Protein Docking

By
Johannes Nangolo
Christopher Roach

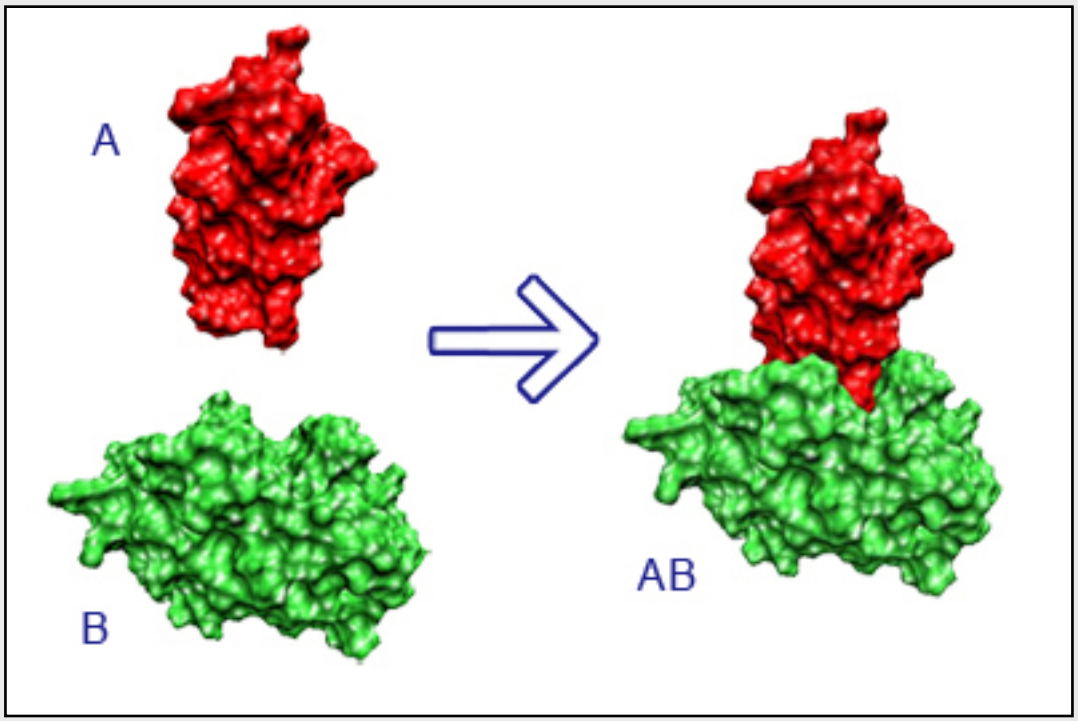
Introduction

What is Protein Docking?

"The protein-protein docking problem is the prediction of a complex between two proteins given the three-dimensional structures of the individual proteins."

- *M.L. Connolly*

What is Protein Docking?



Why is Protein Docking Important?

- **Biomolecular Structure Recognition** - Understanding the process of protein docking is part of understanding the reaction process occurring in organisms.
- **Rational Drug Design** - The efficiency of drugs is often a function of the contact area between the ligand (drug molecules) and the receptor.

Why is This Problem Interesting?

- The problem is far from being solved for several reasons:
  - The physical chemistry of protein-protein interactions is not well understood.
  - Since proteins undergo conformational changes upon binding, the conformational space is very high dimensional.

Why is This Problem Interesting?

- The problem is far from being solved for several reasons (cont.):
  - Each molecule enjoys six degrees of freedom (three rotational and three transformational), thus the number of possibilities of putting two molecules together grows exponentially with the size of the component molecules.
  - A search algorithm that rigorously goes through all the possibilities of binding between two molecules is impossible due to a near infinite search space.

Why is This Problem Interesting?

- There are no known computationally feasible methods to perform conformational searches during docking. It is for this reason, that most research has been done in the area of bound protein docking.

Two Main Protein Docking Problems

- **Bound** - The simpler of the two problems. Bound docking takes proteins from known protein-protein complexes and tries to reassemble them.
- **Unbound** - Deals with the assembly of as of yet unknown protein-protein complexes under the assumption of only small protein conformational changes.

The Components of the Docking Problem

- There are two components to the docking problem:
  - a search procedure, and
  - a scoring function

The Search Procedure

- There are essentially two different approaches to the search procedure:
  - a full search of the solution space , and
  - a gradual guided progression through the solution space.

The Search Procedure

- In the first approach the entire solution space is searched in a systematic way.
- The DOT program is an example of this approach. It does a complete search of the entire solution space by systematically rotating and transforming one molecule about the other.

The Search Procedure

- In the second approach the solution space is only partly explored in a partially random and partially criteria-guided manor, or generates fitting solutions.
- Examples of algorithms that fall into the second approach:
  - Evolutionary Algorithms
  - Monte Carlo
  - Simulated Annealing
  - Molecular Dynamics

The Scoring Function

- Despite which approach to search is taken the final outcome is that a population of solutions is chosen and it is up to the scoring function to choose the best of these solutions.
- The problem is that, despite the tremendous effort to find a fast scoring function that would be able to evaluate the huge number of solutions generated during the search phase, none has been found.

The Scoring Function

- Using the stage in which the scoring method is introduced, a docking algorithm can be grouped into one of two groups: *integrated* or *edge*.
  - An *integrated* docking algorithm integrates the scoring mechanism into the search procedure.
  - An *edge* docking algorithm applies the scoring method at the end of the search procedure.

The Scoring Function

- Geometric matching plays an important role in determining the structure of a complex, and for this reason early scoring mechanisms used nearly exclusively geometric complementarity.
- Geometric complimentarity calculations are highly efficient, and thus they are usually used as a primary filter before more costly evaluation criteria.
- Bottom line--many modern scoring functions use a combination of geometric complimentarity and other criteria.

Research

Docking Research

- Docking method
- The conformation problem - accounting for molecular flexibility.
  - Both molecules are flexible and may alter each other's structure.
- The scoring problem
- The specificity problem

Conformation - Rotamar Library

- During interaction, proteins have to be flexible to fulfill their task in metabolism.
- One kind of flexibility is the local movement of side chains.
- The aim of the project is to investigate flexibility of proteins during docking.
- A Rotamer library describes the conformation of amino-acid side chains and the associated probabilities.

Software

Protein-protein Peptide Docking

- 3D-Dock Suite (BioMolecular Modeling, Cancer Research UK)
- FTDock – Fast Fourier Transformation
- RPScore -Residue level Pair potential Score
- (uses an empirically derived score matrix of amino acid residual pair)
- MultiDock - Multiple copy side-chain refinement Dock
- (uses atomic level interaction – electrostatic force and van der Waals forces)

Bielefeld Protein Docking

- Docking method
- FFT
- Scoring method
- will use
- Rotamar library
- Enegy function (empirical)
- elastic matching
- IPHex –inteligent protein hypothesis explorer
- http://www.techfak.uni-bielefeld.de/ags/ai/projects/docking/software.html

BiGGER

- Docking method
- Grid based Geometrical Search with constraints pruning.
- Scoring method
- Electrostatics Interaction Score
- based on the Coulomb model:
- Hydrophobics Score
- (based on an estimate of the solvation energy variation caused by the formation of the complex)
- Side chain contact filter
- http://www.cqfb.fct.unl.pt/bioin/chemera/

Others

- ClusPro – DOT, ZDOCK, RDOCK (Boston University)
- ESCHER NG – NSC algorithm, parallel (Milan University)
- HADDOCK – Biochemical,Biophysical (Utrecht University Netherlands)
- DOCK - UCSF

The Protein Database

The PDB File

- The format is based on the mmCFI standard
- (macromolecular Crystallographic Information file)
- CIF is a core data dictionary (self describing data) for achieving small molecule of crystallography experiments and their results.
- Consist of : {Encoding, DDL, Dicts, Data files}
- Can be categorized in 3 categories:
    - Atom sites – gives coordinates and related information of the structure, the thermal displacement parameters, the errors in the parameters and include a specification of the component of the asymmetric unit to which an atom belongs.

The PDB Format

- Entity category which describe the chemistry of the components of the structure, as to whether they are polymer, non-polymer or water.
- Struct category which analyze and describe the structure

Title Selection

- Describe the experiment and the biological macromolecules
  - header - uniquely identifies a PDB entry through the idCode field
  - title - title for the experiment
  - compound - macromolecular contents
  - source - biological and/or chemical source
  - keywords - set of terms relevant to the entry
  - expdta - information about the experiment
  - author - people responsible for the contents
  - revdat - history of the modifications
  - jrnl - literature citation for experiment results
  - remark - experimental details

Primary Structure Section

- Contains the sequence of residues in each chain of the macromolecule
  - DBREF - cross-reference links (PDB / Databases)
  - SEQADV - identifies conflicts btw atom record in PDB and other database
  - SEQRES - amino acid or nucleic acid sequence of residues
  - MODRES

Heterogen Section

- Complete description of non-standard residues
  - HET - non-standard residues, such as prosthetic groups, inhibitors, solvent molecules,
  - HETMAN - gives the chemical name of the compound with the given hetID.
  - HETSYN - provides synonyms, if any, for the compound in the corresponding HATMAN
  - FORMUL - presents the chemical formula and charge of a non-standard group.

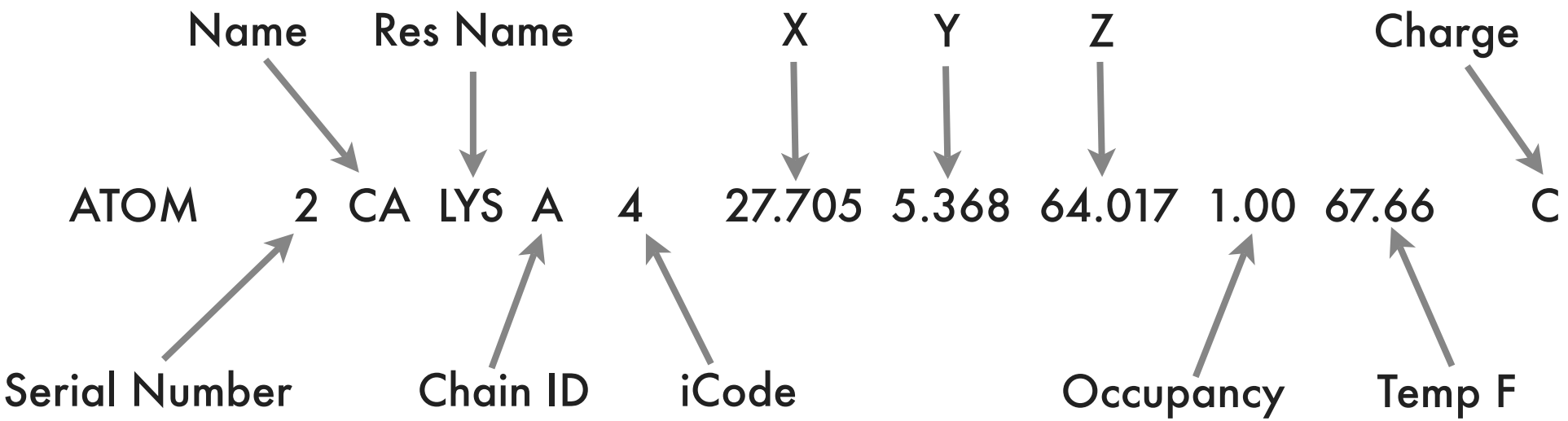
Secondary Structure

- Describes helices, sheets, and turns found in protein
  - Helix - identify the position of helices in the molecule.
  - Sheet - identify the position of sheets in the molecule
  - Turn - identify turns and other short loop turns which normally connect other secondary structure segments.

Others

- Connectivity annotation section
- Miscellaneous feature section
- Crystallographic and coordination section
- Coordination section
- Connectivity section
- Bookkeeping section

Coordination Section
(Atom)



Alpha Shape

- Formalize the idea of shape
- Capture the entire to range of “crude” to “fine” shape representations of point sets
- Example:
  - In a 2-dimension an edge between 2 point is “alpha-exposed” if there exist a circle of alpha radius such that the two points lies on the surface of the circle and the circle contains no other set points.

References

- Coarse and Reliable Geometric Alignment for Protein Docking, Y. Wang, P.K.Agarwal, P.Brown, H. Edelsbrunner, and J. Rudolph, Pacific Symposium on Biocomputing 10:64-75(2005)
- Combinatorial Shape Matching Algorithm for Rigid Protein Docking, V. Choi and N. Goyal, CPM: 15th Symposium on Combinatorial Pattern Matching, 2004
- Principles of docking: An overview of search algorithms and a guide to scoring functions. Halperin, B. Ma, H. Wolfson, and R. Nussinov. PROTEINS-NEW YORK-, 2002.
- http://www.netsci.org/Science/Compchem/feature14.html
- http://www.techfak.unibielefeld.de/ags/ai/projects/docking/welcome.html
- http://shoichetlab.compbio.ucsf.edu/docking.php
- http://www.bmm.icnet.uk/docking/
- http://www.techfak.uni-bielefeld.de/ags/ai/projects/docking/software.html
- http://www.cqfb.fct.unl.pt/bioin/chemera/