies. Scr88401 is a scrambled version of antiSix1_1-128-88401, where the amino acid composition is preserved, but their order is randomly mixed; it was intended to serve as a negative control.

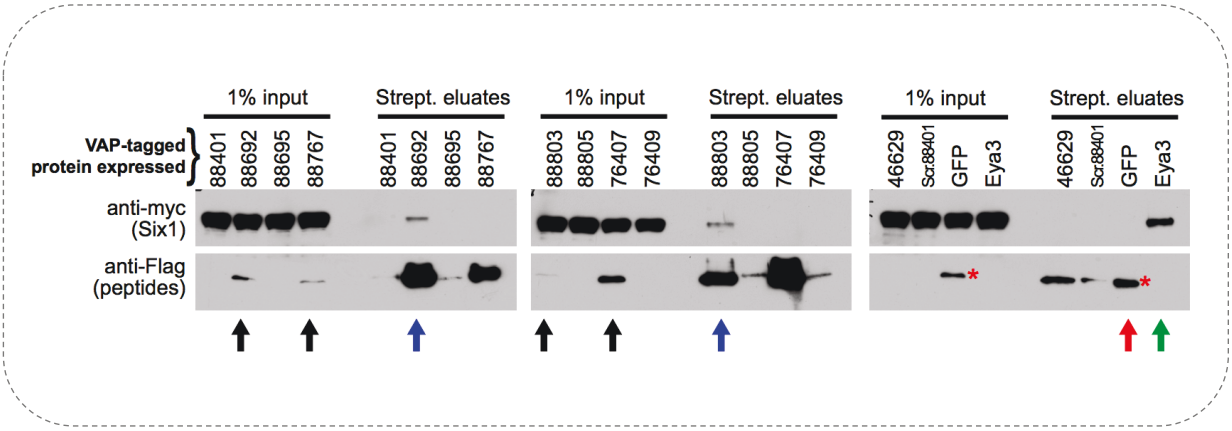


Figure 8.1: Results of the *pull-down* experiments, by *western blot* using anti-myc (to detect SIX1) or anti-Flag (to detect the VAP-tagged proteins) antibodies. Black arrows indicate peptides that expressed to sufficient levels so as to be detected by *western blotting*. Blue arrows indicate the *StrepTactin* eluates (to purify proteins) that contain not only the VAP-tagged peptides, but also the SIX1 protein, revealing an interaction. The red arrow indicates that the negative control sample (VAP-GFP) did not *pull-down* any SIX1 protein. The green arrow shows that the positive control worked as intended, as EYA3 is known to directly bind SIX1 [101] and hence they interacted. The red asterisks mark a band detected with the anti-Flag antibody, which most likely represents a breakdown — degradation — product of the VAP-tagged GFP protein.

8.4.2 Future Work

The experimental validation described in the previous section is strong evidence that two of the peptides interact with the protein they were designed to interact with. However, more work would be needed: (1) to formally prove that the interaction between the synthetic proteins and the target is direct; (2) to identify the actual interaction site; (3) to test whether the synthetic proteins disrupt the interaction

between SIX1 and EYA2. Also, once a synthetic protein is demonstrated to actually disrupt an interaction between SIX1 and EYA2 proteins, it would be necessary to observe the effect on the cell phenotype. That is, to observe how the cell is affected by this induced change.

Therefore, it is possible to continue with the experimental validation, focusing on the two peptides that were indicated to interact with SIX1. After formally proving an interaction and identifying the interaction site, which ideally would be on the n-terminus region of SIX1 (as mentioned in Section 2.2.3), the next experiment would be to *pull-down* SIX1 and EYA2. Then see if expressing those two peptides without the VAP tag will block the interaction.

The following is a brief description of future work to continue with the present research. The tasks are broken into two areas: tasks that will be performed on both the biochemistry and data science sides, and tasks that will be performed only on the biochemistry side.

Future Tasks for Biochemistry and Data Science

Before continuing with further experiments on the biochemistry side, the plan is to work on an optimization phase, mainly with the two peptides that indicated an interaction with SIX1 in the previous experiments. For every change made to a peptide, PIPE will be used to verify that the predicted interaction between the peptide and the target still retains a high score; also, that the predicted interaction with the non-targets continues to be as low as possible.

- Shortening of Peptides. The length of the peptides that indicated an interaction is 35 amino acids. However, a shorter length would be more feasible for further experiments. The preferred size for a third-party company to produce the actual synthetic protein is very short (i.e., from 12 to 24 amino acids) and most of the common peptide drugs are made from small peptides which are shown to be more stable and long lasting. While a peptide drug of length of 35 amino acids would be possible, a shorter length would be preferable. Therefore a mechanism will be implemented to shorten the two peptides, amino acid by amino acid, and see how their predicted interaction scores change. The aim is to end up with at least two peptides shorter than 35 amino acids, but still with a high likelihood of interaction with the target, and a low likelihood of interaction with the non-targets. As explained in Section 4.2.2, the lower limit would be 20 amino acids,

because this is the value for the window size used by InSiPS (through PIPE) to evaluate the similarity of sequence fragments.

- Changing Amino Acid Properties. Once the shortest length of at least two peptides has been achieved, it is planned to evaluate the amino acids, and to manually change them if needed, trying to preserve the properties that were considered for the peptide selection. As mentioned in Section 8.2, the aim would be to avoid unnatural or extreme properties to keep a low level of hydrophobicity; also to prefer peptides with more neutral amino acids.

Future Tasks for Biochemistry

After the optimization phase has been done for at least the two peptides that indicated an interaction with the target, the steps below will be performed through laboratory experiments. This would be the final phase of experimental validation, and will be key to determining whether the synthetic proteins designed by InSiPS would be of help in growing large quantities of satellite cells. Therefore, avoiding premature differentiation by disrupting the SIX-EYA interaction, as explained in Section 2.2.3.

- Validation of the Interaction with the Target. Section 8.4.1 described the experimental validation for nine synthetic protein sequences designed to interact with SIX1. Further validation with *pull-down* experiments will be performed with the peptides after the optimization phase. The objective is to ensure that the optimized peptides still interact with the target.

- Characterizing the Interaction. This will be the formal validation for detection of protein-protein interactions (PPIs) between the optimized synthetic protein sequences and the intended target. The intention will be to detect the affinity and specificity to identify whether the PPI is direct, and also to determine the exact binding site on the target.

- Testing Disruption of the SIX-EYA Interaction. Once a peptide that binds the intended target is identified and formally validated, there will be testing to validate whether the peptide is able to disrupt the interaction between SIX1 and EYA2, by competitively binding the interaction motif of the target protein (i.e. SIX1).

Chapter 9

Conclusions

This research study focused on designing synthetic protein sequences to help in an ongoing study looking a treatment for Duchenne muscular dystrophy (DMD), as described in Chapter 5. Was it possible to design a synthetic protein sequence, able to help in the study of a potential treatment for DMD? More specifically, was it possible to design a synthetic protein, able to disrupt the PPI proposed to regulate the premature differentiation of muscle satellite cells? To answer this question, several computational experiments with InSiPS were carried out on the BGQ supercomputer. As described in Section 6.7, the generated new protein sequences had a high fitness value, up to 0.859335, indicating a high predicted interaction score with the target (0.995597) and a low maximum predicted interaction score with the non-target proteins (0.136865). Such a high target score (predicted interaction score with the target protein) indicated a specificity above 99.98% in the validation of results (described in Section 6.7.4). As mentioned in Section 4.2.5, when scanning the entire protein interaction network of organisms (proteome), it is crucial to maintain a high specificity (true negative rate) and a low false positive (FP) rate, even at a cost of lowering the sensitivity, also known as true positive (TP) rate. The low TP rate would imply predicting fewer true protein interactions, but also the low FP rate at 0.02% (meaning a specificity of 99.98%) would allow us to be very confident of the predicted true interactions. In this research study, the synthetic protein sequence should be predicted to interact with only one target protein and not to interact with the rest of the proteins expressed in the human proteome. Hence the importance of obtaining a high specificity value.

The output data of the generated new protein sequences was analyzed, as described in Chapter 7. The analysis showed that some new protein sequences had

homology with proteins known to bind to the target protein; this was an indicator that the synthetic protein would interact with the target protein as well. The predicted binding sites were analyzed as well to identify whether the interaction site was the one of interest for this investigation. As most of the predicted binding sites for the target protein were imprecise, additional investigation on other proteins from the same family was done; this study showed that some of the predicted interaction sites were in the region of interest, the n-terminus region on SIX1. This data analysis allowed us to determine which new protein sequences would be considered for the production of actual synthetic proteins for the wet-lab experimental validation.

Lastly, in Chapter 8 the process of experimental validation was described. Although these experiments were performed on the biochemistry side, the results and next steps were analyzed and defined as part of a collaborative effort between both biochemistry and data science research lines. Two out of ten synthetic proteins co-purified with the target, which was a sign of interaction. The results of the experiments helped to validate and support the theory that the synthetic proteins would interact with the target protein. However, more formal validation is needed to test the interaction of the synthetic proteins with the target protein. The next experiment will be to test whether the synthetic protein disrupts the protein interaction of interest for this research (the PPI between SIX1 and EYA2 proteins). Looking back at the above question as to whether it was possible to design a synthetic protein able to disrupt the PPI proposed to regulate the premature differentiation of satellite cells, and while it is early yet to give a definitive answer, the current results are promising. They are being considered for further research in looking toward developing a treatment for DMD. While this thesis was still being written, the future work steps mentioned in Section 8.4.2 had already been started. Our era is almost entirely dependent on technology, and data science research can make it possible to bridge the gap between technology and health care. Thus, human collaboration among different disciplines — along with algorithms running in parallel on large computational resources — provide a major opportunity to analyze, understand and manipulate large amounts of data, and in particular, biological data.

9.1 Use of InSiPS to Design Human Proteins

InSiPS was successfully used in the past to design inhibitors of protein interactions for *S. cerevisiae* (a species of yeast), but it had not been used with human proteins. The fact that the current results obtained throughout this collaborative work are being considered for further experimentation demonstrates that InSiPS is also useful for human protein data. In this research project, InSiPS provided a research path with a solution to overcome a challenge for medical science in the search for a treatment to combat a severe disease. As described in Chapter 4, the InSiPS algorithm is stochastic in nature, and the final result depends on the initial population for the genetic algorithm; that is, the initial set of protein sequences made of random amino acids. Therefore, multiple computational experiments were performed before reaching the wet-lab experimental phase. Different to other computational tools for protein design, InSiPS is purely sequence-based and hence does not require protein data that most of the time is not available or is very difficult to obtain. One important consideration in using InSiPS is that the success of a synthetic protein sequence relies on the quality of the input data, specifically on the list of known protein-protein interactions. In this research project, the human protein data was carefully selected and this greatly increased the confidence in the results. Another important consideration when using InSiPS is that it requires a deep examination of the data, in particular, the output data. As a high-performance computational tool, the massive amount of the output data demands a comprehensive analysis by experts from the involved research fields. Therefore, how good the ultimate data is, will be directly related to the quality of the input data, and strictly related to the analysis done in selecting the best protein sequences generated through different runs.

9.2 Importance of Interdisciplinary Research

Based on the experience obtained during this work, interdisciplinary research can be defined as the collaboration between two or more disciplines working on a joint effort to achieve a specific goal in a research project. This type of investigation is fundamental to most projects involving data science. First, the initial problem is proposed by experts in the field and, depending on the complexity of the problem, most of the time their active participation throughout the investigation process is required. Second, the massive amount of data requires specialists, capable of developing or implementing computational tools to process it. Therefore, a shared effort and commitment between the involved fields, by means of a well-founded statement of the problem and the implementation of a reliable methodology, provides a great opportunity to approach an optimal solution for the problem. The procedure followed during this collaborative research between the disciplines of biochemistry and data science is illustrated in Figure 9.1. Interdisciplinary collaboration was essential to this research, where the resultant data was ultimately validated through wet-lab experiments. This gave a broader picture of the whole research study, as the results were not just measured through a formal computational data validation but also through experimental validation; hence the importance of collaboration among the distinct disciplines.

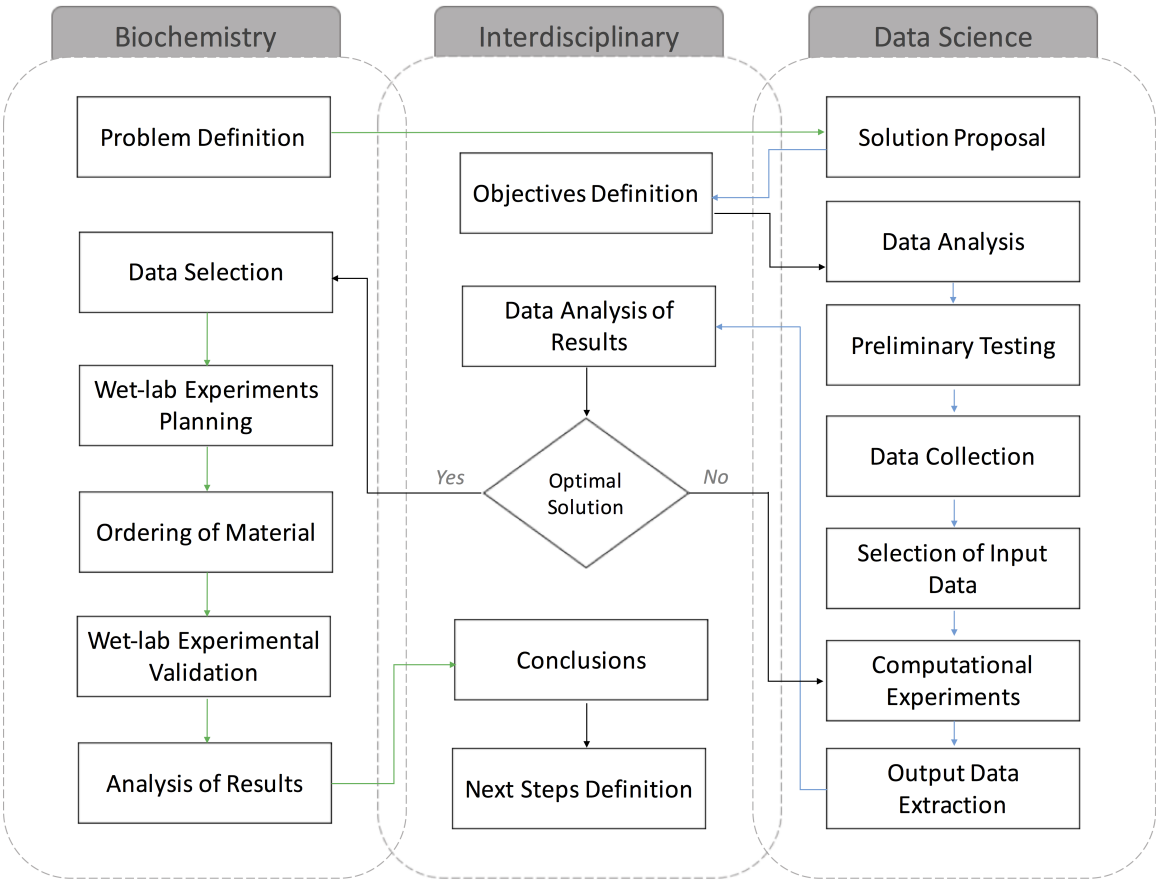


Figure 9.1: Outline of the procedure for interdisciplinary work. Although there was constant communication throughout the research, this illustration highlights in which area (i.e., biochemistry, data science or both) the actual steps of this research project were carried out.

9.3 Summary of Contributions

The main contributions of this thesis were outlined in Chapter 1: (1) two synthetic proteins indicated to interact with the target were experimentally validated and are now being considered for use as protein inhibitors in research to obtain a medical treatment for DMD; and (2) InSiPS had been used in the past to design inhibitors for *S. cerevisiae* [7], but this study represented the first time that InSiPS was used with human protein data. What follows is a more detailed list based on these two contributions that came out of the presented research work.

1. The highest ranked sequences were used for biochemistry experimental validation and research with the current results is ongoing, looking toward developing a treatment for DMD.

2. This research work demonstrates that InSiPS is also useful for human protein data.

3. More of the already generated synthetic protein sequences obtained a high predicted interaction score with the target protein and a low predicted interaction score with the non-target proteins. These protein sequences could also be used for wet-lab experiments and tested in the same way as the protein sequences already tested.

4. The protein interaction (SIX1-EYA2) of interest for this research has also been related to the development of tumors, especially in the spread of breast cancer, as mentioned in Section 2.2.3. If the synthetic proteins demonstrate an ability to disrupt this PPI, it would be of great interest also for other studies.

5. The work presented in this thesis, including the procedure followed for interdisciplinary research, may serve as a reference for other studies aiming to use InSiPS with human (or other species) protein data.

6. InSiPS had been used in the past with a list of 1,701 non-target (cytoplasmic) proteins. It was shown in the computational experiments (Section 6.6.3) that InSiPS is capable of generating synthetic protein sequences with a high fitness value, even when the list of non-target proteins is considerably larger, that is, using the complete set of proteins expressed in the human proteome.

7. It was shown in this thesis that short sequences of 25 and 35 amino acids obtained a high fitness value. Although this factor had not been considered before for InSiPS, a protein sequence short in length is an important issue for third-party companies that would produce the actual synthetic proteins. Therefore, the InSiPS results indicate that it is feasible to create protein sequences of a relatively short length.

9.4 Future Work

A summary of the future work to continue with the present research, highlighting the tasks to be performed in order to move forward from the final results, was given in Chapter 8. First, an optimization phase to shorten the synthetic protein sequences and change the amino acid properties would be performed on both the biochemistry and data science sides. Next, more formal validation to test whether the synthetic protein inhibits the PPI of interest would be done on the biochemistry side. The aim of this section is to include aspects to take into account for future work, to improve forthcoming interdisciplinary research projects; also, some functionality enhancements that could be added to InSiPS and would be of help for future research studies are mentioned.

For Consideration for Forthcoming Research Studies with InSiPS

The list of known PPIs is used to predict the interaction scores between the new protein sequences against target/non-target proteins. Section 6.2.2 described the process to collect the most recent human protein data used for the computational experiments, where two data sets termed *permissive* and *conservative* were described. For this research project, the selected data set was the *conservative* one, where each known PPI had to be confirmed by at least two independent research groups. Therefore, it is highly recommended that future research projects with InSiPS use such a strict process to select the list of known PPIs, trying to collect as much information on known protein interactions as possible.

Proposed Enhancements for InSiPS

InSiPS has now been used for two organisms (i.e., *S. cerevisiae* and *H. sapiens*) where the output results have been scientifically validated. This supports the idea that InSiPS could be used to design synthetic proteins for other research projects for the same or even different species. For this purpose, three enhancements that would facilitate the use of InSiPS for other studies are proposed below.

1. Adding functionality to specify as input data the interaction sites of the target protein. In the present work, the first round of experiments included the whole protein sequence of the target proteins. However, when analyzing the output data, it was found that the predicted interaction sites were different to the expected ones and this led to another round of experiments, including as target protein subsequences of the target protein. Due to the importance of the interaction site in a PPI, especially when designing a synthetic protein as an inhibitory protein, this enhancement would significantly complement this computational protein synthesizer.

2. Addressing the memory issue when including long protein sequences. In this work, two protein sequences of 22,152 and 34,350 amino acids were removed from the list of non-targets due to memory issues; thus, the length of the longest protein sequence for the computational experiments was of 8,797 amino acids. While this might not have affected the overall performance of the results, future studies might want to consider long (target/non-target) protein sequences such as the two that were removed in this work; hence, the importance of troubleshooting this issue.

3. Including functionality to generate synthetic protein sequences shorter than 25 amino acids. The future work for the wet-lab experimental phase in Section 8.4.2 included a task to shorten the already obtained protein sequences, as small peptides (i.e., from 12 to 24 amino acids) are more stable and durable. In Section 6.1 was discussed why synthetic protein sequences of length 25 amino acids, and more often of length 35, were generated. Although it is probably the goal most difficult to obtain, this feature would be a great asset for InSiPS.

Bibliography

[1] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan, X. Luo, and A. Golshani, "**PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs**," *BMC Bioinformatics*, vol. 7, p. 365, 2006.

[2] C. Long, J. R. McAnally, J. M. Shelton, A. A. Mireault, R. Bassel-Duby, and E. N. Olson, "**Prevention of muscular dystrophy in mice by CRISPR/Cas9-mediated editing of germline DNA**," *Science*, vol. 345, no. 6201, pp. 1184–8, 2014.

[3] L. Korngut, C. Campbell, M. Johnston, T. Benstead, A. Genge, A. Mackenzie, A. McCormick, D. Biggar, P. Bourque, H. Briemberg, *et al.*, "**The CNDR: collaborating to translate new therapies for Canadians**," *Canadian Journal of Neurological Sciences*, vol. 40, no. 5, pp. 698–704, 2013.

[4] Y. Wei, K. N. Speechley, G. Zou, and C. Campbell, "**Factors associated with health-related quality of life in children with Duchenne muscular dystrophy**," *Journal of Child Neurology*, vol. 31, no. 7, pp. 879–886, 2016.

[5] Y. Liu, A. Chu, I. Chakroun, U. Islam, and A. Blais, "**Cooperation between myogenic regulatory factors and SIX family transcription factors is important for myoblast differentiation**," *Nucleic Acids Research*, vol. 38, no. 20, pp. 6857–6871, 2010.

[6] Y. Liu, I. Chakroun, D. Yang, E. Horner, J. Liang, A. Aziz, A. Chu, Y. De Repentigny, F. J. Dilworth, R. Kothary, and A. Blais, "**Six1 Regulates MyoD expression in adult muscle progenitor cells**," *PLoS ONE*, vol. 8, no. 6, 2013.

[7] A. Schoenrock, D. Burnside, H. Moteshareie, A. Wong, A. Golshani, F. Dehne, and J. R. Green, "**Engineering inhibitory proteins with InSiPS: the In-Silico Protein Synthesizer**," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, p. 25, ACM, 2015.

[8] A. Schoenrock, F. Dehne, J. R. Green, A. Golshani, and S. Pitre, "**MP-PIPE: a Massively Parallel Protein-Protein Interaction Prediction Engine**," *Proceedings of the International Conference on Supercomputing*, pp. 327–337, 2011.

[9] P.-S. Huang, S. E. Boyken, and D. Baker, "**The coming of age of de novo protein design**," *Nature*, vol. 537, no. 7620, pp. 320–7, 2016.

[10] B. Alberts, A. Johnson, J. Lewis, *et al.*, ***Molecular Biology of the Cell***. New York: Garland Science, 4th ed., 2002. Protein Function, page `https://www.ncbi.nlm.nih.gov/books/NBK26911/`.

[11] G. A. Khoury, J. Smadbeck, C. A. Kieslich, and C. A. Floudas, "**Protein folding and de novo protein design for biotechnological applications**," *Trends in Biotechnology*, vol. 32, no. 2, pp. 99–109, 2014.

[12] M. Suarez and A. Jaramillo, "**Challenges in the computational design of proteins**," *Journal of The Royal Society Interface*, vol. 6, no. Suppl. 4, pp. S477–S491, 2009.

[13] M. W. Gonzalez and M. G. Kann, "**Chapter 4: protein interactions and disease**," *PLoS Computational Biology*, vol. 8, no. 12, 2012.

[14] M. Zhou, Q. Li, and R. Wang, "**Current experimental methods for characterizing protein-protein interactions**," *ChemMedChem*, vol. 11, no. 8, pp. 738–756, 2016.

[15] T. Mahmood and P. C. Yang, "**Western blot: technique, theory, and trouble shooting**," *North American Journal of Medical Sciences*, vol. 4, no. 9, pp. 429–434, 2012.

[16] J. Cohen, "**Bioinformatics—an introduction for computer scientists**," *ACM Computing Surveys*, vol. 36, no. 2, pp. 122–158, 2004.

[17] C. Pichavant, A. Aartsma-Rus, P. R. Clemens, K. E. Davies, G. Dickson, S. Takeda, S. D. Wilton, J. A. Wolff, C. I. Wooddell, X. Xiao, and J. P. Tremblay, "**Current status of pharmaceutical and genetic therapeutic approaches to treat DMD**," *Molecular Therapy*, vol. 19, no. 5, pp. 830–840, 2011.

[18] E. M. Yiu and A. J. Kornberg, "**Duchenne muscular dystrophy**," *Journal of Paediatrics and Child Health*, vol. 51, no. 8, pp. 759–764, 2015.

[19] M. Buckingham and F. Relaix, "**The role of Pax genes in the development of tissues and organs: Pax3 and Pax7 regulate muscle progenitor cell functions**," *Annual Review of Cell and Developmental Biology*, vol. 23, no. 1, pp. 645–673, 2007.

[20] N. A. Dumont, Y. X. Wang, J. von Maltzahn, A. Pasut, C. F. Bentzinger, C. E. Brun, and M. A. Rudnicki, "**Dystrophin expression in muscle stem cells regulates their polarity and asymmetric division.**," *Nature Medicine*, vol. 21, no. 12, pp. 1455–1463, 2015.

[21] A. C. Keefe and G. Kardon, "**A new role for dystrophin in muscle stem cells**," *Nature Publishing Group*, vol. 21, no. 12, pp. 1391–1393, 2015.

[22] D. Montarras, "**Direct isolation of satellite cells for skeletal muscle regeneration**," *Science*, vol. 309, no. 5743, pp. 2064–2067, 2005.

[23] N. C. Chang and M. A. Rudnicki, "**Satellite cells: the architects of skeletal muscle**," *Current Topics in Developmental Biology*, vol. 107, pp. 161–181, 2014.

[24] K. Kawakami, S. Sato, H. Ozaki, and K. Ikeda, "**Six family genes - Structure and function as transcription factors and their roles in development**," *BioEssays*, vol. 22, no. 7, pp. 616–626, 2000.

[25] J. Kumar, "**The sine oculis homeobox (SIX) family of transcription factors as regulators of development and disease**," *Cellular and Molecular Life Sciences*, vol. 66, no. 4, pp. 565–583, 2009.

[26] H. Yajima, N. Motohashi, Y. Ono, S. Sato, K. Ikeda, S. Masuda, E. Yada, H. Kanesaki, Y. Miyagoe-Suzuki, S. Takeda, and K. Kawakami, "**Six family**

genes control the proliferation and differentiation of muscle satellite cells," *Experimental Cell Research*, vol. 316, no. 17, pp. 2932–2944, 2010.

[27] A. N. Patrick, J. H. Cabrera, A. L. Smith, X. S. Chen, H. L. Ford, and R. Zhao, "**Structure-function analyses of the human SIX1-EYA2 complex reveal insights into metastasis and BOR syndrome**," *Nature Structural & Molecular Biology*, vol. 20, no. 4, pp. 447–453, 2013.

[28] M. A. Blevins, C. G. Towers, A. N. Patrick, R. Zhao, and H. L. Ford, "**The SIX1-EYA transcriptional complex as a therapeutic target in cancer**," *Expert Opinion on Therapeutic Targets*, vol. 19, no. 2, pp. 213–225, 2015.

[29] S. Banerjee-Basu, E. S. Ferlanti, J. F. Ryan, and A. D. Baxevanis, "**The Homeodomain Resource: sequences, structures and genomic information**," *Nucleic Acids Research*, vol. 27, no. 1, pp. 336–337, 1999.

[30] R. T. Moreland, J. F. Ryan, C. Pan, and A. D. Baxevanis, "**The Homeodomain Resource: a comprehensive collection of sequence, structure, interaction, genomic and functional information on the homeodomain protein family**," *Database*, vol. 2009, pp. 1–8, 2009.

[31] The UniProt Consortium, "**UniProt: the universal protein knowledgebase**," *Nucleic Acids Research*, vol. 45, no. D1, pp. D158–D169, 2017.

[32] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, "**Detecting protein function and protein-protein interactions from genome sequences**," *Science*, vol. 285, no. 5428, pp. 751–753, 1999.

[33] A. Valencia and F. Pazos, "**Computational methods for the prediction of protein interactions**," *Current Opinion in Structural Biology*, vol. 12, no. 3, pp. 368–373, 2002.

[34] A. Ceol, A. Chatr-aryamontri, E. Santonico, R. Sacco, L. Castagnoli, and G. Cesareni, "**DOMINO: a database of domain–peptide interactions**," *Nucleic Acids Research*, vol. 35, no. suppl_1, pp. D557–D560, 2006.

[35] A. Chatr-aryamontri, R. Oughtred, L. Boucher, J. Rust, C. Chang, N. K. Kolas, L. O'Donnell, S. Oster, C. Theesfeld, A. Sellam, C. Stark, B.-J. Breitkreutz,

K. Dolinski, and M. Tyers, "**The BioGRID interaction database: 2017 update**," *Nucleic Acids Research*, vol. 45, no. D1, pp. D369–D379, 2017.

[36] E. L. Huttlin, L. Ting, R. J. Bruckner, F. Gebreab, M. P. Gygi, J. Szpyt, S. Tam, G. Zarraga, G. Colby, K. Baltier, *et al.*, "**The BioPlex network: a systematic exploration of the human interactome**," *Cell*, vol. 162, no. 2, pp. 425–440, 2015.

[37] I. Coluzza, "**Computational protein design: a review**," *Journal of Physics: Condensed Matter*, vol. 29, no. 14, p. 143001, 2017.

[38] S. Pitre, M. Alamgir, J. R. Green, M. Dumontier, F. Dehne, and A. Golshani, "**Computational methods for predicting protein-protein interactions**," *Advances in Biochemical Engineering/Biotechnology*, vol. 110, no. January, pp. 247–267, 2008.

[39] V. S. Rao, K. Srinivas, G. N. Sujini, and G. N. S. Kumar, "**Protein-protein interaction detection: methods and analysis**," *International Journal of Proteomics*, vol. 2014, no. ii, pp. 1–12, 2014.

[40] A. Vinayagam, J. Zirin, C. Roesel, Y. Hu, B. Yilmazel, A. A. Samsonova, R. A. Neumüller, S. E. Mohr, and N. Perrimon, "**Integrating protein-protein interaction networks with phenotypes reveals signs of interactions**," *Nature Methods*, vol. 11, no. 1, pp. 94–99, 2013.

[41] B. G. Pierce, K. Wiehe, H. Hwang, B. H. Kim, T. Vreven, and Z. Weng, "**ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers**," *Bioinformatics*, vol. 30, no. 12, pp. 1771–1773, 2014.

[42] K. Yugandhar and M. M. Gromiha, "**Protein-protein binding affinity prediction from amino acid sequence**," *Bioinformatics*, vol. 30, no. 24, pp. 3583–3589, 2014.

[43] M. Ohue, Y. Matsuzaki, N. Uchikoga, T. Ishida, and Y. Akiyama, "**MEGADOCK: an all-to-all protein-protein interaction prediction system using tertiary structure data**," *Protein & Peptide Letters*, vol. 21, no. 8, pp. 766–778, 2014.

[44] J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, and Z. Weng, "**Protein-protein docking benchmark 2.0: an update**," *Proteins: Structure, Function and Genetics*, vol. 60, no. 2, pp. 214–216, 2005.

[45] H. Hwang, T. Vreven, J. Janin, and Z. Weng, "**Protein-protein docking benchmark version 4.0**," *Proteins: Structure, Function and Bioinformatics*, vol. 78, no. 15, pp. 3111–3114, 2010.

[46] M. N. Wass, G. Fuentes, C. Pons, F. Pazos, and A. Valencia, "**Towards the prediction of protein interaction partners using physical docking**," *Molecular Systems Biology*, vol. 7, no. 1, p. 469, 2011.

[47] E. Petsalaki, A. Stark, E. García-Urdiales, and R. B. Russell, "**Accurate prediction of peptide binding sites on protein surfaces**," *PLoS Computational Biology*, vol. 5, no. 3, 2009.

[48] A. Baspinar, E. Cukuroglu, R. Nussinov, O. Keskin, and A. Gursoy, "**PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes**," *Nucleic Acids Research*, vol. 42, no. W1, pp. 285–289, 2014.

[49] Y. Murakami and K. Mizuguchi, "**Homology-based prediction of interactions between proteins using averaged one-dependence estimators**," *BMC Bioinformatics*, vol. 15, no. 1, p. 213, 2014.

[50] A. Birlutiu, F. D'Alché-Buc, and T. Heskes, "**A bayesian framework for combining protein and network topology information for predicting protein-protein interactions**," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 3, pp. 538–550, 2015.

[51] Q. C. Zhang, D. Petrey, L. Q. Lei Deng, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, *et al.*, "**Structure-based prediction of protein-protein interactions on a genome-wide scale**," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[52] B. Xu and J. Guan, "**From function to interaction: a new paradigm for accurately predicting protein complexes based on protein-to-protein interaction networks**," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 4, pp. 616–627, 2014.

[53] Y. Guo, L. Yu, Z. Wen, and M. Li, "**Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences**," *Nucleic Acids Research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[54] L. Hu and K. C. Chan, "**Discovering variable-length patterns in protein sequences for protein-protein interaction prediction**," *IEEE Transactions on Nanobioscience*, vol. 14, no. 4, pp. 409–416, 2015.

[55] S. Martin, D. Roe, and J. L. Faulon, "**Predicting protein-protein interactions using signature products**," *Bioinformatics*, vol. 21, no. 2, pp. 218–226, 2005.

[56] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "**Predicting protein-protein interactions based only on sequences information**," *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–41, 2007.

[57] C. Y. Yu, L. C. Chou, and D. T. Chang, "**Predicting protein-protein interactions in unbalanced data using the primary structure of proteins**," *BMC Bioinformatics*, vol. 11, p. 167, 2010.

[58] S.-W. Zhang, L.-Y. Hao, and T.-H. Zhang, "**Prediction of protein–protein interaction with pairwise kernel support vector machine**," *International Journal of Molecular Sciences*, vol. 15, no. 2, pp. 3220–3233, 2014.

[59] U. Ogmen, O. Keskin, A. S. Aytuna, R. Nussinov, and A. Gursoy, "**PRISM: protein interactions by structural matching**," *Nucleic Acids Research*, vol. 33, no. SUPPL. 2, pp. 331–336, 2005.

[60] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "**The protein data bank**," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[61] Z. H. You, K. C. Chan, and P. Hu, "**Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest**," *PLoS ONE*, vol. 10, no. 5, pp. 1–19, 2015.

[62] Z.-H. You, M. Zhou, X. Luo, and S. Li, "**Highly efficient framework for predicting interactions between proteins**," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 721–733, 2017.

[63] R. A. Chica, "**Protein engineering in the 21st century**," *Protein Science*, vol. 24, no. 4, pp. 431–433, 2015.

[64] S. Kang and J. G. Saven, "**Computational protein design: structure, function and combinatorial diversity**," *Current Opinion in Chemical Biology*, vol. 11, no. 3, pp. 329–334, 2007.

[65] R. J. Pantazes, M. J. Grisewood, and C. D. Maranas, "**Recent advances in computational protein design**," *Current Opinion in Structural Biology*, vol. 21, no. 4, pp. 467–472, 2011.

[66] A. Zanghellini, "**De novo computational enzyme design**," *Current Opinion in Biotechnology*, vol. 29, no. 1, pp. 132–138, 2014.

[67] L. Baltzer, H. Nilsson, and J. Nilsson, "**De novo design of proteins–what are the rules?**," *Chemical Reviews*, vol. 101, no. 10, pp. 3153–63, 2001.

[68] F. Yu, V. M. Cangelosi, M. L. Zastrow, M. Tegoni, J. S. Plegaria, A. G. Tebo, C. S. Mocny, L. Ruckthong, H. Qayyum, and V. L. Pecoraro, "**Protein design: toward functional metalloenzymes**," *Chemical Reviews*, vol. 114, no. 7, pp. 3495–3578, 2014.

[69] I. Samish, C. M. MacDermaid, J. M. Perez-Aguilar, and J. G. Saven, "**Theoretical and computational protein design**," *Annual Review of Physical Chemistry*, vol. 62, no. 1, pp. 129–149, 2011.

[70] J. W. Bryson, S. F. Betz, H. S. Lu, D. J. Suich, H. X. Zhou, K. T. O'neil, and W. F. DeGrado, "**Protein design: a hierarchic approach**," *Science*, vol. 270, no. 5238, p. 935, 1995.

[71] B. O. Villoutreix, M. A. Kuenemann, J. L. Poyet, H. Bruzzoni-Giovanelli, C. Labbé, D. Lagorce, O. Sperandio, and M. A. Miteva, "**Drug-like protein-protein interaction modulators: challenges and opportunities for drug discovery and chemical biology**," *Molecular Informatics*, vol. 33, no. 6-7, pp. 414–437, 2014.

[72] M. J. Root, "**Protein design of an HIV-1 entry inhibitor**," *Science*, vol. 291, no. 5505, pp. 884–888, 2001.

[73] H. Saito, T. Inoue, and K. Shiba, "**A synthetic approach for protein evolution and cell engineering**," *2006 IEEE International Symposium on MicroNanoMechanical and Human Science*, pp. 1–6, 2006.

[74] M. R. Arkin, Y. Tang, and J. A. Wells, "**Small-molecule inhibitors of protein-protein interactions: progressing toward the reality**," *Chemistry and Biology*, vol. 21, no. 9, pp. 1102–1114, 2014.

[75] P. Benjamin Stranges and B. Kuhlman, "**A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds**," *Protein Science*, vol. 22, no. 1, pp. 74–82, 2013.

[76] R. V. Devi, S. S. Sathya, and M. S. Coumar, "**Evolutionary algorithms for de novo drug design - A survey**," *Applied Soft Computing Journal*, vol. 27, pp. 543–552, 2015.

[77] A. Paladino, F. Marchetti, S. Rinaldi, and G. Colombo, "**Protein design: from computer models to artificial intelligence**," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, p. e1318, 2017.

[78] R. J. Pantazes and C. D. Maranas, "**OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding**," *Protein Engineering, Design and Selection*, vol. 23, no. 11, pp. 849–858, 2010.

[79] W. Guo, J. A. Wisniewski, and H. Ji, "**Hot spot-based design of small-molecule inhibitors for protein-protein interactions**," *Bioorganic and Medicinal Chemistry Letters*, vol. 24, no. 11, pp. 2546–2554, 2014.

[80] S. J. Fleishman, T. A. Whitehead, D. C. Ekiert, C. Dreyfus, J. E. Corn, E.-M. Strauch, I. A. Wilson, and D. Baker, "**Computational design of proteins targeting the conserved stem region of influenza hemagglutinin**," *Science*, vol. 332, no. 6031, pp. 816–821, 2011.

[81] E.-M. Strauch, S. J. Fleishman, and D. Baker, "**Computational design of a pH-sensitive IgG binding protein**," *Proceedings of the National Academy of Sciences*, vol. 111, no. 2, pp. 675–680, 2014.

[82] A. R. D. Voet, H. Noguchi, C. Addy, D. Simoncini, D. Terada, S. Unzai, S.-Y. Park, K. Y. J. Zhang, and J. R. H. Tame, "**Computational design of a self-assembling symmetrical β-propeller protein**," *Proceedings of the National Academy of Sciences*, vol. 111, no. 42, pp. 15102–15107, 2014.

[83] Y. Park, "**Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences**," *BMC Bioinformatics*, vol. 10, no. 1, p. 419, 2009.

[84] S. Pitre, C. North, M. Alamgir, M. Jessulat, A. Chan, X. Luo, J. R. Green, M. Dumontier, F. Dehne, and A. Golshani, "**Global investigation of protein-protein interactions in yeast Saccharomyces cerevisiae using re-occurring short polypeptide sequences**," *Nucleic Acids Research*, vol. 36, no. 13, pp. 4286–4294, 2008.

[85] A. Amos-Binks, C. Patulea, S. Pitre, A. Schoenrock, Y. Gui, J. R. Green, A. Golshani, and F. Dehne, "**Binding site prediction for protein-protein interactions and novel motif discovery using re-occurring polypeptide sequences**," *BMC Bioinformatics*, vol. 12, no. 1, p. 225, 2011.

[86] S. Pitre, M. Hooshyar, A. Schoenrock, B. Samanfar, M. Jessulat, J. R. Green, F. Dehne, and A. Golshani, "**Short co-occurring polypeptide regions can predict global protein interaction maps**," *Scientific Reports*, vol. 2, pp. 1–10, 2012.

[87] A. Schoenrock, B. Samanfar, S. Pitre, M. Hooshyar, K. Jin, C. A. Phillips, H. Wang, S. Phanse, K. Omidi, Y. Gui, M. Alamgir, A. Wong, F. Barrenäs, M. Babu, M. Benson, M. A. Langston, J. R. Green, F. Dehne, and A. Golshani, "**Efficient prediction of human protein-protein interactions at a global scale**," *BMC Bioinformatics*, vol. 15, no. 1, pp. 1–22, 2014.

[88] A. Schoenrock, "**Realizing the potential of protein-protein interaction prediction for studying single and evolutionarily similar organisms**

and engineering inhibitory proteins with InSiPS: the in-silico protein synthesizer," Doctoral dissertation, Carleton University, 2016.

[89] K. Dick, F. Dehne, A. Golshani, and J. R. Green, "**Positome: a method for improving protein-protein interaction quality and prediction accuracy**," in *Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–8, IEEE, 2017.

[90] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "**CD-HIT: accelerated for clustering the next-generation sequencing data**," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[91] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "**CD-HIT Suite: a web server for clustering and comparing biological sequences**," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[92] C.-X. Deng, "**BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution**," *Nucleic Acids Research*, vol. 34, no. 5, pp. 1416–1426, 2006.

[93] S. P. Jackson and J. Bartek, "**The DNA-damage response in human biology and disease**," *Nature*, vol. 461, no. 7267, p. 1071, 2009.

[94] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "**Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[95] E. L. Huttlin, R. J. Bruckner, J. A. Paulo, J. R. Cannon, L. Ting, K. Baltier, G. Colby, F. Gebreab, M. P. Gygi, H. Parzen, *et al.*, "**Architecture of the human interactome defines protein communities and disease networks**," *Nature*, vol. 545, no. 7655, pp. 505–509, 2017.

[96] J. Wang, K. Huo, L. Ma, L. Tang, D. Li, X. Huang, Y. Yuan, C. Li, W. Wang, W. Guan, *et al.*, "**Toward an understanding of the protein interaction network of the human liver**," *Molecular Systems Biology*, vol. 7, no. 1, pp. 536–536, 2014.

[97] T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, *et al.*, "**An atlas of combinatorial transcriptional regulation in mouse and man**," *Cell*, vol. 140, no. 5, pp. 744–752, 2010.

[98] X. Fan, L. F. Brass, M. Poncz, F. Spitz, P. Maire, and D. R. Manning, "**The α subunits of Gz and Gi interact with the eyes absent transcription cofactor Eya2, preventing its interaction with the six class of homeodomain-containing proteins**," *Journal of Biological Chemistry*, vol. 275, no. 41, pp. 32129–32134, 2000.

[99] J. Sun, Z. Karoulia, E. Y. Wong, M. Ahmed, K. Itoh, and P. X. Xu, "**The phosphatase-transcription activator EYA1 is targeted by anaphase-promoting complex/Cdh1 for degradation at M-to-G1 transition**," *Molecular and Cellular Biology*, vol. 33, no. 5, pp. 927–936, 2013.

[100] J.-F. Rual, K. Venkatesan, H. Tong, T. Hirozane-Kishikawa, *et al.*, "**Towards a proteome-scale map of the human protein-protein interaction network**," *Nature*, vol. 437, pp. 1173–1178, 2005.

[101] H. Zhang and E. Stavnezer, "**Ski regulates muscle terminal differentiation by transcriptional activation of Myog in a complex with Six1 and Eya3**," *Journal of Biological Chemistry*, vol. 284, no. 5, pp. 2867–2879, 2009.