

TEL AVIV UNIVERSITY

SACKLER SCHOOL OF MEDICINE

DEPARTMENT OF HUMAN GENETICS AND MOLECULAR MEDICINE

IMPROVED DOCKING SCHEMES UTILIZING
BIOLOGICAL INFORMATION

THESIS SUBMITTED FOR THE DEGREE "DOCTOR OF PHILOSOPHY"

BY

INBAL LANDSBERG (HALPERIN)

SUBMITTED TO THE SENATE OF TEL AVIV UNIVERSITY

FEBRUARY 2005

This work was carried out under the supervision of:

1. Prof. Ruth Nussinov,
Department Of Human Genetics and Molecular Medicine,
Sackler School of Medicine,
Tel-Aviv University.
2. Prof. Haim Wolfson,
School of Computer Science,
Sackler Faculty for Exact Sciences,
Tel-Aviv University.

Acknowledgements

I thank my thesis advisors, Prof. Ruth Nussinov and Prof. Haim Wolfson, for introducing me to the fascinating field of docking and for their guidance and patience. I thank all the members of the Structural Bioinformatics group at Tel Aviv University for creating a kind and sharing as well as a professional working environment. I thank my mom, my dad, my family and friends for their support during these years. I thank my husband, David, for his endless true love.

Table of Contents

Abstract	Page	5
Chapter 1: Introduction	Page	7
The importance of the field of research	Page	8
The current state of the Docking field	Page	9
The core conception of this thesis	Page	10
Putting the thesis conception into practice	Page	11
Learning by induction	Page	13
The relationship between the six articles composing this thesis	Page	14
Chapter 2: The collection of articles	Page	16
Principles of docking: An overview of search algorithms and a guide to scoring functions	Page	17
SiteLight: binding-site prediction using phage display libraries	Page	52
Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking.	Page	68
Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families.	Page	80
Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking.	Page	119
Approaching the CAPRI challenge with an efficient geometry based docking	Page	125
Chapter 3: Discussion	Page	132
The main results of this thesis	Page	133
The scientific contribution of this thesis	Page	138
Future directions	Page	141
List of Shortcuts and Terms	Page	144
References	Page	146

Abstract

Protein-protein interactions are the key to understanding and controlling all the major processes that govern life as well as diseases and death: metabolism, signal transduction, regulation, replication etc. The task of reconstructing all protein-protein complexes occurring in one organism is so enormous it cannot be addressed solely by a labor intensive experimental approach. Docking schemes, a computational approach for predicting a protein-protein complex based on the structures of its components, are efficient enough for this application. Nevertheless, the performance of docking schemes is still not satisfying.

The goal of this work is to improve the reliability of docking schemes. Current docking schemes are based on shape and chemical features such as: geometric complementarity, hydrogen bonds, atom-atom interactions, electrostatic interactions and solvation energy. Many different measurements were developed over the last two decades to describe these features. Diverse ways to combine these to an effective scoring function were attempted. Despite these continuous efforts docking schemes did not reach the desired reliability. The performance of docking schemes is by far more dependable when a-priori knowledge of the binding site location is utilized. Therefore, this work is directed by the conception that a fruitful approach to improve docking schemes is to guide them by additional information that distinguishes, to some extent, the binding site from the rest of the protein surface. To achieve this purpose three novel general methods for binding site prediction were developed. These methods are suitable for guiding docking. Each method is based on a different source of biological information: 1. phage

display libraries derived peptides 2. Structurally conserved couples at interfaces 3. Inter-molecular correlated mutations. For each of these methods an appropriate dataset was created to assess the method's potential to predict binding sites and to guide docking.

In addition to developing general methods for binding site prediction, 19 particular blind docking experiments targets were also guided. These experiments were conducted as part of the CAPRI (Critical Assessment of PRediction of Interactions) challenge. These specific case studies shade light on the diverse rules that direct protein interactions, and on the potential, limitations and requirements of geometric based docking schemes. Some of these observations later set the basis for the development of general methods for binding site prediction.

The main contribution of thesis is the development of three methods for binding site prediction. Two of them, SiteLight and the conserved pairs, are based on original, innovative ideas. SiteLight, is the first attempt to utilize a commonly used experimental method, the phage display libraries, to predict three dimensional binding sites based on peptide mimicry. The conserved pairs' method is the first attempt to exploit the coupling of conserved residues across interfaces. The innovations of the third method, the inter-molecular correlated mutations method, are: 1. fitting the correlated mutation measurement to protein-protein interactions using knowledge based complementarity matrices. 2. Examining, for the first time, the applicability of inter-molecular correlated mutations to binding site prediction.

Chapter 1

Introduction

The importance of the field of research: Docking

Protein-protein interactions are the core of all the basic processes of life as well as diseases and death: metabolism, signal transduction, regulation, replication etc. Revealing the network of protein-protein interactions as well as the fine details of each interaction is one of the most important tasks of bioinformatics in the post-genomic era. This task is a key step in the process of rational drug design and in finding new drug targets. Determining experimentally a three dimensional structure of a single protein, is still a labor intensive task. The task is even more demanding when a complex of more than one macromolecule is at stake. Current efforts of the structural genomic initiative are referred to covering the space of single proteins / domains structures ¹. The number of available single-macromolecules structures in public the database (more than 30,000) exceeds the number of available complex structures by at least a factor of ten ². Moreover, the number of missing complex structures is even higher if we take into account that each macromolecule is typically involved in more than one interaction. Generating such a large number of complexes is only feasible, at this time, using an automated computer-aided approach such as docking schemes. In conclusion, the growing number of individual protein structures in the databases and the relatively small number of solved complexes makes docking a highly important method.

The current state of the Docking field: Current docking schemes reliability is not sufficient for large-scale automatic complex generation

Docking has been a subject of research for more than two decades³. In recent years the interest in the field expanded dramatically. More and more research groups have been joining this field, and conferences are devoted to this subject. Nevertheless, this field is still in its beginning. It has not been modified yet from an academic theoretical technique to a practical commonly used technique.

Some docking schemes are able to rank correct solutions within the top hundred or even within the top ten places for some docking targets³. Nevertheless, for most complexes the highest ranked structures are still false positives, *i.e.*, solutions with a high rmsd (root mean square deviation) from the "true" complex, a high score, and a low rank. Reliability of current docking schemes was assessed in the CAPRI (Critical Assessment of PRediction of Interactions) challenge⁴. The CAPRI is designed to test protein docking algorithms in blind predictions of the structure of protein-protein complexes. In the first 5 rounds 2495 results were submitted for 19 protein targets by more than 30 research groups. Among these results only a small percentage yielded reasonable results: 2.3%, 3.2% and 6.0% of the submitted solutions scored high, medium and acceptable respectively. The majority of the results (88.5%) were incorrect. These results are probably an overestimation. Some of the techniques used in the CAPRI challenge are specific for particular case studies or involve manual intervention. The performance of automatic docking schemes on a large scale is expected to be even lower.

The results of the CAPRI challenge indicate the complexity and difficulty of the docking problem. Part of these difficulties originates from the insufficient knowledge about the rules that governs protein interactions. The relative contribution of different forces to protein interactions (hydrophobicity, hydrogen bonding, electrostatic interaction

etc) is still a matter of dispute. Moreover, the docking problem presents a double challenge: not only is it an unsolved biological-chemical riddle but it is also a computational-mathematical challenge. Describing a protein in the atomic level or describing its surface creates large complex systems. In order to efficiently cope with the number of ways to combine two of these systems in a three dimensional space, sophisticated algorithms and data structures are required.

The core conception of this thesis: Binding site prediction is a promising approach for improving docking schemes

A part of the research was devoted to focusing on the element of a docking scheme which has the most potential for improvement. For this purpose the docking field was thoroughly reviewed. This broad examination of docking methods and comparison of their performance led to the conclusion that the most fertile approach to improving docking schemes is to focus on the scoring function. In principle, progress in the performance of docking schemes can be achieved by improving each of the three main elements of a docking scheme: surface representation, search stage, scoring stage. The basic methodology for surface representation, Connolly's surface representation ⁵ (*i.e.* rolling a "water" probe ball over the Van der Waals surface, smoothing the surface) seems to be stable and perform well. Sophisticated modifications over this method improved docking schemes only slightly. Improvement of the second element (*i.e.* search stage) can be divided into two strategies: 1. A more efficient sampling of the search space 2. Reduction of the search space. The former improves mainly the running times and not reliability. Despite the importance of keeping docking schemes fast and efficient the

main problem with docking schemes still remains their low reliability. The later strategy (i.e. reduction of the search space) results in the reduction of the number of false positives. Consequently, "near native" solutions are expected to be ranked higher. Most of the potential for improvement appears to be in the third element, the scoring function. Current scoring functions are based on geometric complementarity, hydrogen bonds, atom-atom interactions, electrostatic interactions and solvation energy. Many measurements that describe these features were combined in diverse ways in order to develop an effective scoring function. Despite the continuous efforts docking schemes did not reach the desired results. Therefore, we have chosen the following approach to reduce the search space: guiding docking schemes based on binding site prediction methods. Binding site prediction methods should be general, i.e. applicable to any protein, since the long-term goal of docking schemes is large-scale automatic prediction of complex structures.

Putting the thesis conception into practice: Development of three new general methods for binding site prediction

At the heart of this thesis are three novel general methods for binding site prediction. Each of them is based on a different source of information: 1. Phage display libraries. 2. Structurally conserved residue pairs 3. Inter-molecular correlated mutations. The central principles of these methods are presented below.

Phage display libraries. *SiteLight*⁶, is a novel computational tool for binding site prediction using phage display libraries. A phage display library is a collection of random peptide sequences of a specified length (typically 6-15 residues). A target molecule of

interest scans such a library to yield binding peptides. Some of these peptides are expected to bind the target molecule at the binding site of a binding partner of interest. These peptides are expected to mimic the binding site of this binding partner. Mapping the peptides derived from the library scan onto the surface of the binding partner reveals this mimicry, thereby pointing to the binding site. An algorithm for efficient mapping of a peptide sequence onto a three-dimensional (3D) protein surface was developed. This is the first study that attempts to validate the applicability of phage display libraries for automated binding site prediction on 3D structures.

Structurally conserved residue pairs. Conservation is a commonly used method for binding site prediction ⁷. The expectation that binding will impose conservation constrains renders another assumption: Positions that were conserved for binding purposes should be coupled across interfaces. Though this assumption is a fundamental one it has never been inspected or used to enhance the prediction power of conservation methods. In this thesis this assumption is first tested. For this purpose, a statistical model was built, which found the pairing assumption to be truthful ⁸. In addition, improvement of docking schemes using the structurally conserved pairs was demonstrated.

Inter-molecular correlated mutations. The term *correlated mutations* refers to co-occurring mutations. Once a residue is changed, despite the functional constraints operating on it, this mutation can be compensated by an additional mutation of a complementary residue. Correlated mutations have been repeatedly exploited for intra-molecular, but not inter-molecular, contact map prediction ^{9; 10; 11}. This gap is due to several obstacles, such as the availability of 3D complexes, paralog discrimination and the availability of sequence pairs, that are required for inter- but not intra-molecular analyses. In this research fusion protein families, which bypass some of these obstacles, were chosen for analysis ¹². This dataset enabled for the first time to assess the performance of six previously suggested correlated mutation measurements. In addition,

a new method for measuring correlated mutations that was designed especially for protein-protein interfaces was developed. This method is based on residue complementarity. In addition, three modifications over existing methods were developed and their performance assessed.

Learning by induction: Taking part in particular docking experiments, as part of the CAPRI challenge, inspires the creation of new methods for binding site prediction

The CAPRI (Critical Assessment of PRediction of Interactions) challenge is designed to test protein docking algorithms in blind predictions of the structure of protein-protein complexes⁴. It encourages the development of docking schemes by sharing knowledge among researchers and establishing collaborations, as well as motivating researchers by competition and the presentation of challenging docking targets. The greater part of major research groups in the docking field from all over the world take part in this challenge. The Structural bioinformatics group in Tel-Aviv University, which I am part of, participated in all of the rounds of the CAPRI challenge that took place so far. In these rounds 19 blind docking experiments were conducted. After revealing the complexes determined by crystallography, the submitted solutions were assessed. My part in the CAPRI challenge included: 1. Establishing geometric constraints based on literature search. 2. Manual screening of top ranking solutions intended for selecting 10 for submission.

These particular case studies shade light on the diverse rules that direct protein interactions, and on the potential, limitations and requirements of geometric based docking schemes. Some of these observations later set the foundations for the

development of general methods for binding site prediction. For example, the Dockerin-Cohesin case study was the trigger to developing the correlated mutations method for inter-molecular residue pairs. The target of alpha-amylase with three of its Abs highlighted the need to predict the binding site of an antigen. Antigenicity prediction methods are not ripe for this purpose yet. Standard conservation methods are inappropriate since the antigen and the antibody do not co-evolve over a long time. This deficiency of methods for antigen binding site prediction intrigued the attempt to use artificial evolution sources. This alternative set the basis for *SiteLight*. Evaluation of the results of the first rounds of the CAPRI challenge emphasized the need for more stringent constrains. A correct, but not perfect, prediction of the binding site on each protein separately is not always enough to yield successful docking solutions. This observation led to the concept of creating pair constrains. Working with inter-molecular pairs instead of residues on a single side of binding site create much stringent constrains of the possible orientations of two proteins. Inter-molecular pairs are a key feature in two of the developed methods for binding site prediction: the structurally conserved residue pairs' method and the inter-molecular correlated mutations method. The utilization of inter-molecular pairs makes these methods especially suitable for improving docking schemes.

The relationship between the six articles composing this thesis

This work is composed of research projects that study protein docking from different points of view. These projects are presented in six articles ^{3; 6; 8; 12; 13; 14}. The foundations for the conception that guided this thesis were laid by the article, titled "Principles of docking: An overview of search algorithms and a guide to scoring

functions" ³. This article compares existing docking algorithms and reviews existing scoring schemes. It led to the conclusion that a fruitful approach to improve docking schemes is to guide them by additional biological information. The additional information should be able to distinguish, to some extent, the binding site from the rest of the protein surface. The approach set by the first article was later implemented by developing three novel methods for binding site prediction. These are titled "SiteLight: binding-site prediction using phage display libraries" ⁶, "Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking" ⁸ and "Inter-molecular contacts prediction with correlated mutations: a study on fusion proteins and the Cohesin-Dockerin families" ¹². Each method can be used as a separated "stand alone" technique undependable on the others. Since the methods are based on different sources of information they can also supplement and enhance each other. These can be used as additive filters in docking schemes. Some of the ideas for the developed binding site prediction methods originated from the CAPRI challenge (see section "Learning by induction"). The experiments that were part of the CAPRI are summarized in two articles titled "Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking" ¹³ and "Approaching the CAPRI challenge with efficient geometry based docking" ¹⁴.

Chapter 2

The Collection of Articles