Computational prediction and analysis of macromolecular interactions

https://hdl.handle.net/2144/17063

Boston University
Approved by

First Reader

Sandor Vajda, Ph.D.
Professor of Biomedical Engineering

Second Reader

Dmytro Kozakov, Ph.D.
Research Associate Professor of Biomedical Engineering
ACKNOWLEDGMENTS

I would like thank my wife Laurel and my parents, without whose love and support I would not have made it this far. This work would not have been possible without the other members of the Structural Biology Lab, past and present.
COMPUTATIONAL PREDICTION AND ANALYSIS OF
MACROMOLECULAR INTERACTIONS

SCOTT EDWARD MOTTARELLA
Boston University Graduate School of Arts and Sciences
and College of Engineering, 2016

Major Professor: Sandor Vajda, Professor of Biomedical Engineering

ABSTRACT

Protein interactions regulate gene expression, cell signaling, catalysis, and many other functions across all of molecular biology. We must understand them quantitatively, and experimental methods have provided the data that form the basis of our current understanding. They remain our most accurate tools. However, their low efficiency and high cost leave room for predictive, computational approaches that can provide faster and more detailed answers to biological problems. A rigid-body simulation can quickly and effectively calculate the predicted interaction energy between two molecular structures in proximity. The fast Fourier-transform-based mapping algorithm FTMap predicts small molecule binding 'hot spots' on a protein's surface and can provide likely orientations of specific ligands of interest that may occupy those hot spots. This process now allows unique ligands to be used by this algorithm while permitting additional small molecular cofactors to remain in their bound conformation. By keeping the cofactors bound, FTMap can reduce false positives where the algorithm identifies a true, but incorrect, ligand pocket where the known cofactor already binds. A related algorithm, ClusPro, can evaluate interaction energies for billions of docked conformations of macromolecular
structures. The work reported in this thesis can predict protein-polysaccharide interactions and the software now contains a publicly available feature for predicting protein-heparin interactions. In addition, a new approach for determining regions of predicted activity on a protein's surface allows prediction of a protein-protein interface. This new tool can also identify the interface in encounter complexes formed by the process of protein association—more closely resembling the biological nature of the interaction than the former, calculated, binary, bound and unbound states.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................... iv

ABSTRACT ........................................................................................................................ v

TABLE OF CONTENTS ..................................................................................................... vii

LIST OF TABLES ............................................................................................................. x

LIST OF FIGURES .......................................................................................................... xi

LIST OF ABBREVIATIONS .......................................................................................... xii

INTRODUCTION ............................................................................................................. 1

1.1 Determination of Molecular Interactions .................................................................. 2

1.2 Rigid Body Sampling in PIPER ................................................................................ 3

1.2.1 Scoring Function ............................................................................................... 5

1.2.2 Pairwise Potentials using DARS ........................................................................ 6

1.3 'Hot Spot' identification with Computational Solvent Mapping ............................... 7

1.4 Automated protein docking....................................................................................... 9

1.5 Contributions........................................................................................................... 10

EXTENDED PROTEIN MAPPING ................................................................................ 11

2.1 Introduction ............................................................................................................. 11

2.2 The Parameterization Server ................................................................................ 12

2.3 Case Study .............................................................................................................. 17

2.4 Conclusions........................................................................................................... 21
4.2.4 Quality measures.............................................................................................. 64

4.3 RESULTS ............................................................................................................... 65
  4.3.1 Encounter complex interaction site.............................................................. 65
  4.3.2 Prediction of the interaction site using the known partner protein .............. 68
  4.3.3 Binding site detection using a panel of proteins .......................................... 73

4.4 DISCUSSION ......................................................................................................... 77

BIBLIOGRAPHY ............................................................................................................. 85

CURRICULUM VITAE ................................................................................................... 96
LIST OF TABLES

Table 1. List of Heparin Bound Protein Structures and their Unbound Forms. ............... 32
Table 2. List of bound and unbound protein conformations used for docking prediction as well as rank and RMSD of best prediction and the rank of the binding site. ............ 37
Table 3. List of bound and unbound protein conformations used for docking prediction by Samsonov et al as well as rank and RMSD of best prediction and the RMSD of the top ranked pose. ........................................................................................................ 48
Table 4. Success rates at cutoff values 0.25 and 0.5 for both recall and precision........ 75
LIST OF FIGURES

Figure 1. The standard FTMap 16 probe library............................................................... 13

Figure 2. Mapping of an unbound structure of thrombin (PDB ID: 1HXF) using the small
molecule C2A from a ligand-bound thrombin structure (PDB ID: 2C8Z) as a user-
supplied additional probe.............................................................................................. 17

Figure 3. FTMap predicted hot spots on thrombin.......................................................... 19

Figure 4. FTMap predicted hot spots of thrombin with dimethyl sulfoxide cofactor
included.......................................................................................................................... 20

Figure 5. Thrombin binding site with FTMap predicted hot spots and bound full ligand
C7M (silver)................................................................................................................ 21

Figure 6. Best results and contact maps of the five test systems..................................... 40

Figure 7. Atom-atom contacts between heparin chains and protein residues................. 41

Figure 8. Results for Annexin V (PDB IDs 1G5N for the bound structure, 2IE7 for the
unbound)........................................................................................................................ 42

Figure 9. Results for human fibronectin (Unbound PDB 1FNH).................................. 44

FIGURE 10. Encounter complexes and predictions of the interaction site for the EIN/HPr
complex (Fawzi et al., 2010)...................................................................................... 67

FIGURE 11. Success rates using partner proteins for prediction...................................... 71

FIGURE 12. Predicting the binding sites of the nitrogenase molybdenum-iron protein.. 72

FIGURE 13. Success rates using a panel of 12 proteins for prediction............................ 74

Figure 14. Predicted binding sites of the nuclear transport factor 2 (NTF2, PDB ID 1oun,
chains A and B, forming a homodimer, shown as wheat surface)............................ 77
**LIST OF ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE</td>
<td>Analytic Continuum Electrostatic</td>
</tr>
<tr>
<td>ACF</td>
<td>Atom Contact Frequency</td>
</tr>
<tr>
<td>ACP</td>
<td>Atomic Contact Potential</td>
</tr>
<tr>
<td>AM1</td>
<td>Austin Model 1</td>
</tr>
<tr>
<td>AMBER</td>
<td>Assisted Model Building with Energy Refinement</td>
</tr>
<tr>
<td>BCC</td>
<td>Bond Charge Correction</td>
</tr>
<tr>
<td>C</td>
<td>Cutoff</td>
</tr>
<tr>
<td>CAPRI</td>
<td>Critical Assessment of Prediction of Interaction</td>
</tr>
<tr>
<td>CHARMM</td>
<td>Chemistry at Harvard Macromolecular Mechanics</td>
</tr>
<tr>
<td>COM</td>
<td>Center of Mass</td>
</tr>
<tr>
<td>CS</td>
<td>Consensus Site</td>
</tr>
<tr>
<td>DARS</td>
<td>Decoys as the Reference State</td>
</tr>
<tr>
<td>EI</td>
<td>Enzyme-Inhibitor</td>
</tr>
<tr>
<td>EIN</td>
<td>N-Terminal Domain of Enzyme I</td>
</tr>
<tr>
<td>FBDD</td>
<td>Fragment Based Drug Discovery</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>FGF</td>
<td>Fibroblast Growth Factor</td>
</tr>
<tr>
<td>FN</td>
<td>Fibronectin</td>
</tr>
<tr>
<td>FT</td>
<td>Fourier Transform</td>
</tr>
<tr>
<td>GAFF</td>
<td>General AMBER Force Field</td>
</tr>
<tr>
<td>GAG</td>
<td>Glycosaminoglycan</td>
</tr>
</tbody>
</table>
GAMESS ........................................... General Atomic and Molecular Electronic Structure System
HGF ........................................................................................................ Hepatocyte Growth Factor
HPr ................................................................. Histidine-Containing Phosphocarrier Protein
HS .............................................................. Heparan Sulfate
IC\textsubscript{50} ................................................. Half Maximal Inhibitory Concentration
IFT ................................................................. Inverse Fourier Transform
IRMSD .............................................................. Interface RMSD
LE .............................................................. Ligand Efficiency
MD .............................................................. Molecular Dynamics
MMFF ............................................................. Merck Molecular Force Field
MSCS .............................................................. Multiple Solvent Crystal Structures
NIP .............................................................. Normalized Interface Frequency
NMR .............................................................. Nuclear Magnetic Resonance
P .............................................................. Precision
PDB .............................................................. Protein Data Bank
PRE ...................................................... Paramagnetic Relaxation Enhancement
R .............................................................. Recall
RCF .............................................................. Residue Contact Frequency
RMSD .............................................................. Root Mean Squared Deviation
SF .............................................................. Scatter Factor
SMILES ....................................................... Simplified Molecular-Input Line-Entry System
CHAPTER ONE

INTRODUCTION

Proteins are the functional units of most cellular mechanisms. They function by interacting with various substrates on the atomic level to mediate biological changes. Proteins act in gene regulation, such as histones, polymerases, and transcription factors; signal transduction, such as cytokines, neurotransmitters, and G proteins; metabolism, such as transferases and ligases; as well as numerous other biological processes. A great effort has been invested into understanding their structures and their functions for over 60 years. Our understanding of biological processes has allowed us to accurately describe methods by which these proteins function and the faults when they do not. Understanding the dysfunction of proteins has been the first step to many significant medical advances.

Protein interactions occur between a protein and a substrate. That substrate can be any other molecule, such as a small molecule, a polymer, or another protein. The structural features of the proteins involved in each interaction vary substrate to substrate. Smaller substrates tend to bind in deeper cavities of the protein surface and consist of several specific areas that contribute a disproportionate amount to the binding free energy (Hajduk et al., 2005). Larger substrates like proteins cover a larger amount of surface, thus proteins typically have a variety of distributed features specific to one partner (Ezkurdia et al., 2009). Describing the features of these interactions is a large facet of molecular biology and several methods exist to build models of the resulting complexes.
1.1 Determination of Molecular Interactions

Two common methods for determining the interface residues of protein complexes are X-ray crystallography and mutagenesis. X-ray crystallography begins by forming a crystal out of a sample of a single complex. It then attempts to back-calculate the spatial coordinates of each atom from the diffraction of X-rays that pass through the crystal. The result is a three-dimensional model containing the approximate positions of the atoms in the complex, one of the most detailed examples available for describing protein interactions today. Mutagenesis functions by altering the sequence of a target protein, typically by mutating a single amino acid to alanine, to determine the role of each residue one by one in a particular interaction. With this method, residues that are essential for a particular interaction to occur can be identified and annotated for how much of a disruption the mutation incurred (Cunningham et al., 1989). While these methods, and others, provide valuable information for the description of protein interactions, none is effective enough or efficient enough to describe all complexes, leaving room for newer approaches to be developed.

Over the last 30 years, the increasing power of computational simulations has been applied to this problem. Algorithmic approaches that were either too time consuming or too mathematically involved have been developed to make use of the information determined by biochemical means and to fill the gaps left by them. Sequence-based algorithms search for patterns of residues involved in protein interactions. Molecular dynamics (MD) simulations calculate the motion of structures on the atomic scale as they come into proximity of each other. Rigid body sampling attempts
to calculate the interaction energy of two structures relative to each other without
calculating the intramolecular atomic motion of either structure. Sequence-based
algorithms are typically the fastest and MD simulations can be the most physically
accurate, but rigid body sampling strikes a balance between the two that provides
significantly more information than sequence-based algorithms in significantly less time
than MD simulations.

### 1.2 Rigid Body Sampling in PIPER

PIPER is a rigid body sampling algorithm that has been developed previously
(Kozakov et al., 2006). It uses Fast Fourier Transforms (FFT) to evaluate an approximate
energy function in discretized six dimensional space. PIPER, a consistent top performer
in the Critical Assessment of Prediction of Interaction (CAPRI)(Kozakov et al., 2013),
uses a dense packed grid with spacing of 1Å for protein structures and 0.8Å for small
molecules and a quasi-uniform, deterministic, layered Sukharev grid sequence to sample
rotational space with 70,000 rotations for protein structures and 500 for small molecules.
With one structure fixed, the second is moved and rotated in this discretized space while
measuring the energy of each position.

This energy function is defined on the grid and expressed as the sum of P
correlation functions for all possible translations $\alpha, \beta, \gamma$ of the ligand for a particular
rotation:

$$E(\alpha, \beta, \gamma) = \sum_p \sum_{i,j,k} R_p(i, j, k)L_p(i + \alpha, j + \beta, k + \gamma)$$
where \( R_p(i, j, k) \) and \( L_p(i, j, k) \) are the components of the correlation function defined on the receptor and the ligand, respectively. This expression can be efficiently calculated using \( P \) forward and one inverse Fast Fourier transform, denoted by FT and IFT, respectively:

\[
E(\alpha, \beta, \gamma) = IFT \left\{ \sum_{p} FT^* \{R_p\} FT \{L_p\} \right\}(\alpha, \beta, \gamma)
\]

\[
FT\{F\}(l, m, n) = \sum_{i,j,k} F(i, j, k) \exp^{-2\pi i(l/N_1+m/N_2+n/k/N_3)}
\]

\[
IFT\{f\}(i, j, k) = \frac{1}{N_1N_2N_3} \sum_{l,m,n} f(l, m, n) \exp^{2\pi i(l/N_1+m/N_2+n/k/N_3)}
\]

where \( i = \sqrt{-1}, N_1, N_2, \) and \( N_3 \) are the dimensions of the grid along the three coordinates. The values of each grid coordinate are pre-calculated for the receptor protein and each rotation of the ligand protein for each of the \( P \) components of the energy function described below. If \( N_1 = N_2 = N_3 = N \), then a direct calculation of the interaction would require \( N^3 \) multiplications for each grid point at \( N^3 \) grid translations, requiring \( O(N^6) \) compute time to evaluate all possible poses. Using a fast Fourier transform, each grid is converted to the frequency domain where the same values can be calculated in \( O(N^3) \) time including each grid translation \( \alpha, \beta, \gamma \). The values are then returned to the original domain. The advantage of the fast Fourier transform is that it breaks down the inputs to perform the calculations and then combines the results to produce the transform of the whole input. This can be performed recursively in \( O(N^3 \log N^3) \) time. (Cooley and Tukey, 1965).
1.2.1 Scoring Function

The energy function is given as the sum of terms representing shape complementarity, electrostatic, and desolvation contributions, the latter described by a pairwise potential:

\[
E = E_{\text{shape}} + \omega_2 E_{\text{elec}} + \omega_3 E_{\text{pair}}
\]

\[
E_{\text{shape}} = E_{\text{attr}} + \omega_1 E_{\text{rep}}
\]

\[
E_{\text{elec}} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \frac{q_i q_j}{\left(r_{ij}^2 + D^2 \exp \left(-\frac{r_{ij}^2}{4D^2}\right)\right)^{1/2}}
\]

\[
E_{\text{pair}} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \varepsilon_{ij}
\]

where \(N_R\) and \(N_L\) denote the numbers of atoms in the receptor and ligand, respectively.

According to these expressions, the shape complementarity term, \(E_{\text{shape}}\), accounts for both attractive and repulsive interactions, using a stepwise implementation of the van der Waals energy. The electrostatic term, \(E_{\text{elec}}\), is given by a simplified generalized Born-type expression. The pairwise term, \(E_{\text{pair}}\), is limited to a narrow range, \(r_{ij}\), where \(d < r_{ij} < D\) and \(\varepsilon_{ij} = 0\) outside of the predetermined range. The coefficients \(\omega_1\), \(\omega_2\), and \(\omega_3\) weight the different contributions to the scoring function. While the repulsive interaction eliminates atomic overlaps, the weight term is reduced to allow for slight variations that might occur naturally as the two structures come into contact. This is especially important when using unbound models (i.e., separately crystallized) to predict a bound conformation. As a rigid body docking method, PIPER acknowledges that these
variations exist, but assumes most are moderate. Certain cases, however, such as those labeled “difficult” in the protein-protein docking benchmark set (Hwang et al., 2010a), include a greater conformational change in the bound versus unbound state including changes in the backbone conformation. While these “difficult” cases are included in the set utilized in this text, no new methodology has been developed to address this issue specifically.

1.2.2 Pairwise Potentials using DARS

PIPER makes use of a structure-based pairwise potential named DARS (Decoys As the Reference State)(Chuang et al., 2008). DARS uses the frequency of atom-atom interactions by atom type from a set of known interactions over a generated set of decoy interactions to favor more frequently occurring natural states. These states are assumed to be low energy. The statistical potential between two atoms of types I and J within the framework of the inverse Boltzmann approach can be expressed:

\[ \varepsilon_{IJ} = -kT \ln \left( \frac{p_{ij}^{nat}}{p_{ij}^{ref}} \right) \]

where \( k \) is the Boltzmann constant, \( T \) is temperature, \( p_{ij}^{nat} \) is the naturally occurring probability of atom types I and J contacting, and \( p_{ij}^{ref} \) is the probability of atom types I and J contacting in the reference state (Sippl, 1990).

To calculate \( p_{ij}^{nat} \), a nonredundant database of native protein-protein complexes was considered (Glaser et al., 2001). After removing conflicting complexes with the test set, atoms were labelled with the 18 atom types introduced for the atomic contact
potential (ACP)(Zhang et al., 1997). Contact frequencies were then calculated for all atom types within 6Å over the test set where:

\[ p_{ij}^{nat} = \frac{\nu_{ij}^{obs}}{\sum_{i,j} \nu_{ij}^{obs}} \]

where \( \nu_{ij}^{obs} \) is the number of times atom types I and J are observed contacting in the database.

The strength of DARS as a statistical potential for improving rigid body docking results lies in the generation of an unbiased reference set. A good reference set should maintain all atom type related properties except pairwise interactions (Godzik, 1996). To that end, protein structures were docked using PIPER considering only the van der Waals interaction energy to create complexes with only shape complementarity. The atom type contact frequencies were calculated for all atom types within 6Å over this reference set to provide pairwise potentials for atom-type contacts in the context of rigid body protein docking.

1.3 'Hot Spot' identification with Computational Solvent Mapping

Computational solvent mapping is a method developed to calculate the regions on the binding surface, called 'hot spots', which contribute a disproportionate amount to the binding free energy (Hajduk et al., 2005). Multiple experimental techniques have been developed to find the 'hot spots' of target proteins. One such method, called the Multiple Solvent Crystal Structures (MSCS) method, solves the structure of the protein using X-ray crystallography after the protein crystals have been soaked in various aqueous
solutions of probe compounds, typically organic solvents (Mattos and Ringe, 1996). A variety of solvents bind in the 'hot spots' and the superimposed structures will highlight these regions with overlapping probe clusters.

The FTMap algorithm was developed to reproduce the mapping results from MSCS using PIPER to find favorable interactions of multiple probe molecules (Brenke et al., 2009). FTMap uses a small library of sixteen fragment-sized probe molecules with varying size, shape, and polarity. 2,000 bound positions are taken from PIPER for each of the probe molecules and are minimized using the Chemistry at Harvard Macromolecular Mechanics (CHARMM) potential with the Analytic Continuum Electrostatic (ACE) model. Each probe molecule is then clustered independently with a greedy approach that first considers the lowest energy structure, and then clusters the structures within 3Å root-mean-square deviation (RMSD) of that position. This process is repeated on the remaining structures, ignoring clusters with fewer than ten members. Clustering is repeated for the remaining probe molecules.

Consensus sites (CSs) are determined by aggregating overlapping probe clusters. The clusters from all probe molecules are marked by their centers of mass (COM) and neighboring clusters are considered when their COM is within 4Å RMSD. A greedy approach is again taken, where the cluster with the most neighbors are combined into the first CS, removed from consideration, then the process is repeated on the next cluster with maximum neighbors. In this way, a consensus site represents the regions on a protein surface likely to bind various molecular fragments similar to the MSCS method.
1.4 Automated protein docking

Protein docking attempts to predict the position and orientation of two bodies interacting from the structural data of each independent protein (unbound). It is a challenging problem that attempts to evaluate billions of possible conformations, scoring them in a fashion that returns the near-native structure of the complex (bound). The PIPER algorithm discussed previously, as well as many other algorithms like it, is capable of returning conformational 'hits' (<10Å RMSD) within the top 10 predicted complexes. However, many systems require hundreds, or even thousands of returned complexes to retain a successful prediction. The docking method ClusPro, making use of the rigid-body sampling algorithm PIPER, offers a solution to detect and prioritize likely conformations from a set of predicted orientations.

After running PIPER, the complexes with the lowest evaluated interaction energy are kept. Depending on options defined by a user, this can be either the top 1,000 or 1,500 complexes. The interface RMSD between the ligand components of these structures is then determined by calculating the RMSD between any ligand atom in one complex that exists within 10Å RMSD of the receptor protein to the corresponding ligand atom in the second complex. The receptor atoms remain fixed throughout, thus variations exist only in the ligand atoms. Using interface RMSD as a distance metric, the complexes are clustered in a greedy fashion by considering the highest ranked complex, the complex with the most neighbors, as a cluster center and representing all structures within 9Å RMSD. These structures are removed from consideration and the process is repeated until no structures remain. For the purpose of protein docking, clusters with fewer than 10
members are ignored. Clusters are ranked by membership, with the cluster containing the most members returned as the most likely docked orientation. Only the top 30 clusters are kept for consideration. Each cluster is represented by its original center. After clustering, the representative complex for each cluster is run through a van der Waals minimization using CHARMM with the receptor backbone fixed. This step is necessary to remove steric clashes that are allowed to exist in PIPER's predicted complexes. The receptor structure is then reset and one final minimization occurs before final results are returned.

1.5 Contributions

The work in Chapter 2 was performed with Chi Ho Ngan and Tanggis Bohnuud who developed the original version of small molecule parameterization. I have updated that version to add support for cofactors as well as improvements in speed and efficiency.
CHAPTER TWO

EXTENDED PROTEIN MAPPING

2.1 Introduction

The FTMap algorithm was designed to predict hot spots on a protein surface, where a hot spot is a focused region that contributed significantly to the binding free energy (Brenke et al., 2009). They are identified as regions that bind multiple different small organic compounds. Within a hot spot, ligands and drug like molecules bind more tightly than the rest of the ligand binding site in general. Thus, hot spots represent the site of initial focus in fragment based drug discovery (FBDD). It has been shown that FTMap is capable of determining ligand hot spots such as those used for FBDD using only a small set of preselected, fragment sized molecules.

The FBDD process begins with the identification of fragment-size compounds that typically bind in a hot spot. Effectiveness of a fragment is measured by ligand efficiency (LE), a score of binding energy divided by heavy atom count. This measure grants the best values to compounds that bind most tightly for their size. An entire library of molecules will be run against a target of interest to discover suitable compounds to serve as core fragments for further development. Fragment-sized compounds for drug discovery are kept in small libraries that can be protein family specific, have chemical features suitable for eventual expansion, consist of moieties common to human consumable medicines, or be prescreened for legal purposes and patent licensing. Having the fragment-bound protein structures provides details such as directions for fragment
expansion or potential for optimal moiety-protein interactions. However, standard NMR and X-ray crystallography techniques can be costly and time consuming.

It was shown that the initial fragments for successful FBDD experiments typically bind in FTMap identified hot spots (Hall et al., 2012). Using FTMap for the initial screening in fragment based drug discovery provides many advantages. As a computational approach, the process is economical and efficient. Compounds from a fragment library that do not bind the target in an identified hot spot are not likely to provide good core fragments for further analysis. Also, using the FTMap identified hot spots to limit the search area ignores regions where fragment compounds may physically bind, but are most likely useless for extending the core fragment into a lead molecule. In fact, using the bound pose of the core fragment as well as additional identified hot spots provides precise information for how to extend the compound. Finally, fragments identified in these additional hot spots provide useful clues as to which chemical moieties should be used to extend the lead compound into those distinct sub-pockets. To extend the FTMap algorithm to allow for an automated computational analog to fragment discovery, it was necessary to allow for compounds beyond the standard FTMap library to be mapped.

### 2.2 The Parameterization Server

The FTMap server requires specific parameters for input molecules in order to accurately predict interactions between a target protein and a probe molecule. For additional compounds to be run through the algorithm, these properties must be
calculated first. As described, FTMap begins with a rigid body docking of each compound independently using PIPER. The standard FTMap probe library contains subfragment sized molecules with molecular weights below 100 and few rotational degrees of freedom. Typical fragment sized compounds for FBDD have molecular weights between 150 and 250 with one or two rotatable bonds. To explore the conformational space in rigid body docking, it was necessary to generate multiple rotamers of the input compounds.

Figure 1. The standard FTMap 16 probe library. This set contains mostly sub-fragment-sized molecules and very few rotatable bonds as defined by Confab.
To generate molecular conformations, the application Confab was used to systematically explore the torsion angles of a molecule (O'Boyle et al., 2011b). Confab requires a chemically feasible 3D structure as input. If the user provides only a SMILES string as input, a full structure is generated using Open Babel (O'Boyle et al., 2011a). In addition, explicit hydrogen atoms are added using Open Babel for all inputs where necessary. With the input structure, Confab identifies all acyclic, single bonds between heavy atoms where each of the atoms binds to at least one additional heavy atom and labels them as rotatable. Bonds in ring structures, bonds to hydrogen atoms, or bonds that extend to only hydrogen atoms (such as the bond to the carbon atom of a methyl group) are ignored. Confab then iterates through a set of acceptable torsion angles for each rotatable bond in a torsion-driven approach. Torsion angles that create symmetric topologies are ignored for efficiency.

Conformational energies are calculated throughout conformer generation using the MMFF94 force field based on three energy terms: torsion, van der Waals, and electrostatics. An initial low energy conformation is calculated by preferentially minimizing the energy of the innermost rotatable bond, then progressing outward by bond. For further conformations, these energies are compared against the lowest energy conformation generated. Should a lower energy conformation be found, it is used in place of the initial low energy conformation. Conformations with a difference in energy above a threshold cutoff are ignored. For parameterization of user supplied compounds, the default limit of 50 kcal/mol is used. Conformations are explored in a random, non-redundant order and stored in a tree structure based on RMSD from previously calculated
conformations. This tree structure, along with one final RMSD check to remove duplicates, allow for quickly selecting only diverse conformations. This approach provides adequate coverage of the conformational space in reasonable time (O’Boyle et al., 2011b).

Due to limitations on the FTMap server and in fairness to all users, compounds that generate more than 100 conformations are rejected. Generally, structures that are appropriately sized for FBDD have three or fewer rotatable bonds and generate a reasonable number of conformations. Each conformation is then prepared individually for the FTMap server with calculation of parameters needed for the rigid body grid search and for the minimization by CHARMM. These parameters are generated by the program antechamber (Wang et al., 2006), part of the AMBER package, based on the general AMBER force field (GAFF)(Wang et al., 2004). GAFF is a complete force field; parameters are either available for all atom types or can be computed using empirical rules. This is sufficient for most organic compounds consisting of carbon, nitrogen, oxygen, sulphur, phosphorous, hydrogen, or halogen atoms, especially molecules intended for drug design.

The charge model Austin Model 1 with a bond charge correction (AM1-BCC)(Jakalian et al., 2002) was selected to calculate atomic charges based on quality and speed. This begins by assigning a predetermined charge from AM1 to each atom based on features such as formal charge and electron delocalization. Each partial charge is then adjusted using the bond charge correction based on the general atom and bond types in the molecule. Due to the simplicity of the acceptable compounds, charges generated with
the AM1-BCC model were of a similar quality to charges generated by an *ab initio*
method.

To reduce the runtime of the FTMap algorithm, user supplied molecules are then
screened to remove non-polar hydrogen atoms. These atoms are merged with their bound
heavy atom and the values of the heavy atom adjusted to represent the removed
hydrogens. The final steps of parameterization involve creation of the appropriately
formatted input files including the generated parameters for the input compounds.
Finally, the generated parameters are passed to the FTMap algorithm (Ngan et al., 2012).

As stated previously, the FTMap algorithm has been optimized and proven
effective at detecting protein hot spots using only the standard 16 probe library, including
FBDD cases (Hall et al., 2012). As such, when a user passes an additional molecule, it is
treated separately from the standard probes. The standard probes are used to detect the
protein hot spots as in a normal run of FTMap. Each additional compound passed by the
user is run through the rigid body docking step and the initial clustering step where poses
are clustered separately by compound. The additional compounds are then excluded from
the final cross clustering step in hot spot detection. The user supplied compounds that
then overlap with the FTMap predicted hot spots are returned, represented by their cluster
centers showing the lowest energy pose.

By allowing for the user supplied molecules to be identified separately by the
FTMap algorithm, it became possible for multiple new compounds to be included in a
single run. The rigid body stage still limits one ligand per run, but the new feature allows
for cofactors to remain in their bound pose on the receptor if they are known. By
occupying a small molecule binding site on the receptor, bound cofactors can remove true hot spots from discovery by FTMap, even though they are biologically correct. In cases where a bound cofactor is known, this can remove confounding results that show reasonable binding sites for which the bound ligand is previously known.

2.3 Case Study

Figure 2. Mapping of an unbound structure of thrombin (PDB ID: 1HXF) using the small molecule C2A from a ligand-bound thrombin structure (PDB ID: 2C8Z) as a user-supplied additional probe. The lowest energy cluster of C2A (with the cluster shown as cyan sticks) overlaps the main consensus site (blue lines) from the mapping of the unbound thrombin structure using the standard probe set and has an almost identical pose to the ligand from the bound structure (white sticks). Nitrogen and chlorine atoms are colored blue and green, respectively.

To demonstrate the function of the updated FTMap server and parameterization feature, we show the mapping of thrombin. Results are accessible on the FTMap
homepage under the ‘Examples’ tab. The unbound structure of thrombin (PDB ID: 1HXF) (Zhang and Tulinsky, 1997) was mapped using the 16 standard probes and an additional small molecule (HETATM ID: C2A) used in an FBDD campaign and co-crystallized with thrombin (PDB ID: 2C8Z) (Howard et al., 2006). FTMap identifies the important hot spots based on the consensus clusters of the 16 standard probes and reports the lowest energy cluster representatives of C2A within 4 Å of each of the consensus clusters, using geometric center distances in the calculations. Figure 2 shows one of the lowest energy poses of C2A generated by computational mapping on top of the most populated consensus cluster. It is interesting to note that this pose is almost identical to the bound pose of C2A, co-crystallized with thrombin (PDB ID: 2C8Z). Although C2A is a weak binder with an IC$_{50}$ of only 300 µM, FTMap was capable of detecting the interaction. The pose identified by computational mapping of C2A is particularly interesting because the chlorophenyl group occupies the S1 site fully and the NH$_2$ group protrudes from the pocket, indicating the possibility of expanding the molecule. Indeed, C2A was subsequently joined with a 12 µM ligand to generate a 220 nM inhibitor (Howard et al., 2006). Therefore, in this case, computational mapping recapitulated important protein-ligand interactions and this type of information can be very useful for screening candidate molecules in FBDD.
Figure 3. FTMap predicted hot spots on thrombin. The predicted site on the top left, ranked fourth is occupied by a single dimethyl sulfoxide. Without that cofactor present, FTMap identifies its binding pocket as a hot spot, and when C2A is included as a probe compound, six poses are predicted in that area.
Figure 4. FTMap predicted hot spots of thrombin with dimethyl sulfoxide cofactor included. The cofactor is seen to the top left of the protein. No hot spot is detected in the area as it remained occupied throughout the mapping process.

The fourth ranked predicted hot spot on thrombin is a separate binding site occupied by a single dimethyl sulfoxide. In the standard run of FTMap using thrombin this binding site contains six predicted poses of C2A and lowers the rank of each of the subsequent predicted hot spots that is occupied by the full drug molecule C7M. By parameterizing the dimethyl sulfoxide and including it in its bound pose on the unbound thrombin protein, FTMap no longer predicts a hot spot in that area. Subsequently, eleven of twelve predicted poses of C2A overlap the top predicted binding site and, as shown above, the top ranked pose aligns well to the known bound pose of that fragment. Also, the rank of the predicted hot spots increases giving a higher rank to each of the additional
hot spots in the full ligand binding site. Where the full binding site originally extended to hot spots ranked fifth and seventh, including the cofactor changes those ranks to fourth and sixth, respectively.

Figure 5. Thrombin binding site with FTMap predicted hot spots and bound full ligand C7M (silver). Here can be seen how the top predicted hot spot (cyan) along with subsequent predicted hot spots can be used to identify areas to extend a fragment to a full lead compound. Of note, the predicted hot spots in lavender and peach have increased in rank by having the bound cofactor occupy its site on the other side of the protein (not shown).

2.4 Conclusions

The protein mapping algorithm FTMap is a computational analog of experimental screening approaches, based on NMR or X-ray crystallography, to the identification of binding hot spots of proteins. The demonstrated robustness of mapping contrasts the uncertainty of finding bound poses of small ligands by traditional docking methods.
FTMap was originally implemented as a server that used only a standard set of 16 small molecules as probes. While this probe set is sufficient for the reliable identification of hot spots, it was shown that finding the distribution of further small molecules around the hot spot region can be very useful for FBDD. The extension of FTMap described here enables the user to submit arbitrary small molecules for mapping. FTMap identifies the hot spots using the standard probes, and for each additional probe provides representative poses of the lowest energy clusters located close to the hot spots. These results help to find bound poses for the user-specified molecules, show if a compound is not likely to bind in the hot spot region, and provide input for the design of larger ligands.
CHAPTER THREE

PREDICTION OF HEPARIN BINDING

3.1 Introduction

Glycosaminoglycans (GAGs) are a diverse group of polysaccharides that participate in many biological processes through the regulation of their protein partners (Bernfield et al., 1999; Gandhi and Mancera, 2008; Lindahl, 2007; Bourin and Lindahl, 1993). They are produced by almost every cell type and are most frequently found in the extracellular space and on cell surfaces where they play important roles in mediating cell-extracellular matrix interactions and cell-cell communication, and in regulating extracellular matrix structure and function (Spencer et al., 2010; Raman et al., 2005; Sasisekharan et al., 2006; Capila and Linhardt, 2002). Heparan sulfate (HS) represents a structurally varied family of GAGs. HS is a polysaccharide of disaccharide units of alternating hexuronic acid and D-glucosamine. Variations in disaccharide sulfation and hexuronic acid structure are responsible for the differences between and even within HS chains (Bernfield et al., 1999; Lindahl, 2007; Fugedi, 2003; Esko and Lindahl, 2001). Heparin, a particular member of the HS family, consists of highly-sulfated disaccharides, and is frequently used as a model compound in experimental and theoretical studies of protein-HS interactions (Forster and Mulloy, 2001). Heparin is very important in its own right, and has long been known for its capability as an anticoagulant, which has been used in humans for almost one hundred years (Capila and Linhardt, 2002). The mechanism of heparin action involves catalyzing the inactivation of thrombin by antithrombin III by causing a conformational change in antithrombin III (Bernfield et al.,
1999; Bourn and Lindahl, 1993). The ability of heparin to bind to proteins involved in regulating other cell processes such as cell proliferation suggests other medically relevant interactions (Lindahl, 2007). Indeed, over the past two decades a growing number of biological activities have been discovered to be regulated by the interaction of heparin/HS with proteins that play major roles in cancer, wound healing, infectious diseases, and inflammatory processes. Thus, it is not surprising that a 2002 review on heparin-protein interactions has been cited over 1000 times (Capila and Linhardt, 2002).

Given the importance of interactions between heparin/HS and a very large variety of proteins, considerable effort has been invested in the identification of protein regions that form such interactions. Structures have been determined by X-ray crystallography for a number of important protein-heparin/HS complexes, including complexes of antithrombin III, annexin V, and fibroblast growth factors (Gandhi and Mancera, 2008; Mulloy and Linhardt, 2001). However, the crystallization of protein-heparin/HS complexes is challenging, primarily due to inhomogeneity of GAG fragments and the nature of ionic interactions that may allow for multiple binding orientations (Imberty et al., 2007). Because of these problems, X-ray structures are available only for a small fraction of proteins that are known to interact with heparin/HS, and computational methods have been frequently used for predicting the structures of specific complexes. Early attempts were made to identify sequence determinants required for binding, and although some consensus sequences have been found, they are neither necessary nor sufficient (Cardin and Weintraub, 1989). A minimum requirement that appears to be common to all heparin binding sites is the concentration of basic residues in a particular
area of the protein surface (not necessarily close to each other in sequence), oriented in a geometry that matches the pattern of sulfate groups along the heparin/HS chain (Forster and Mulloy, 2006).

The combination of the charged residues and the requirements of a geometry that can accommodate an elongated heparin/HS molecule is expected to facilitate the identification of heparin binding sites and orientations using docking methods, and a large number of such calculations have been described in the literature. In most cases docking has been predictively applied to specific proteins without much validation, although in some cases the results were qualitatively confirmed by NMR titration or site-directed mutagenesis experiments (Forster and Mulloy, 2006; Gandhi et al., 2008; Ricard-Blum et al., 2004; Mulloy and Forster, 2008; Pita et al., 2008; Kern et al., 2003; Ballut et al., 2013; Carpentier et al., 2013; Hung et al., 2013). The number of proper method validation studies, involving known protein-heparin/HS structures as a test set, appears to be relatively limited. In an early but still very influential study, Bitomsky and Wade docked mono- and disaccharide probes using a number of programs (Bitomsky and Wade, 1999), namely GRID(Goodford, 1985), AutoDock(Morris et al., 1998), and DOCK(Ewing et al., 2001). The test set included three proteins: two structures taken from complexes of heparin with fibroblast growth factors, and the third the unbound structure of antithrombin III. Searches were performed both globally, i.e., considering the entire protein surface, and locally, restricting considerations to a box around the known heparin binding site. For each probe, results were given in terms of “interaction probabilities”, defined as the normalized number of contacts between each residue and a
set of low energy probe poses. While this measure is less informative than RMSD (root mean square deviation) from the native state used in most traditional docking tests, the relatively high interaction probabilities obtained for binding site residues indicate that the methods were able to correctly localize the heparin binding sites. However, it was also apparent that probes had comparably low energies in several different orientations. Rigid body docking of hexasaccharides was also performed. It was observed that in all docking runs, the crystal structure did not represent the most favorable conformation with respect to the force fields of the docking programs. These results were confirmed by more recent validation studies. Forster and Mulloy used AutoDock to dock an NMR structure of heparin to separately crystallized structures of the three proteins considered by Bitomsky and Wade (Forster and Mulloy, 2006). While the methodology used by Forster and Mulloy appears to be fairly rigorous, they did not present results beyond qualitative statements. More detailed results were given by Samsonov et al., who analyzed how solvent inclusion affects the results obtained by a number of docking programs (Samsonov et al., 2011). Although the main result of the paper was that inclusion of solvent improves results and that docking generally yields near-native structures, it was also reported that, on the average, there were only 1.8 to 4.5 correct poses among the top 10 lowest energy structures. It was also shown that, in terms of the energy, the average rank of the best (lowest RMSD) pose was between 11 and 33, depending on the target protein. Thus, there seems to be an agreement in the literature that docking generates near-native conformations, but selecting them among the ones generated is still a problem.
Essentially all methods used in the above studies have been developed for docking small molecules (with a few rotatable bonds) to traditional drug target proteins with well-defined binding pockets. In addition, the methods generally assume that the approximate location of the binding pocket is known, and hence restrict the search to a box around the putative binding site. In contrast, heparin/HS is substantially larger than the small ligands considered as potential drugs, and it also has more rotatable bonds. In addition, heparin/HS generally binds to shallow solvent exposed crevices or even slightly protruding regions on the protein surface rather than in deep pockets. Although the binding region must be highly charged, a number of sites on a protein may satisfy this condition, each allowing for several orientations of the ligand. Thus, it is not clear whether the small molecule molecular docking methods offer the best tools for heparin/HS docking, and whether these methods can predict position and orientations with an accuracy that is useful in applications.

We have encountered the problem of heparin docking problem as a participant in the CAPRI worldwide protein docking experiment. While the CAPRI targets are generally protein-protein complexes, Target 57 in Round 27 of the experiment required docking a six-sugar heparin to the unbound structure of fragment 423-700 of the protein BT4661, a polysaccharide binding protein from the *Bacteroides Thetaiotaomicron* heparin utilization locus. The predictor groups in CAPRI were given the atomic coordinates of the protein and a generic heparin conformation, and were expected to model the complex as accurately as possible. The BT4661-heparin complex is now deposited in the Protein Data Bank (PDB ID 4AK2), but at the time of the CAPRI
assignment no information was available on the location of heparin binding, and the ligand-free structure of the protein is still unpublished. Since the 423-700 fragment of the protein BT4661 has two domains with a deep crevice between the domains, and many positive charges on the surface, locating the most likely binding pose with any confidence was far from trivial. We tried several approaches described in previous papers on heparin docking. First we mapped the protein using FTMap, a program developed for the identification and characterization of ligand binding sites (Brenke et al., 2009), which can be considered as a substantially improved version of the GRID program (Vajda and Guarnieri, 2006). FTMap found several pockets that could favorably accommodate small ligands, but were not very likely to bind an elongated heparin molecule. Next, as in a number of previous studies, we docked mono and disaccharide fragments of heparin using AutoDock, but the program returned too many potential binding sites, leaving us with a high level of uncertainty.

After experimenting with FTMap and AutoDock, we decided to use the docking program PIPER (Kozakov et al., 2006), also implemented in our heavily used automated docking server ClusPro (Comeau et al., 2004; Kozakov et al., 2010). ClusPro was the first automated protein-protein server, and based on the results of the last three CAPRI evaluation meetings, it consistently has been the most accurate among such servers (Lensink et al., 2007; Lensink and Wodak, 2010; Lensink and Wodak, 2013). Models built by ClusPro have been reported in over 250 publications. Since PIPER and ClusPro were developed for protein-protein docking, neither had the parameters needed for heparin docking, and thus had to be appropriately expanded. This approach was fairly
successful, at least when compared to the results of the other 30 groups that participated in CAPRI and submitted predictions for Target 57. Based on the evaluator’s report (Lensink and Wodak, 2013), we obtained two medium quality models, with ligand RMSD (after superimposing the receptors) below 5 Å and interface RMSD (for residues within 10 Å of the interface) around 2 Å. There were two other groups that achieved comparable accuracy, but our model predicted the highest fraction of native contacts.

Based on our successful prediction at CAPRI we decided to establish a general method and to add heparin docking as an advanced option of the ClusPro server. The goal of this chapter is to describe this new development and the resulting capabilities. As shown by the CAPRI results, ClusPro is competitive with other approaches. However, the method also has clear limitations, and can be used only as the first step toward determining heparin-protein interactions. In fact, ClusPro performs rigid body docking of a generic tetrasaccharide heparin probe, and we do not attempt to predict detailed interactions at the atom-atom level. Our objectives are to predict the location of the heparin binding site, defined in terms of amino acid residues that are in contact with the ligand, and possibly some of the most likely orientations of the heparin chain. As will be shown, the binding sites identified by the server are highly reliable, and some of the top poses are relatively close to the native orientation of the heparin. Thus, we hope that adding heparin docking to ClusPro will be useful to the research community as the first step toward constructing more accurate interaction models. The orientations generated by ClusPro can be filtered using data from low resolution experiments, and the models can be refined by molecular dynamics (MD) or Monte Carlo simulations that use the specific
heparin/HS molecule, allow for flexibility, and may explicitly account for the effects of the solvent.

### 3.2 Methods

#### 3.2.1 The ClusPro server

Heparin docking was added as an advanced option to our protein-protein docking server ClusPro. The server has been in operation since 2004 (Comeau et al., 2004), and it was substantially upgraded to Version 2.0 in 2007 (Comeau et al., 2007). The first step of the current version is a global rigid body docking using PIPER, which is based on the fast Fourier transform (FFT) correlation approach (Kozakov et al., 2006). In PIPER the smaller protein (i.e., in our current application the heparin), is considered the ligand, which is moved about the protein receptor on a 3D grid with spacing 1.0 Å, using 70,000 rotations at each grid point, as described in chapter one. For a protein of average size this means close to a billion function evaluations. The advantage of the FFT approach is that a scoring function, written as a sum of correlation functions, can be evaluated extremely efficiently, and thus sampling the very large number of conformations is computationally feasible. For heparin docking, the interaction energy type scoring function is calculated using three terms: van der Waals attractive, van der Waals repulsive, and electrostatic energy. Weights for each of these terms were selected to maximize accuracy in a test set of heparin binding proteins (see Results). The DARS term described previously is ignored as the reference state was not selected to serve in protein-heparin docking. As will be described, in the heparin docking mode we perform sampling using two different
weight sets, and from each set we retain the 900 lowest energy structures.

The second step performed by ClusPro is clustering the retained structures using pairwise RMSD as the distance measure (Kozakov et al., 2005). The biophysical meaning of clustering is isolating highly populated low-energy basins of the energy landscape (Kozakov et al., 2013). Several studies, including ours (Comeau et al., 2004), have demonstrated that clustering algorithms generally perform better for isolating near native structures as compared with selecting low-energy conformations, if used in conjunction with exhaustive energy-based sampling such as in PIPER. The clustering of the 1800 poses starts with the lowest energy pose and grouping all poses within 9 Å. From the remaining poses not already grouped, the lowest energy pose is selected and the process repeats until no poses remain. Finally, the structures are refined by minimizing the CHARMM energy of the complexes. All heparin atoms and all side chain atoms within 5 Å of the heparin are allowed off-grid flexibility during minimization, then the protein atoms are reset and a final minimization resolves collisions. While the minimization generally removes potential steric clashes, it does not substantially change the conformation of the complexes, and thus the RMSD of our ClusPro submissions from the native complexes is fully determined by the rigid-body docking and clustering steps (Kozakov et al., 2013).

### 3.2.2 Test set selection

To select a test set of heparin-binding proteins, the Protein Data Bank (Berman et al., 2000) was searched for structures containing both 2-O-sulfo-alpha-L-idopyranuronic
acid and N,O6-disulfo-glucosamine; the two most common sugars making up heparin chains. Several structures contained a single 4-deoxy-2-O-sulfo-alpha-L-threo-hex-4-enopyranuronic acid in place of 2-O-sulfo-alpha-L-idopyranuronic acid when that sugar was the terminal sugar in the heparin chain. From these structures, we selected only proteins for which an unbound form with sequence similarity of 95% or better was also available. Additional cases were removed from consideration for containing essential ions in the heparin binding site, for structures containing missing or incomplete atomic coordinates, and for multiple conformations of the same protein target. These strict selection criteria resulted in a small test set of five proteins listed in Table 1.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Unbound PDB</th>
<th>Bound PDB</th>
<th>Heparin Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human 3-O-Sulfotransferase-3</td>
<td>1T8T (Moon et al., 2004)</td>
<td>1T8U(Moon et al., 2004)</td>
<td>2</td>
</tr>
<tr>
<td>E2 Domain of amyloid precursor-like protein 1</td>
<td>3Q7L(Xue et al., 2011b)</td>
<td>3QMK(Xue et al., 2011a)</td>
<td>2</td>
</tr>
<tr>
<td>NK1 fragment of human hepatocyte growth factor/scatter factor (HGF/SF)</td>
<td>1NK1(Chirgadze et al., 1999)</td>
<td>1GMN(Lietha et al., 2001)</td>
<td>2.5</td>
</tr>
<tr>
<td>Plasma serine protease inhibitor</td>
<td>1LQ8(Huntington et al., 2003)</td>
<td>3DY0(Li and Huntington, 2008)</td>
<td>2.5</td>
</tr>
<tr>
<td>Basic fibroblast growth factor</td>
<td>1BFG(Ago et al., 1991)</td>
<td>1BFC(Faham et al., 1996)</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1. List of Heparin Bound Protein Structures and their Unbound Forms. Length of Heparin Chain Present in Bound Structures Given in Terms of Disaccharides.

3.2.3 Heparin parameterization

The heparin molecule was extracted from each bound structure, and parameterized for use in the ClusPro program. The molecules were given unique atom naming and each atom was assigned a value for partial charge, radius, mass, and
hydrogen bond acceptor/donor status. Each atom was also assigned a mol2 atom type and bond lengths, bond angles, and bond torsions were calculated for each unique bond. The quantum chemistry system GAMESS (Schmidt et al., 1993) was used to compute Austin Model 1 (AM1) atomic charges; the information was then piped through molecule manipulation scripts (http://charles.karney.info/b2d-scripts/) using the molecular mechanics suite ANTECHAMBER (Wang et al., 2006) to perform bond charge corrections (BCCs) and to generate the final AM1-BCC atomic charges. ANTECHAMBER generates GAFF-based (Wang et al., 2004) topology files and parameter files in the CHARMM format (Brooks et al., 1983). GAFF is a complete force field, i.e., parameters are either available for all atom types or can be computed using empirical rules. The molecular parameters were then copied to the ClusPro server to allow for each heparin molecule to be used as a ligand in interaction predictions. This parameterization process is an older version of the same method described in Chapter 2.

### 3.2.4 Selecting a quality measure

Since we will cross-dock heparin/HS ligands that may have different chemical structures, we need a quality measure that allows for superimposing such molecules. We will therefore use RMSD defined in terms of a subset of atoms from each heparin chain. Each subset contains a single carbon atom from each sugar residue of the heparin chain. This carbon atom is a member of the six-member ring in each sugar bonded to the oxygen in the ring as well as a second oxygen not in the ring. It is typically named C₁ in all heparin sugars and is the only carbon in the ring bound to two unique oxygen atoms. In
addition, this atom is along the heparin backbone if we consider the backbone to be a string that passes through the connectors between sugar residues and around which the residues may rotate, and hence the RMSD defined in terms of these atoms will describe the overall position and orientation of the docked heparin molecule relative to the native ligand. We note that the RMSD measure involving a single carbon atom from each heparin subunit is similar to a backbone RMSD frequently used for describing the accuracy of results in protein-protein docking, and allows for heparin chains of varying length to be compared using a method similar to a sliding window. It is also more forgiving to differences in heparin conformation while remaining a valid indicator of heparin position and orientation.

3.2.5 Binding Site Prediction

Using the representative poses from ClusPro, an additional round of clustering was added to form binding sites. In this step heparin poses are clustered by proximity as before, but this time each pose is represented as a single point at its center of mass rather than the positions of all C$_1$ carbon atoms. Clusters that fail to meet a minimum population threshold are not considered. Another difference is that the clusters are ranked by the highest number of protein-heparin contacts per contacting protein atoms rather than by cluster population. Thus, the cluster making most contacts with the protein at any single atom is ranked highest.
3.3 Results

3.3.1 Selecting scoring function coefficients

In a standard run of the ClusPro server, a ligand (usually the smaller molecule) is systematically positioned about a receptor and the receptor-ligand interaction energy is calculated by several parameters in our rigid docking program PIPER. The best poses are then clustered to select the most likely positions and orientations of the ligand to the receptor. This functional approach was originally optimized for protein-protein docking, including minor adjustments for specific interactions such as for enzyme-inhibitor or antibody-antigen pairs. Our objective here is to optimize ClusPro for predicting protein-heparin interactions, involving the selection of scoring function coefficients and the selection of an appropriate heparin probe.

The molecular mechanics scoring function of PIPER we use for heparin docking has the general form
\[ E = w_{\text{attr}} E_{\text{attr}} + w_{\text{rep}} E_{\text{rep}} + w_{\text{elec}} E_{\text{elec}}, \]
where \( E_{\text{attr}} \) and \( E_{\text{rep}} \) denote the attractive and repulsive contributions to the van der Waals interaction energy, and \( E_{\text{elec}} \) is an electrostatic energy term (Kozakov et al., 2006). The coefficients \( w_{\text{attr}}, w_{\text{rep}}, \) and \( w_{\text{elec}} \) specify the weights of the corresponding terms, and are optimally selected for different types of docking problems (Kozakov et al., 2013). In order to select appropriate weight coefficients for heparin docking we considered the test set of heparin-bound protein structures listed in Table 1. The heparin structure was extracted from each complex, parameterized, loaded into ClusPro, and docked to the unbound (separately crystallized) protein structure. We attempted, but were unable, to discover a single set of weights capable of successfully finding the binding site on all test systems, so an approach that
takes two weight sets and concatenates them prior to clustering was developed. In both sets $w_{rep} = 0.40$ and $w_{elec} = 300$, and only the weights of the attractive van der Waals contributions differ: $w_{attr} = -0.40$ in one set, and $w_{attr} = -0.10$ in the other, the latter generating complexes with weaker shape complementarity. Since the FFT method evaluates the different energy terms separately, generating results using two different sets of weights does not increase the computing time. For each set we retain the 900 lowest energy docked structures, and the resulting 1800 structures are clustered as described previously.

3.3.2 Selection of a heparin probe

In applying ClusPro to heparin docking, the next step was to find a heparin molecule capable of representing all heparin chains in any protein-heparin interaction. Each of the heparin ligands was then docked to all five unbound protein structures in the test set. The result of this all-against-all comparison of the structures show that just one heparin molecule, from PDB structure 3QMK, is capable of predicting the bound pose of the five test cases within 10 Å. The results of docking this heparin structure against the five targets in the test set are shown in Table 2.
### Table 2. List of bound and unbound protein conformations used for docking prediction as well as rank and RMSD of best prediction and the rank of the binding site.

<table>
<thead>
<tr>
<th>Bound PDB ID</th>
<th>Unbound PDB ID</th>
<th>Chain</th>
<th>Best Probe Rank</th>
<th>RMSD (Å)</th>
<th>Binding Site Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1T8U</td>
<td>1T8T</td>
<td>B</td>
<td>5th</td>
<td>7.995</td>
<td>1st</td>
</tr>
<tr>
<td>3QMK</td>
<td>3Q7L</td>
<td>A</td>
<td>5th</td>
<td>5.961</td>
<td>1st</td>
</tr>
<tr>
<td>1GMN</td>
<td>1NK1</td>
<td>A</td>
<td>3rd</td>
<td>8.959</td>
<td>1st</td>
</tr>
<tr>
<td>3DY0</td>
<td>1LQ8</td>
<td>A</td>
<td>1st</td>
<td>7.510</td>
<td>1st</td>
</tr>
<tr>
<td>1BFC</td>
<td>1BFG</td>
<td>A</td>
<td>2nd</td>
<td>3.499</td>
<td>1st</td>
</tr>
</tbody>
</table>

#### 3.3.3 Docking results for the test set

As shown in Table 2, docking of the selected heparin probe the binding site for heparin can be predicted by the consensus of calculated heparin poses. In addition, ClusPro is capable of returning a high ranking, low RMSD result for heparin binding. These best poses are shown in the first column of Figure 6 for each test system. In these figures, the unbound protein is shown with the actual heparin pose (obtained by aligning the bound protein to the unbound protein) in green sticks, while the best predicted pose is shown in thin, cyan sticks. For each system, the predicted pose forms more atom-atom interactions than the actual heparin pose. This is a consequence of the ClusPro algorithm's energy calculation, which results in several predictions in the correct relative position but closer to the target protein. However, in the case of 1LQ8 the actual heparin pose bridges a gap between two projecting amino acid residues (Arg26 and Asn230),
while the predicted pose runs perpendicular down the valley between those residues, entirely beneath the actual heparin bridge.

As shown in Table 2, the five largest clusters generally include heparin poses that have the right orientation (with less than 10 Å RMSD from the native), but not necessarily the best rank. An atom-atom contact map is provided to show useful features even in such cases if no a priori information is available. For this map, each interaction of 4 Å or less from all returned representative poses is counted, and each protein atom is colored by its number of contacts. The second column in Figure 6 shows the contact map for the five proteins in the test set. Clearly shown along with the correct pose (obtained by aligning the bound protein to the unbound protein) are the contact maps. In each case, the hottest (reddest) area of the contact map correctly predicts the location of the bound heparin. Whiter areas have fewer contacts while dark areas have none. Only in the case of 1NK1 was there a secondary contact site (not shown), though it was clearly less important than the primary contact site shown here.
3.3.4 Analysis of predicted contacts

To observe the accuracy of predicting heparin binding sites on unbound protein structures, the contacts per protein residue were charted. These charts are given below in Figure 7. For these charts, the number of contacts made (atom-atom distance ≤ 4.0 Å; Y-axis) were counted per residue of the protein (X-axis). Counts of the contacts in the actual bound case (orange) were normalized to the same scale as the predicted unbound case (blue). These graphs show a general matching trend that any contacting residues are correctly predicted to have atom-atom contacts with the heparin chain. They also show many cases where a residue is predicted to form contacts with the heparin, though no true contacts are formed. In fact, in some cases ClusPro, which tends to optimize contacts, generates models that have better shape complementarity than the real protein-heparin complex. This problem will be discussed further in the chapter.
Figure 7. Atom-atom contacts between heparin chains and protein residues. Predicted heparin chains in blue, actual chains (normalized) in orange. The x-axis is protein residue sorted by residue number where only residues that had at least one atom-atom contact are included for clarity.

3.3.5 Applications to further proteins

The extended ClusPro server was further tested against additional proteins that did not meet our original criteria for inclusion into the test set. In one such system,
Annexin V, calcium ions present in the binding site create a favorable pocket for binding that would not otherwise exist (Capila et al., 2001). These ions were also found in a PDB entry of the unbound protein (Colloc’h et al., 2007). This unbound conformation with the included ions was run with the new heparin docking approach. The best result, shown in Figure 8, had an RMSD of 6.66 Å and was ranked second.

Figure 8. Results for Annexin V (PDB IDs 1G5N for the bound structure, 2IE7 for the unbound). Calcium ions are shown as green spheres. A. The best predicted structure is shown as thin cyan sticks. The native binding mode is shown in green. Notice that as in many other cases, the predicted structure is closer to the surface than the native one. B. Heat map based on heparin docking. Red (hot) areas have large number of contacts with the docked heparin poses, white areas have fewer, and dark areas have none.

We also docked heparin to segment FN12-FN14 of human fibronectin for which the structure of the bound conformation was not available, but one heparin binding site and one potential heparin binding site were described in the literature (Sachchidanand et al., 2002). Results for this fibronectin model are shown in Figure 9 (PDB ID 1FNH). To
run this protein model, the structure was divided into separate domains FN13 and FN14 as described. The FN13 domain contains the heparin binding site confirmed by mutation studies, and consisting of residues R98, R99, R101, R115, and K117 (residue numbering is based on the X-ray structure of the FN12-FN14 fragment). R146 may also be part of the binding site, but it is slightly removed from the other residues, in line with the mutagenesis data where change of R146 to S reduces heparin binding only 3-fold, whereas mutation of any of the other five residues reduces it at least 10-fold. For this domain, the correct binding site was the only ClusPro predicted binding site and it accurately centers on the implicated bound residues. The top three clusters predict an orientation very similar to the cluster ranked 1 (shown in magenta in Figure 9), and they all interact with the side chains R98 and R99, considered key to heparin binding.
Figure 9. Results for human fibronectin (Unbound PDB 1FNH). Subunit FN13 is on the left and FN14 is on the right. A. The predicted heparin binding residues are shown in blue. For FN13 we show the predicted heparin poses ranked 1 (magenta) and 8 (green). For FN14 we show the poses ranked 1 (cyan) and 2 (yellow). B. The same as A, but the protein is shown as a cartoon, with the R and K residues of the heparin binding site shown as sticks. Only the top ranked heparin poses are shown.

The FN14 domain contains a predicted binding site with several positively charged residues. These residues are more scattered than the heparin binding residues on FN13. Indeed, peptides from FN14 spanning residues 204–210, 217–235 and 257–271
have been implicated independently, and thus there is a level of uncertainty concerning the binding site on FN14. Accordingly, ClusPro predicts more diverse heparin poses than for FN13, and based on cluster sizes the binding to FN14 is substantially weaker. In fact, when docking to the entire FN12-14 segment, the top 7 clusters are at the heparin binding site of FN13, and only cluster 8 is located on FN14. Mapping of the latter segment separately, the top three poses are found at the site represented by poses 1 and 2 in Figure 9A. Poses 4 and 5 (not shown) are shifted as the continuation of pose 1 (cyan) toward the intersegment region on the back of the protein. Based on these five clusters, we predict that the residues most likely involved in heparin binding are K216, R230, R232, and K261. While R225 (shown on the right from the other binding residues) seems to be separated, its long side chain can also interact with the bound heparin. In addition, R207, R209, and K257 are oriented toward the intersegment region, and most likely can interact with longer heparin chains. Based on the repeated appearance of poses similar to the top ranked ones (magenta and cyan in Figure 9) among the docked structures, we consider it likely that such orientations occur. It is generally assumed that a heparin chain of 12 to 16 saccharide units bridges the two FN13 and FN14 segments. We show the poses ranked 8 for FN13 (green) and ranked 2 for FN14 (yellow) on Figure 9A, because their orientation extends the top ranked poses toward the other FN segment. Indeed, the two poses have the right directionality and distance (19.2 Å) to be connected by adding a tetrasaccharide unit bridging the intersegment region, without any major interaction with the protein. The resulting model is more curved than a recent model obtained by rigid docking of a standard heparin dodecasaccharide to the FN12-FN14 structure, where some interactions
also occur with the FN12 segment (Carpentier et al., 2013). In contrast, predicting only the tetrasaccharides binding to FN13 and FN14, our model does not reach FN12. Most of our top poses on FN13 are placed between the side chains of R98, R99, R101, and R115. According to an NMR study, only these residues are required for the binding of an octasaccharide to the FN13-FN14 units. Of course, the predicted importance of the four residues does not prove that the model is more correct, and further experimental studies, particularly exploring the heparin binding residues on FN14, are required for a better understanding of the binding mode.

3.3.6 Comparison to other computational approaches

The method presented in this chapter has been developed from our protein-protein docking program ClusPro specifically for heparin docking. Thus, for comparisons we have to restrict considerations to published heparin docking studies, as we have no intention to apply this option to any other problem. Although predictive heparin docking is performed in a substantial number of papers, very few studies evaluates the algorithm on test problems involving known protein-heparin structures. We have found such systematic evaluations in two papers, by Bitomsky and Wade and Samsonov et al. However, comparison to our results is still far from straightforward, since in each test problem both papers considered the known structure of a protein-heparin complex, separated the two molecules, and docked them back. Although some of the methods allowed for heparin flexibility, the protein was always kept fixed in its heparin-bound conformation. Furthermore, Samsonov et al. restricted the search to a very tight box
centered on the true bound position of the ligand. In contrast, our method has been
developed to solve realistic heparin docking problems, and thus in each test case we
considered the unbound (i.e., separately crystallized) protein and the “standard”
tetrasaccharide built into ClusPro, and performed a global search over the entire protein
surface without any assumption on the location of the binding site. We were able to find
unbound protein structures for most complexes considered by Bitomsky and Wade and
Samsonov et al., and used these structures in our calculations. Based on the known
complex structures we were able to determine that our method correctly identified the
binding site in all cases. However, since we solved a very different problem using much
less information on the targets, assessing the accuracy of docking in terms of RMSD
from the native complex structures does not provide a very informative comparison to
earlier results.

Bitomsky and Wade tested three docking algorithms, GRID/GROUP (Goodford,
1985), AutoDock, and DOCK (Shoichet and Kuntz, 1991), using three different
approaches. Two test cases were considered with known heparin-bound structures of
proteins, namely the complexes of fibroblast growth factor proteins bFGF (PDB ID
1BFC) and aFGF (PDB ID 2AXM, chain A)(DiGabriele et al., 1998). In both cases the
bound protein structures served as the receptors. The heparins were first docked as rigid
body chains, then a single disaccharide was used as a probe, and lastly the flexible sugar
chains were docked to the proteins. The initial rigid body approach was a simple test of
the algorithms. The disaccharide probe was used to identify the heparin binding site.
Finally, the flexible docking performed a complete test of these algorithms. Docking our
standard tetrasaccharide to the separately crystallized fibroblast growth factor structures, we were able to identify the correct binding sites in both cases. For bFGF (unbound PDB ID 1BFG), our best pose was ranked first with an RMSD of 3.51 Å using the previously defined backbone RMSD measure. As noted, Bitomsky and Wade docked the bound heparin structure extracted from the complex to the heparin-bound structure of bFGF, and reported 4.1 Å as the best RMSD from the native complex, both with AutoDock and DOCK. For aFGF (unbound PDB ID 3K1X)(Fernandez et al., 2010), our best pose was ranked third with an RMSD of 6.56 Å. Again, the authors used the original bound heparin conformation as the ligand, and docking to the heparin-bound structure of the protein, they were able to achieve their best result at 4.1 Å with AutoDock and 14.2 Å with DOCK.

<table>
<thead>
<tr>
<th>Bound PDB ID</th>
<th>Unbound PDB ID</th>
<th>Best Probe Rank</th>
<th>Best Probe RMSD (Å)</th>
<th>Top Rank RMSD (Å)</th>
<th>Binding Site Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1OJN60</td>
<td>1N7O61</td>
<td>5th</td>
<td>4.35</td>
<td>6.84</td>
<td>1st</td>
</tr>
<tr>
<td>1RWH41</td>
<td>1RWA61</td>
<td>2nd</td>
<td>4.32</td>
<td>9.71</td>
<td>1st</td>
</tr>
<tr>
<td>1G5N62</td>
<td>2IE754</td>
<td>2nd</td>
<td>6.66</td>
<td>8.60</td>
<td>1st</td>
</tr>
<tr>
<td>1T8640</td>
<td>1T8662</td>
<td>5th</td>
<td>8.00</td>
<td>8.36</td>
<td>1st</td>
</tr>
<tr>
<td>3E7163</td>
<td>2FUQ64</td>
<td>7th</td>
<td>2.55</td>
<td>12.20</td>
<td>1st</td>
</tr>
<tr>
<td>2HYW65</td>
<td>2HYW65</td>
<td>9th</td>
<td>8.05</td>
<td>17.23</td>
<td>1st</td>
</tr>
<tr>
<td>2BRS66</td>
<td>1H8U67</td>
<td>12th</td>
<td>4.21</td>
<td>13.38</td>
<td>1st</td>
</tr>
<tr>
<td>1BFB48</td>
<td>1BFG47</td>
<td>2nd</td>
<td>3.50</td>
<td>8.54</td>
<td>1st</td>
</tr>
<tr>
<td>3IN968</td>
<td>3IMN68</td>
<td>4th</td>
<td>2.50</td>
<td>6.07</td>
<td>1st*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2nd</td>
<td>5.13</td>
<td>8.68</td>
<td>1st*</td>
</tr>
</tbody>
</table>

Table 3. List of bound and unbound protein conformations used for docking prediction by Samsonov et al as well as rank and RMSD of best prediction and the RMSD of the top ranked pose. System 3IN9/3IMN contained two GAG binding sites, both are shown. The binding site for 3IN9 spanned the two bound disaccharides, resulting in both bound poses in the top ranked binding site.

Samsonov et al. examined ten proteins that were co-crystallized with various glycosaminoglycans (GAGs) rather than heparin, and hence were not selected for our
original test set. Of these ten systems, nine were found to have unbound structures available (see Table 3). The authors docked the minimized GAG structures to the bound protein structures using a box around the bound ligand position, thus performing local rather than global docking. In contrast, we performed global docking of the "standard" tetrasaccharide to unbound proteins. ClusPro ranked first the correct binding site in all cases, and for most proteins it also generated poses with fairly low RMSD from the bound structure. However, in some cases the best-ranked orientations had higher RMSD values. As we will further discuss, this result emphasizes that lower ranked clusters should also be considered for refinement.

### 3.4 Discussion

Heparin and other glycosaminoglycans are garnering attention as important regulators of function of many proteins, as well as potential drug molecules. Heparin has been used as an anticoagulant for almost a century and its mechanism of action suggests that many other uses may be possible (Lindahl, 2007; Bourin and Lindahl, 1993). Since crystallizing protein-heparin complexes for structure determination is generally difficult, computational docking can be a useful approach for understanding specific interactions. The size of heparin chains has provided a challenge for methods originally developed for docking small molecules such as AutoDock or DOCK, as well as methods that specifically target the prediction of protein-protein interactions such as our PIPER program and the ClusPro server. Our representative heparin probe is a short polymer, with size much larger than the small molecules handled by AutoDock, and significantly
smaller than the typical protein uploaded to ClusPro. This required new parameters
adjusted for molecules of this size. With appropriate updates we have extended the
capabilities of the ClusPro server to include heparin molecules. To our knowledge, there
does not already exist an interaction prediction software specific to heparin chain binding
site prediction.

A single heparin tetrasaccharide was selected to serve as the default representative
for heparin docking using ClusPro, and the approach has proven to be successful against
all test targets in predicting the heparin binding location. This observation is in good
agreement with previous reports. As noted by Mulloy and Linhardt (2001), a relatively
short heparin is a well-behaved ligand for rigid docking and the site on a protein surface
where heparin binds can be identified with good reliability. Indeed, the conformation of
heparin appears to be unusually well defined compared with other polysaccharides: in
both solution studies of unbound heparin and crystal structures of complexes, heparin is a
ribbon-like molecule with sulfo groups arranged in clusters along opposite sides of the
polysaccharide chain. Forster and Mulloy (2006) used several heparin structures for
docking, but noted that in no case has the difference between the two conformations
given rise to substantially different predictions. The number of rotatable bonds, 22 even
in this short chain, could generate over one billion conformations. These two facts
precluded using multiple conformations of the heparin chain. However, it was also noted
that the exact details of the interactions with proteins cannot be accurately predicted by
docking calculations, whether or not flexibility is taken into account (Mulloy and
Linhardt, 2001). Accordingly, our objective is the reliable identification of the heparin
binding site and possibly determining the most likely orientations of the tetrasaccharide. Using ClusPro we generate and evaluate a very large number of docked structures, and determine a consensus binding area with good reliability. As expected, the method does not provide a simulation of the interaction, or give any details of interactions between specific atoms in either structure. However, we expect that the results provide information for further exploration of the interactions between the protein and the specific heparin/HS molecule using more detailed simulation tools, primarily molecular dynamics. In fact, without the initial docking and clustering for the identification of the binding site it may be premature to focus on details.

Of the heparin chains examined, the selected tetrasaccharide probe was the shortest. With the longer chains, it was typical to see the tails peel away from the protein surface and move out into space towards the position of the neighboring protein in the crystal lattice. This caused heparin conformations with only half of the chain bound, followed by an unusual bend away from the protein surface. This was not seen in the shorter heparin molecules that were bound to the protein surface for the entire length of the chain. Therefore, we selected a heparin tetrasaccharide probe without the bend for predicting the binding site on proteins. In addition, this chain length is representative of a majority of available heparin bound structures in the Protein Data Bank, as longer structures are not free of the artifacts just described. In view of this limitation on the available structures, we do not generate three dimensional models of chains longer than four saccharide units, as we have no basis for comparison.

In spite of correctly predicting the binding sites, we encountered difficulties that
required some attention. First, several target proteins contained multiple chains and ClusPro favored the crevices found in the region between chains. Our solution was to run only a single chain at a time, though masking the region between chains (an advanced ClusPro option) would also have worked. Second, several structures contained near identical chains, and ClusPro found the same site on an identical chain, even if there was no heparin molecule present there in the bound protein structure. This caused very large RMSD values from the native complex. Lastly, on the HGF target protein (PDB ID 1NK1), a second location attracted many heparin poses and caused a predicted secondary binding site. This secondary site, however, occurs between two distinct protein domains and looked very similar to the inaccurate results found between multiple protein chains. It would be easy to exclude this fault by masking such a site or by cutting the chains into distinct domains.

More generally, we already mentioned that ClusPro frequently predicts heparin binding closer to the surface than the one observed in x-ray structures. This is particularly troubling for longer heparin chains because in many cases it is assumed that some part of the chain does not directly interact with the protein, but it is simply held in place by heparin segments that do. The assumed bridging of the fibronectin segments FN13 and FN14 is an example. Indeed, it is recognized that an important function of heparin/HS is to mediate and regulate protein-protein interactions, and this generally involves much longer heparin/HS molecules than the tetrasaccharide we routinely dock. Thus, it appears that it would be reasonable to modify the scoring function by further reducing the van der Waals attraction term, or penalizing the removal of solvent from charged residues, i.e.,
reducing direct interactions with the protein. However, a scoring function with such properties would not necessarily be able to place the short (tetrameric) heparin fragments. Thus, the optimal scoring function would depend on the length of the heparin chain. In principle, this would not be a problem, because we could optimize the interaction energy weight in a length-dependent matter. The real problem is that, apart from a few very special cases, there are no structures available for proteins with heparin chains longer than 8 saccharide units, and thus we have no data to parameterize and test the potential. Additional complications arise when one considers that most HS chains have considerably more structural variability than heparin, with long stretches of un- and undersulfated regions that might be involved in protein binding through mechanisms not captured by our model.

In spite of the above limitations, we hope that the new heparin binding site detection option in ClusPro will provide preliminary assistance in many areas where heparin is a focus of research. As shown, the server reliably identifies the heparin binding site, including the list of contact residues, on any protein. In combination with other methods, both computational and experimental, ClusPro may fill an imperative, though seemingly absent, role providing users with accurate information quickly and easily. For example, this information could be used to direct a molecular dynamics simulation targeting specific residues of a protein. These simulations could also be extended to show the formation of a heparin mediated quaternary structure. The data from ClusPro could also be used for site-directed mutagenesis with the computationally predicted heparin binding residues serving as targets for mutation. Finally, this information could be useful
in hypothesis generation. Heparin and heparan sulfate fulfil a variety of roles including structure, signaling, and growth. By incorporating data about the interaction of these GAGs and individual components of such complex processes and provide a working theory for how the components perform their respective functions.
CHAPTER FOUR

SPECIFICITY OF ENCOUNTER COMPLEX INTERACTION SITES

4.1 INTRODUCTION

Protein-protein interactions play an essential role in many biological functions, including the control of immune response, enzymatic activity, transcription, and signal transduction. One of the most investigated problems related to protein-protein association is determining interactions sites, i.e., regions on a protein surface that interact with one or more protein partners. The importance of this problem is shown by the large number of methods that have been developed for the prediction of such surface regions as described in recent reviews (Esmaielbeiki et al., 2016; Porollo and Meller, 2012; Aumentado-Armstrong et al., 2015; Xue et al., 2015). Esmaielbeiki et al. (2016) reported the performance of 64 different methods, whereas Xue et al. (2015) provided links to 19 web based servers.

In spite of the extensive research, it is generally recognized that none of the current methods works extremely well (Esmaielbeiki et al., 2016), and even some fundamental questions related to protein-protein recognition remain unanswered. First of all, it is not clear whether proteins have well defined “special” sites that preferentially interact with other proteins, or the interactions are primarily pairwise, defined by steric, hydrophobic, electrostatic, and chemical complementarity between two partner proteins (Tonddast-Navaei and Skolnick, 2015). Accordingly, current prediction methods are based on two very different models. The so called “monomer-based” methods do not account for any pairwise interaction, and the detection of putative binding sites is
typically based on physicochemical properties of the surface region (e.g., interface propensity, hydrophobicity, or desolvation properties), on geometric properties (e.g., shape of the surface region or residue mobility), or on evolutionary conservation of surface residues (Wilkins et al., 2012). In contrast, “dimer-based” prediction methods assume that better results can be obtained by algorithms that aim to predict the atomic structures of protein–protein complexes from the structures of their component proteins, and hence can, in principle, predict the binding interfaces in a binding partner specific way (Xue et al., 2015; Hwang et al., 2014). Emerging methods that involve coevolution-based residue–residue contact predictions (Ovchinnikov et al., 2014; Hopf et al., 2014) are also based on the pairwise recognition model. Both monomer-based and dimer-based models have strong support for some classes of protein complexes. For example, the active site region of enzymes is clearly the preferred interaction site for protein substrates and inhibitors, and is relatively well defined as more concave and more hydrophobic than the rest of the surface. On the other extreme, specific antibodies can be raised against many different regions of a protein.

Dimer-based methods are expected to provide more specific results because most proteins interact with multiple other proteins, some involving unique interfaces (Bertin et al., 2007). Thus monomer-based interface prediction algorithms are inherently limited, as they cannot predict interfaces specific to different binding partners (Hwang et al., 2014). It was even suggested that consideration of specific binding partners is essential for reliably predicting binding sites (Xue et al., 2015). However, there is a price for higher specificity: docking based methods require structural information on both partner proteins
and are more dependent on conformational changes (Esmaielbeiki et al., 2016). In addition, while docking may generate a set of structures that include ones that are close to the native complex, scoring functions are not accurate enough for reliable selection of the best models for all complexes. In fact, according to the experience of the ongoing Critical Assessment of Prediction of Interactions (CAPRI) experiment, the success rate of protein-protein docking hovers around 60%, i.e., no acceptable predictions were submitted for almost half of the targets, in spite of the rules allowing for 10 models (Mendez et al., 2003; Mendez et al., 2005; Lensink and Wodak, 2010; Lensink et al., 2007; Lensink and Wodak, 2013). Docking based binding site prediction methods avoid these problems by considering ensembles of low energy structures rather than trying to identify the one with the lowest energy (Hwang et al., 2014; Fernandex-Recio et al., 2004). The ensemble of the docking solutions generated by the simulation is subsequently used to project the docking energy landscape onto the protein surface by selecting contact atoms or contact residues from all structures. It was found that regions with high population of contacts occurring in the low energy docked structures tend to overlap with the actual binding sites. In particular, the normalized interface propensity (NIP) method measures the likelihood of a surface residue to be buried by frequently predicted docked poses. Similarly, the residue contact frequency (RCF) method measures the proportion of interactions each individual atom makes over all predicted docked poses (atom contact frequency or ACF) (Hwang et al., 2010b), then sums those values for each residue (Hwang et al., 2014). Several recent studies confirm
that frequently occurring inter-residue contacts provide good predictions of native interactions (Oliva et al., 2013; Vangone et al., 2013; Oliva et al., 2015).

In this chapter we pursue three different aims. First, we develop an interaction site prediction algorithm and software based on PIPER (Kozakov et al., 2006), which is recognized as one of the best docking programs currently available (Lensink and Wodak, 2010; Lensink and Wodak, 2013), and add the method as a new option to our heavily used docking server ClusPro 2.0 (Comeau et al., 2007). While we use a docking based approach similar to some earlier methods (Hwang et al., 2014; Fernandez-Recio et al., 2004), our goal is finding contiguous binding regions rather than (possibly isolated) binding site residues, and hence the frequently occurring contacts are clustered as described in the Methods. The prediction protocol will be demonstrated by applying it to enzyme-inhibitor and “other complexes” subsets of version 4.0 of the well-established protein docking benchmark (Hwang et al., 2010a). Our second aim is to emphasize that considering an entire ensemble of low energy structures (Hwang et al., 2014; Fernandez-Recio et al., 2004) in the binding funnel around the native complex, docking based method actually predict the contacts in a transition state also known as the encounter complex (Berg et al., 1985; von Hippel and Berg, 1989). The encounter complex can be thought of as an ensemble of conformations in which the two molecules can rotationally diffuse along each other, or participate in a series of “microcollisions” that properly align the reactive groups. This transition state is followed by conformational rearrangements leading to the native complex. Thus, while the ensemble of low energy docked structures provides information on the encounter complex and hence on the process of protein-
protein association, considering these structures rather than the complex itself generally limits the specificity of information on the actual interaction site. The third and possibly most interesting aim is to investigate the impact of accounting for specific pairwise interactions when determining the accuracy of interactions sites. To study this impact we consider proteins of interest and dock their known partner proteins as in other docking based interaction site prediction methods (Hwang et al., 2014; Fernandez-Recio et al., 2004). However, we also repeat the analysis using a fixed panel of 12 proteins rather than the specific interaction partner, thereby exploring how interactions with the specific binding partner improve the predictions of interactions sites. Of course, one has to keep in mind that we actually study the interactions in encounter complexes formed either between specific protein pairs or between the target protein and the unrelated probe proteins. As will be described, the comparison of results provides interesting insight on the nature and specificity in protein-protein recognition.

4.2 MATERIALS AND METHODS

4.2.1 Benchmark set of complexes

For the purpose of exploring binding site prediction we use the Protein-Protein Docking Benchmark 4.0 (Hwang et al., 2010a). This curated, non-redundant data set consists of protein pairs whose structures have been determined experimentally for both their bound complex and for each of the unbound partners. The complexes are sorted into three categories: enzyme-inhibitor (EI), antibody-antigen, and “others”; and are ranked as easy, medium, or difficult by the change in the root mean square deviation (RMSD)
between bound and unbound protein conformations. We excluded the antibody-antigen pairs from our data set as their mode of binding is different from the other cases. The remaining set included 51 enzyme-inhibitor pairs and 99 complexes in the “others” category. The advantage of using the protein benchmark set is that it also includes the unbound proteins for each complex, and we use such proteins for docking, thereby solving a more real-life problem than just separating the complex and using the resulting structures for interaction site prediction. The disadvantage of the benchmark set is that it has been developed for testing docking methods, and hence the complexes are given as two interacting proteins, although in reality many complexes include more than two proteins, frequently binding at different sites. For example, as will be shown, many receptors are homodimers, resulting in two identical or nearly identical binding sites, but the complex in the benchmark set may include only one ligand. If it works correctly, the prediction algorithm predicts both sites on a homodimer, but only one of the predictions is in the benchmark set and hence the results could be considered incorrect. Therefore we check the actual X-ray structure of the complex in the Protein Data Bank (PDB), and retain interaction sites with all ligands (Berman et al., 2000). No interaction site is added if it is not represented in the PDB file, although the binary complex still can be, and frequently is, part of a larger structure. However, no assessment of such additional interaction can be made without a rigorous analysis of the related publications, and it still has the element of potential subjectivity.
4.2.2 Interaction site prediction using the known partner protein

For each protein pair in the benchmark, the prediction of interaction sites begins with the rigid body docking of the unbound structures. For each docking run, we consider one protein to be the receptor, i.e., the protein on whose surface we will predict the binding interface, and the other protein to be the ligand. For a complete run with known interacting partners, this process is completed twice, with each protein alternating in receptor and ligand roles. We make use of the PIPER docking algorithm described above (Kozakov et al., 2006), which is based on the fast Fourier transform (FFT) correlation approach. The program is implemented in the heavily used server ClusPro 2.0 (Comeau et al., 2007), which yields good results when docking X-ray structures of two proteins with at most moderate backbone conformational changes upon binding (Kozakov et al., 2010; Kozakov et al., 2013). For binding site prediction we retain the 1500 lowest energy poses, and extract all atom-atom contacts that exist between any solvent accessible atom on the receptor protein and any atom in any predicted pose of the ligand protein. A contact is said to exist for any two respective atoms that are within 6Å of each other. For the atoms on the surface of the receptor protein, these contacts are summed and each atom is annotated with the number of contacts it is predicted to make. On the other end of these predicted contacts are the predicted ligand atoms in Euclidean space. These ligand atoms are moved to a 1Å grid such that the value at each grid point is the total number of surrounding ligand atoms in all 1500 docked structures. A grid point is considered to be a contact grid point if it is within 6Å of a receptor atom that, in turn, has at least one ligand atom within its 6Å neighborhood.
We collect the contact grid points from all of the 1500 structures and cluster them using a greedy clustering algorithm that begins by considering the highest value grid point in the set. From this center point, the cluster is expanded in three dimensions. Expansion is halted in any direction if the grid point is too far (more than 9Å) from the starting center point, or if the consensus of all neighboring grid points does not meet a secondary minimum threshold, requiring 5 out of the 26 surrounding grid points to be in the contact set. The last condition guarantees that the region defined by the cluster remains contiguous. Once the grid points that satisfy the conditions to be in the first cluster are collected, they are removed, and we select the highest value grid point among the rest to be the center of the second cluster, and construct the cluster using the same greedy algorithm. Five clusters are constructed in this ways.

The binding site prediction algorithm provides two key pieces of information. First is the structure of the receptor protein with the individual atoms labeled with the number of predicted contacts that the atom makes. This part simply projects the density of contacts onto the surface of the receptor, and does not assure contiguity. In a typical case, a select few atoms form a large number of contacts, while most of the surrounding atoms form fewer contacts. Second, areas where predicted contact grid points tend to cluster are returned as the predicted binding sites, each represented as a collection of grid points. These areas depict the likely location of the surface layer in the bound ligand protein, specifically its atoms forming an interaction with the receptor protein. The five clusters are considered as five independent predictions of the interaction site. As will be shown, these clusters are not necessarily adjacent to each other and may predict different
binding sites; however, the adjacent clusters can be merged for better coverage of sites that are too large to fit the 9Å maximum radius of the predicted sites.

4.2.3 Binding site detection using a panel of proteins

The method was extended to determine if protein binding sites could be identified in the absence of its binding partner. Borrowing from the FTMap algorithm (Brenke et al., 2009; Kozakov et al., 2015) for detecting binding hot spots for small molecule binding sites, in which a set of 16 small organic molecules serve as probes for identifying surface atoms that contribute significantly to the binding free energy, the method of binding site detection was applied docking a panel of 12 diverse “probe” proteins rather than the specific partner protein. The protein probes were selected from the data set of “others” proteins, each consisting of a single chain and chosen based on diversity of size and shape, using the principal moments of inertia to describe each structure. Each probe protein was used as the ligand and with the target protein as the receptor was run through the PIPER algorithm individually. The contact grid for each probe molecule was generated, then all 12 grids were summed for each receptor, with the number of contacts for each grid point summed across all probe ligands. Binding sites were then predicted with the same greedy clustering algorithm as before using the combined grid.
4.2.4 Quality measures

To measure the accuracy of binding site prediction for each protein in the benchmark, we considered the “true” contact grid points, defined by the native structure of the complex, as to be in the 6Å radius neighborhood of a receptor atom that has at least one ligand atom within 6Å, and calculated the precision and recall (also called sensitivity) of each predicted site for each protein. In this context, precision $P = \frac{T_p}{T_p + F_p}$ measures the fraction of predicted contact grid points that are correct, i.e., are in the set of “true” contact grid points, whereas recall $R = \frac{T_p}{T_p + F_n}$, also known as sensitivity, measures the fraction of “true” contact grid points that are retrieved in the predicted set. Using both measures describes both the quality of our predictions and prevents the possibility of under-defining the binding site. Correctly predicted sites will be defined at two cutoff values $C=0.5$ and $C=0.25$. $C=0.5$ means that at least 50% percent of the grid points predicted to be in the binding site are correct, i.e., are in the set of “true“ contact grid points, and that we also capture at least 50% of the “true” contact grid points that are within the predicted site. As will be discussed, several factors limit the achievable accuracy, and to show how the success rates depend on the required precision and recall values we also provide results for the less demanding cutoff of $C=0.25$. As shown in an exhaustive table summarizing the performance of 64 different binding site prediction methods, in most cases precision and recall are both in the 0.25 to 0.5 range (Esmaielbeiki et al., 2016).

In addition to the prediction of interaction sites, we also discuss the distribution of docked structures around the native state, expressed in terms of interface root mean
square deviation (IRMSD), frequently used as one of the measures of accuracy in docking (Mendez et al., 2003). For the calculation of IRMSD, the interface residues of the ligand protein are selected within 10 Å of any receptor atom in the native complex. For each docked structure, the ligand is superimposed onto the ligand in the X-ray structure of the complex, and the RMSD is calculated between the C\(_\alpha\) atoms of the ligand interface residues in docked and native structures.

4.3 RESULTS

4.3.1 Encounter complex interaction site

Before the analysis of the benchmark set of protein pairs, we show that constructing the predictions of contacts on an ensemble of low energy docked structures, docking based methods actually predict the binding site in encounter complexes rather than the native complex. As an example, we consider the complex formed the N-terminal domain of Enzyme I (EIN) and the histidine-containing phosphocarrier protein (HPr) (Kozakov et al., 2015; Fawzi et al., 2010). The encounter complexes formed between these two molecules have been studied by NMR paramagnetic relaxation enhancement (PRE), a technique that is exquisitely sensitive to the presence of lowly populated states in the fast exchange regime (Fawzi et al., 2010; Clore, 2008; Clore and Iwahara, 2009). By monitoring the residue-specific intermolecular PRE rates observed on the \(^{15}\)N-labeled EIN as a function of the concentration of paramagnetically labeled HPr, Fawzi et al. (2010) were able to identify and dissect two distinct classes of encounter complexes. The Class I encounter complex was in equilibrium with, and sterically occluded by, the
specific complex (i.e., directly competed sterically with the specific complex). In contrast, Class II of encounter complexes was found to bind at two other regions of EIN and was able to coexist with the specific complex to form a ternary complex ensemble. We have recently used a docking based approach to study the interaction between EIN and HPr (Kozakov et al., 2014). In order to capture conformations in all three regions reported by Fawzi et al. (2010), we retained 10,000 of the low energy conformations of the complex generated by the docking program PIPER (Kozakov et al., 2014), and have shown that the non-native structures found by the docking represent encounter complexes. The proof involved the back-calculation of PRE profiles. It was shown that the theoretical profile calculated only from the coordinates in the X-ray structure of the native EIN-HPr complex substantially deviates from the experimental PRE values for a number of residues, whereas calculating the theoretical PRE profiles based on all 10,000 low energy structures generated by the docking substantially improved the agreement with the data.
FIGURE 10. Encounter complexes and predictions of the interaction site for the EIN/HPr complex (Fawzi et al., 2010). (A) Models of the complex predicted by the center of the 10 most populated clusters of docked structures. The receptor (EIN), superimposed in the 10 complexes, is shown as grey surface. The ligand (HPr) structures are shown as cartoons, with the structure in yellow representing the native binding mode. (B) The top five predictions of the binding site shown as meshes colored blue, cyan, green, orange, and red in the order of construction. The native structure of HPr is shown as yellow cartoon. (C) Same as (B), but rotated by 90 degrees around the vertical axis toward the reader for a better view of the five predictions of the EIN binding sites.

As detailed in Methods, for the prediction of the interactions sites we retain the 1500 lowest energy structures generated by PIPER. We argued that these low energy states represent physical reality, namely encounter complexes. To demonstrate the encounter ensemble for the EIN/HPr complex we used the ClusPro server, which clusters the low energy docked structures using pairwise IRMSD as the distance measure and 9 Å as the cluster radius (Comeau et al., 2004). Figure 10A shows the centers of the 10 most populated clusters. For this complex, in all clusters the ligand structure is close to the one in the native complex (shown as yellow cartoon in Figure 10A) after superimposing the
receptors, with IRMSD ranging from 1.18 Å to over 10 Å. These structures represent rigid body rotations and small translations as the ligand explores the energy well around the native binding mode in the process of association. The binding site determination program then uses the contacts between the fixed receptor and the atoms of ligand structures in the encounter ensemble, and clusters the contacts as described above in the Methods. Figures 10B and 10C show the five clusters of contact grid points, represented by meshes colored blue, cyan, green, orange, and red. We note that HPr primarily interacts with EIN via a short helix of residues 345 to 352, which fits into a fairly deep crevice of EIN, and via a longer helix formed by residues 316 to 329. The top predicted site (blue, P=0.72, R=0.67) covers the smaller helix, where P and R denote the precision and recall for this cluster. The interface of EIN with the longer helix of HPr is covered by cluster 2 (cyan, P=0.62, R=0.60) and cluster 3 (green, P=0.84, R=0.49) of the contact grid points. Cluster 4 (orange, P=0.36, R=0.16) and cluster 5 (red, P=0.27, R=0.16) are not in the actual interface, as shown by the low precision and recall values. However, these regions form contacts in some of the encounter complexes, which shows that the method based on considering the contacts formed in low energy docked structures provides information on interactions that occur in encounter complexes rather than the target protein and its specific partner, and hence tends to overpredict the actual binding site.

4.3.2 Prediction of the interaction site using the known partner protein

The prediction method based on generating encounter complexes by docking was applied to our data set of 150 protein-protein interactions. Each interaction was run twice,
with each protein serving as the receptor or ligand, respectively, in alternating runs, for a total of 300 predicted binding sites (102 for the enzyme-inhibitor and 198 for the “other” complexes). The top five predicted binding sites from each run were retained. We examined the interactions in each protein separately and defined success as the ability to predict the contact grid points with the specified precision and recall using the five clusters of grid points as predictions. Figures 11A and 11B, respectively, show percentages of successful predictions for enzyme-inhibitor and “other” type complexes. As described in Methods, success rates were calculated using either $C=0.25$ or $C=0.5$ as the cutoff value for both precision and recall. In enzyme-inhibitor pairs, the top ranked prediction of the binding site at $C=0.25$ is accepted as correct for 78% of the complexes. Considering any of the top five predicted sites, this success rate increases to 94%. At the more demanding threshold of $C=0.5$, the success rates were 51% for the top site and close to 80% when considering any of the top five sites. Success rates were lower for the “other” type proteins: at $C=0.25$, 52% and 79%, respectively, when considering the top ranked and any of the top five predictions. At $C=0.5$, the corresponding success rates were 32% and 54%. As we will discuss, there are inherent limitations on the accuracy achievable by docking site prediction, and these success rates are in line with other current methods for binding site prediction (Esmaielbeiki et al., 2016; Hwang et al., 2014; Fernandez-Recio et al., 2004; Fernandez et al., 2005). However, we emphasize that most other methods predict contact atoms or contact residues rather than the contiguous interaction regions considered in this work, and hence more rigorous comparison of the performance is difficult. More generally, as noted in a recent comprehensive review, in
this field there are no established benchmarks and measures of performance (Esmaielbeiki et al., 2016).

Figures 11C and 11D, respectively, also show success rates for the 2nd protein in the complex, depending on the ranking of the correctly predicted site of the 1st protein for complexes in enzyme-inhibitor and “others” categories at the threshold of C=0.25. In the top chart, the bars show success rates when the interaction site in 1st protein is predicted by the top ranked prediction, and the site in the 2nd protein is predicted either by the top prediction, any of the top 5 predictions, or any prediction. In the lower chart, the bars show success rates when the interaction site in 1st protein is predicted by any of the five top ranked predictions, and the site in the 2nd protein is predicted either by the top prediction, any of the top 5 predictions, or any prediction. For enzyme-inhibitor complexes (Figure 11C), the top prediction correctly identifies the binding site on both proteins for 32% of the complexes. Almost two-thirds of cases had the correct binding sites ranked within the top five predictions for both sides, and over 90% of cases had at least one binding site is correct within the top five predictions. In the “others” category (Figure 11D), the top prediction is correct on one of the two proteins for 51% of the complexes, but the success rate is reduced to 12% if we expect the predictions to be correct for both proteins. However, over 35% of cases have both sides of the interaction correctly predicted when considering the top five binding sites, and the success rate reaches 73% when we request the correct prediction to be within the top five sites for only one of the proteins.
FIGURE 11. Success rates using partner proteins for prediction. (A) Enzyme-inhibitor complexes. Cutoff values are either 0.25 or 0.5 for both precision and recall. (B) “Other” complexes. Cutoff values for both precision and recall are as in (A). (C) Comparing the ranks of predicted binding sites for enzyme-inhibitor complexes with precision and recall both exceeding 0.25. The upper bar shows success rates when the interaction site in 1st protein is predicted by the top ranked prediction, and the site in the 2nd protein is predicted either by the top prediction, any of the top 5 predictions, or any prediction. The lower bar shows success rates when the interaction site in 1st protein is predicted by any of the five top ranked predictions, and the site in the 2nd protein is predicted either by the top prediction, any of the top 5 predictions, or any prediction. (D) Same as (C) but for “other” type complexes.

Most obvious problems with the binding site prediction occurred when the target protein interacts with multiple partner proteins, possibly even all at once. Figure 12 shows one example, the complex formed by the nitrogenase molybdenum-iron protein and nitrogenase iron protein (PDB ID 1n2c). The receptor, nitrogenase molybdenum-iron protein (PDB ID 3min, chains A, B, C, and D), which is shown in surface representation,
is a dimer of dimers, and binds two copies of the ligand, nitrogenase iron protein (PDB ID 2nip, chains A and B), which is itself a dimer. As shown in Figure 12, the interaction sites predicted by clusters 1, 3, and 5 are at one of the ligand binding site, whereas clusters 2 and 4 are at the other. However, the structure in the Protein Docking Benchmark Set (Hwang et al., 2010a) includes only one of the ligands, and hence using it for validating the prediction we may miss the agreement with the top cluster. This type of problems can be avoided by using the structure from the PDB for validation, but this will not work if not all ligands are included in the PDB file. This possibility is clearly a source of errors, but its magnitude is difficult to estimate without an exhaustive analysis of the original literature describing the proteins in the test set.

FIGURE 12. Predicting the binding sites of the nitrogenase molybdenum-iron protein. The protein, a dimer of dimers, shown in surface representation, has two identical binding sites that interact with two copies of nitrogenase iron protein shown as yellow cartoons. One site is predicted by clusters 1, 3, and 5 of contacts, whereas clusters 2 and 4 predict the other site.
4.3.3 Binding site detection using a panel of proteins

As described in the Methods, here we use a panel of 12 protein probes for the binding site detection instead of the specific partner proteins. Each probe protein consists of a single chain and has been chosen based on diversity of size and shape. For each target protein pair in our benchmark set the interaction sites were predicted by using the probe library against each of the partner proteins independently, resulting in 300 test cases. Figure 13 shows the prediction success rates using the panel of probe proteins. The results are in the same arrangement as they were presented in Figure 11 for the calculations using specific protein partners, and hence can be easily compared.
FIGURE 13. Success rates using a panel of 12 proteins for prediction. (A) Enzyme-inhibitor complexes. Cutoff values are either 0.25 or 0.5 for both precision and recall. (B) “Other” complexes. Cutoff values for both precision and recall are as in Fig 2A. (C) Comparing the ranks of predicted binding sites for enzyme-inhibitor complexes with precision and recall both exceeding 0.25. The upper bar shows success rates when the interaction site in 1st protein is predicted by the top ranked prediction, and the site in the 2nd protein is predicted either by the top prediction, any of the top 5 predictions, or any prediction. The lower bar shows success rates when the interaction site in 1st protein is predicted by any of the five top ranked predictions, and the site in the 2nd protein is predicted either by the top prediction, any of the top 5 predictions, or any prediction. (D) Same as (C) but for “other” type complexes.

A comparison is also given in Table 4, although without showing results for the 2nd protein. As will be further discussed, using the panel the top cluster provides acceptable predictions for about 10% fewer proteins than the use of the specific partner; however, the difference is reduced and in some cases completely disappears when any of
the top five clusters are considered as predictions. According to Figures 11 and 13, overall a similar number of complexes had the binding site of at least one partner correctly predicted as when using the known interacting partners as the set of probe molecules. The rate of success suggests that probe proteins can be used as a suitable replacement for cases where the binding partner is unknown or whose structure is unavailable.

<table>
<thead>
<tr>
<th>Complex</th>
<th>Rank</th>
<th>Success Rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cutoff=0.25</td>
<td>Cutoff=0.5</td>
</tr>
<tr>
<td></td>
<td>Partner</td>
<td>Panel</td>
</tr>
<tr>
<td>Enzyme-Inhibitor</td>
<td>1 78 67</td>
<td>51 45</td>
</tr>
<tr>
<td>Others</td>
<td>1 52 40</td>
<td>31 20</td>
</tr>
</tbody>
</table>

Table 4. Success rates at cutoff values 0.25 and 0.5 for both recall and precision. Rank indicates the number of predicted sites considered.

A large number of cases remain predictable for only one side of the interaction, suggesting that some interactions may only be described in terms of one binding partner that has the role of the receptor, and thus has a well-defined binding site that is recognized by the partner protein, but the reverse is not necessarily the case. The complex of the nitrogenase molybdenum-iron protein and the nitrogenase iron protein (PDB ID 1n2c), shown in Figure 12, is such an example. While the small ligand finds the two
binding sites on the nitrogenase molybdenum-iron protein, the latter is too large and too complex to predominantly bind to the interface region of the much smaller partner, the nitrogenase iron protein. Interestingly, the binding site of the latter is not found when docking the receptor, but it is identified as the top cluster when we use the panel of probe proteins. In fact for a number of proteins, docking the nonspecific panel of probe proteins yields better prediction in cluster 1 than docking the specific partner protein. At C=0.5, this occurs 11 times among the 198 interactions in “other” complexes, and 26 times among the 102 enzyme-inhibitor interactions. The majority of interaction sites that are correctly predicted by docking the partner protein are also predicted well by docking the protein probes, and the quality of results does not diminish significantly. In Figure 14, a side by side comparison of the two methods of prediction show that predicted contacts clustering in the exact same regions for the two methods, even in this instance where multiple ligand binding sites exist.
Figure 14. Predicted binding sites of the nuclear transport factor 2 (NTF2, PDB ID 1oun, chains A and B, forming a homodimer, shown as wheat surface). (A) Using the known partner protein, the GDP-bound form of the Ras-family GTPase Ran (PDB ID 1qg4, chain A, shown as slate cartoon), the predicted binding sites are ranked 1 (blue) and 3 (green) combined represent one binding site of Ran, while sites ranked 2 (cyan) and 4 (orange) combine to represent another. The native complex (PDB ID 1a2k, chains A, B, C, and D are superimposed for reference. We note that the protein benchmark set does not include Chain D, whereas the PDB structure has an additional Ran protein, denoted as chain E of the complex. (B) With the probe proteins, predicted sites rank 1 (blue) detects one binding site of NTF2, while predicted sites 2 (cyan) and 5 (red) constitute the second binding site.

4.4 DISCUSSION

This chapter described converting the well-established and heavily used protein docking method PIPER into a tool for the predicting regions on protein surfaces that are likely to interact with other proteins. Our results confirm that to obtain reliable predictions is far from simple, a fact that is already well known and clearly stated in the literature (Esmaielbeiki et al., 2016). It is also well known that the reliability of docking methods, that are the starting points of docking-based binding site prediction methods, is
also limited. The PIPER program we use performs global and systematic sampling on a
dense grid, but it assumes rigid body association. Although the “smooth” scoring function
allows for moderate overlaps, results are still unreliable if one or both proteins
substantially change conformation upon binding. There are other methods, based e.g., on
Monte Carlo search algorithms that allow for some flexibility, but sampling with these
methods is computationally much more expensive, and hence generally restricted to
regions of the presumed interface regions. Thus, the use of such methods assumes the
knowledge of the interaction sites, which is an obvious contradiction if the goal of the
analysis is finding such sites. Therefore, the use of a global docking method is well
justified. However, the biggest problem with any docking method is the limited accuracy
of the scoring function: even after generating near-native docked structures, the selection
of best models in terms of the RMSD from the native structures is generally difficult and
far from reliable.

As discussed, essentially all docking based binding site prediction methods
generate an entire ensemble of low energy structures, and extract predicted receptor-
ligand contacts from a large number of such structures. Assuming that the energy
function represent the physical interactions that occur between the two proteins, we
argued that the low energy docked structures are actually encounter complexes that occur
as a transition state in the process of protein-protein association as the interacting protein
pair evolves toward the native conformation of the complex. These encounter complexes
form an ensemble of conformations within the energy funnel around the native state such
that the two molecules still can rotationally diffuse along each other, or participate in a
series of “microcollisions”. This already requires attractive interactions between the receptor and the ligand, but assumes that the final stabilizing pairwise interactions between the two molecules are not yet developed. It is clear that such tumbling of the molecules leads to an interface region, which is broader than the one seen in the native complex, thereby affecting the potential accuracy of the method.

We also mentioned that our problem formulation differs somewhat from that of other similar methods as we try to identify contiguous surface regions that can serve as binding sites rather than just predicting atoms or residues that are likely to be in the contact interface. However, when the contact atoms are determined, the contiguous interaction sites can be easily obtained by appropriate clustering. The questions that we still have to answer are the size and number of the interactions sites. One of potential use of the predicted site is focusing the search in docking, thereby enabling the application or more accurate but computationally more expensive docking methods (Mamonov et al., 2016). To make this approach more effective requires restricting the size of the predicted sites, and we selected 9 Å as the maximum radius. This value corresponds to the effective range of electrostatic and solvation forces, and has been used by us as a clustering radius in protein-protein docking. Concerning the number of predicted sites, it is clear that we need to consider more than one. The reason is that a large fraction of proteins have two or more binding sites, and retaining a single prediction is unable to account for this. Accordingly, there is a visible jump in the success rate when we add the second prediction (Table 4). While adding further sites leads to smaller improvement, for a substantial number of proteins the site of interactions considered in the benchmark set
overlaps with the 3rd, 4th, or even 5th predicted site, in many cases because the binary complex is part of a larger structure and the target protein has a number of interaction sites, but also due to the limited accuracy of the docking algorithm. In view of these arguments, we return five clusters of contact grid points that potentially can predict five different interaction sites but can be merged if adjacent to each other, as shown in some of our examples.

The next question is the method of determining the success rate. As is well-established, we calculate precision and recall for each of the predicted site as the measure of accuracy. But what are the values that define a successful prediction? Answering this question we emphasize that our method predicts interaction sites in encounter complexes rather than in the specific complex. In the process of tumbling toward its final native state, the ligand in these encounter complexes has many nonspecific contacts with the receptor, and in the best case these are the contacts predicted by our method. Thus, we can expect substantial overprediction of the contacts, and thus partial overlap with the “true” interaction site should be considered as success. Based on this argument, the cutoff values C=0.5 and C=0.25 we have selected for both precision and recall seems to be reasonable, and are in line with the precision and recall values found for most binding site prediction methods. Accepting these parameters (i.e., five predicted sites, required precision and recall at C=0.25), our results look acceptable. Indeed, the success rates for enzyme-inhibitor and “other” type complexes are 94% and 83%, respectively, and even at C=0.5, 79% and 54% of predictions are successful for the two types of complexes.
We noted that encounter complexes are partially stabilized by nonspecific interactions. If this is the case, what is the impact of specific pairwise interactions on the prediction success rate? This question was explored by replacing the specific partner protein in the docking by a fixed panel of “probe” proteins. As shown in Table 4, there was about 10% drop in the success rate when considering only the top prediction. However, since many proteins have multiple sites, we also considered up to five predictions. Adding these sites, the difference between docking the specific protein partner and docking a fixed set of proteins decreases and at C=0.25 completely disappear. Since many of the additional sites are due to non-specific contacts, this result is not surprising. Nevertheless, the analysis shows that accounting for specific pairwise interactions leads to about 10% improvement when considering only the top predicted site, but this impact tends to be lost when by adding more predicted sites.

Since the impact of pairwise interactions under the success criteria used here are relatively small, the question is whether one could get better result by using monomer-based prediction tools. While it is difficult to give an exact answer, according to the results shown in Table 4 we think that the dimer-based approach is still somewhat better, but the difference is small, and both approaches have similar inherent limitations. The majority of monomer-based interaction site prediction approaches use machine-learning algorithms and utilize the existing, incomplete library of multimeric structures for training. In these studies, many different features such as sequence conservation, solvent accessible surface area, residue propensity, and geometric properties were employed. Though there were some signals in each of the mentioned features, none was strong
enough to make the interaction site residues dramatically stand out from the non-interacting ones. Notice that this shortcoming is caused by the same problem that limits the accuracy of scoring functions used for docking, namely the limited accuracy of representing the physical forces that drive molecular interactions. Even relatively small omission of properly representing such interactions leads to problems because the differences between the properties of interacting and non-interacting regions on protein surfaces are small (Esmailbeiki et al., 2016; Tonddast-Navaei and Skonick, 2015). It was even suggested that the majority of interface residues can participate in interactions with different proteins.

The last question to answer is that, in view of the above limitations, how can binding site prediction methods be improved? It is clear that both monomer-based and dimer-based methods can be substantially improved only by utilizing information offered by evolutionary considerations that are additional to the purely structural approaches. In monomer-based methods the evolutionary information can be and has been extensively used at the sequence level, because interface residues are generally more conserved than the rest of the protein surface (Res et al., 2005; Valencia and Pazos, 2003), and also at the level of structure as templates if appropriate homologs are available (Ma et al., 2003; Korkin et al., 2005). In dimer-based approaches most current success was achieved by adding co-evolutionary information (Ovchinnikov et al., 2014; Hopf et al., 2014). It is thus expected that the various methods that take advantage of evolutionary considerations will be integrated with structure-based methods.
The predictions from this approach could provide a useful source of information for various additional experiments. As has already been mentioned, the predicted sites could be entered into a secondary round of binding prediction for either focused rigid-body docking or alternate methods such as molecular dynamics or a Monte Carlo approach. This approach for interaction prediction could also be used for larger scale models where whole organelles or cells are modeled. In these simulations, a large variety of proteins are simulated in a finite space with interactions between all proteins being possible. Using interactions pre-calculated from this method could be used to quickly evaluate the random protein-protein interactions in a simulated cell, both between proteins that are not expected to biologically interact and to identify when two interacting proteins will form an encounter complex. The data from predicted binding sites could also be utilized in early stages of drug development, narrowing structural features to target to interrupt protein-protein interactions.

Beyond computational simulations, this information could be used for site-directed mutagenesis to identify residues essential for a given interaction. Additionally, the information could be valuable to affinity chromatography and protein assays where steric hindrance of protein interactions either through binding to beads or by protein modification such as attaching epitopes could be identified and avoided. In a similar fashion, the effect of protein labels could be evaluated based on proximity to predicted sites. Besides complete steric hindrance of protein-protein interactions by the addition of labels and tags, partial hindrance could be explained in part by blocking regions of the
encounter complex that do not prohibit binding, but do reduce the chances of an interaction forming.
BIBLIOGRAPHY


Lunin, V.V., Li, Y.G., Linhardt, R.J., Miyazono, H., Kyogashima, M., Kaneko, T., Bell, A.W., and Cygler, M. (2004). High-resolution crystal structure of Arthrobacter


CURRICULUM VITAE

Education

Boston University, Boston, MA
PhD in Bioinformatics May 2016
Structural Bioinformatics Lab with Sandor Vajda
Dissertation: Computational prediction and analysis of macromolecular interactions

Master of Science in Bioinformatics May 2013

Rochester Institute of Technology, Rochester, NY
Bachelor of Science in Bioinformatics with honors May 2010
Minors in Computer Science and Music

Research Experience

Graduate Research Assistant
Laboratory of Sandor Vajda, Boston University – Boston, MA March 2011 – Present

• Created an automated small molecule parameterization pipeline to enable use of any small molecule to serve as a probe in protein mapping server FTMap (with Ngan CH and Bohnuud T)
• Optimized protein mapping server FTMap/Param for prediction of small molecule binding in target protein hot spots (with Ngan CH and Bohnuud T)
• Created server feature for prediction of protein-heparin binding including parameterization of heparin and heparan sulfate polymers using ClusPro
• Developed algorithm for prediction of protein-protein binding sites by adapting protein mapping to be used with protein sized probes

**Bioinformatician/Contractor**

UCB Pharmaceuticals, Bioinformatics Department – Rochester, NY
February 2010 – June 2010

• Worked in a pharmaceutical informatics department creating software to perform latest techniques in drug development, specifically looking into protein-target interactions

**Undergraduate Research Assistant**

Laboratory of Paul A. Craig, RIT – Rochester, NY
July 2006 – May 2010

• Worked on Structural Biology Extensible Visualization Scripting Language to create a scripting language capable of being utilized in multiple software visualization tools
• Developed algorithm to translate between specialized scripting languages for structure visualization tools
Awards

Best Poster at CHI Discovery on Target  
October 2014

NSF IGERT Fellowship  
July 2010 – May 2012

Skills

Proficient in Python, Perl, C, Java, PHP

Experience with SQL databases, parallel computing, genetics, and genetic engineering,

Presentations and Posters

Discovery On Target – Boston, MA  
October 2014

Poster – Predicting macromolecular binding sites on protein binding partners

Awarded Best Poster Prize

ACS National Meeting – San Francisco, CA  
August 2014

Presentation – Predicting macromolecular binding sites on protein binding partners

ACS National Meeting – New Orleans, LA  
April 2013

Presentation – Protein mapping using FTMap with User-Defined Probe Sets

Publications

Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella SE, Luo L, Xia B, Beglov D, Vajda S. 2015. The FTMap family of web servers for determining and


