Integrative modeling of biomolecular complexes

Prof. Alexandre M.J.J. Bonvin
Bijvoet Center for Biomolecular Research
Faculty of Science, Utrecht University
the Netherlands
a.m.j.j.bonvin@uu.nl

Overview
- Introduction
- Information sources
- General aspects of docking
- Information-driven docking with HADDOCK
- Incorporating biophysical data into docking
- Conclusions & perspectives

The protein-protein interaction Cosmos

Adding the 3rd dimension
- Experimental Structures
- Computational Models

**Structural coverage of interactomes**

Unique interactions in interactomes

- ~7,500 binary interactions in *E.coli*
- ~44,900 binary interactions in *H.sapiens*

- with complete structures
- with partial (domain-domain) or complete models
- with structures for the interactors (suitable for docking)
- without structural data

**Molecular Docking**

**Methodology**

Data incorporation

Sampling

Scoring

**Data Integration during Sampling**

Global Search

Information-driven Search
What is Integrative Modeling?

Why integrative modeling?

For Experimentalists
- New hypothesis to drive experiments
- Speed up structure determination
- Increase our understanding of function

For Modelers
- Decrease high false positive rate
- Ease accuracy assessment

Related reviews
Experimental sources: mutagenesis

Advantages/disadvantages
+ Residue level information
- Loss of native structure should be checked

Detection
- Binding assays
- Surface plasmon resonance
- Mass spectrometry
- Yeast two hybrid
- Phage display libraries, ...

Experimental sources: cross-linking and other chemical modifications

Advantages/disadvantages
+ Distance information between linker residues
- Cross-linking reaction problematic
- Detection difficult

Detection
- Mass spectrometry
- Surface plasmon resonance
- Yeast two hybrid
- Phage display libraries, ...

Experimental sources: H/D exchange

Advantages/disadvantages
+ Residue information
- Direct vs indirect effects
- Labeling needed for NMR

Detection
- Mass spectrometry
- NMR $^{15}$N HSQC

Experimental sources: NMR chemical shift perturbations

Advantages/disadvantages
+ Residue/atomic level
+ No need for assignment if combined with a.a. selective labeling
- Direct vs indirect effects
- Labeling needed

Detection
- NMR $^{15}$N or $^{13}$C HSQC
Experimental sources: NMR orientational data (RDCs, relaxation)

Advantages/disadvantages
+ Atomic level
- Labeling needed

Detection
- NMR

Other potential experimental sources

- Paramagnetic probes in combination with NMR
- Cryo-electron microscopy or tomography and small angle X-ray scattering (SAXS) ==> shape information
- Fluorescence quenching
- Fluorescence resonance energy transfer (FRET)
- Infrared spectroscopy combined with specific labeling
- ...

Predicting interaction surfaces

• In the absence of any experimental information (other than the unbound 3D structures) we can try to predict interfaces from sequence information?

• WHISCY:
  WHAT Information does Surface Conservation Yield?

http://www.nmr.chem.uu.nl/whiscy

De Vries, van Dijk Bonvin. Proteins 2006

Predicting interaction surfaces

• Several other approaches have been described:
  – HSSP (Sander & Schneider, 1993)
  – Evolutionary trace (Lichtarge et al., 1996)
  – Correlated mutations (Pazos et al., 1996)
  – ConsSurf (Armon et al., 2001)
  – Neural network (Zhou & Shan, 2001) (Fariselli et al., 2002)
  – Rate4Site (Pupko et al., 2002)
  – ProMate (Neuvirth et al., 2004)
  – PPI-PRED (Bradford & Westhead, 2005)
  – PPISP (Chen & Zhou, 2005)
  – PINUP (Liang et al., 2006)
  – SPPIDER (Kufareva et al, 2007)
  – PIER (Porolo & Meller, 2007)
  – SVM method (Dong et al., 2007)
  – ... and many more since then
  – Our recent meta-server: CPORT (de Vries & Bonvin, 2011)

See review article (de Vries & Bonvin 2008)
Interface prediction servers

- PPISP (Zhou & Shan, 2001; Chen & Zhou, 2005)
  http://pipe.scs.fsu.edu/ppisp.html
- ProMate (Neuvirth et al., 2004)
  http://biportal.weizmann.ac.il/promate
- WHISCOY (De Vries et al., 2005)
  http://www.nmr.chem.uu.nl/whiscy
- PINUP (Jiang et al., 2006)
  http://sparks.informatics.iupui.edu/PINUP
- PIER (Kufareva et al., 2006)
  http://abagyan.scripps.edu/PIER
- SPPIDER (Porollo & Meller, 2007)
  http://sppider.cchmc.org

Consensus interface prediction (CPORT)

haddock.science.uu.nl/services/CPORT

Combining experimental or predicted data with docking

- a posteriori: data-filtered docking
  - Use standard docking approach
  - Filter/rescore solutions

- a priori: data-directed docking
  - Include data directly in the docking by adding an additional energy term or limiting the search space

Overview

Introduction
Information sources
General aspects of docking
Information-driven docking with HADDOCK
Incorporating biophysical data into docking
Conclusions & perspectives
Docking

- Choices to be made in docking:
  - Representation of the system
  - Sampling method:
    - 3 rotations and 3 translations
    - Internal degrees of freedom?
  - Scoring
  - Flexibility, conformational changes?
  - Use experimental information?

Explicit representation of the system

- \( x,y,z \) coordinates of each atom for both molecules
- Search method will be in real space

Grid-based representation of the system

- Discretise of the 3D structure of a protein onto a grid
  - "Shape representation" of the protein (source: biggs / Krippahl)
  - Resolution defined by grid spacing
  - Docking will require to match the shapes ("geometric matching")
  - Search in real or Fourier space

Mixed representations of the system

- Ligand and/or part of the interacting region is explicitly represented
- Remaining of structure is mapped onto a grid
- Interaction explicit atoms <-> grid
- E.g. AutoDock, ICM
Surface representation of the system: spherical harmonics

- Surface of protein described by an expansion of spherical harmonics, e.g.

\[ r(\theta, \phi) = \sum_{l=0}^{15} \sum_{m=-l}^{l} a_{lm} \psi_{lm}(\theta, \phi) \]

(source: HEX / Richie)

Surface representation of the system: surface patches

- Molecular shape representation: identify relevant "puzzle" pieces from the surface (e.g. convex or concave patches)
- Try to find matching patches (geometric hashing)
- E.g.: PatchDock (Nussinov & Wolfson)

(source: PatchDock / Nussinov & Wolfson)

Overview

- Introduction
- Information sources
- General aspects of docking
  - Representation of the system
  - Search methods
- Information-driven docking with HADDOCK
- Incorporating biophysical data into docking
- Conclusions & perspectives

Systematic search

- Sample rotations (3) and translations (3)
- For each orientation calculate a score
- Can be very time consuming depending on scoring function
- Translational search often carried out in (2D or 3D) Fourier space by convolution of the grids
- Examples:
  - FFT methods: Z-DOCK, GRAMM, FTDOCK...
  - Direct search: Bigger (uses fast boolean operations)
Protein Docking Using FFT

- Rotate
- Discretize
- Fast Fourier Transform
- Complex Conjugate
- Correlation function

Systematic search

- Search can be carried out stepwise:
  - from low to high resolution
  - from crude to more sophisticated scoring
- A decreasing number of solutions is kept at each stages
- Final solutions often further refined (EM, MD...)

“Energy-driven” search methods

- Conformational search techniques aiming at minimizing some kind of energy function (e.g. VdW, electrostatic...):
  - Energy minimization
  - Molecular dynamics
  - Brownian dynamics
  - Monte-Carlo methods
  - Genetic algorithms
  - ...
- Often combined with some simulated annealing scheme
“Energy-driven” search methods

- Still require some sampling of starting conditions:
  - How to position molecules?
  - Should be within interaction (attraction) range
  - E.g. “anchor points” in ICM (Abagyan)

Sample all combinations and for each several rotations


Dealing with flexibility

- Flexibility makes the docking problem harder!
  - Increased number of degrees of freedom
  - Scoring more difficult
- Difficult to predict a-priori conformational changes
- Current docking methodology can mainly deal with small conformational changes
- Treatment of flexibility depends on the chosen representation of the system and the search method

Dealing with flexibility: “soft docking”

- Deal with small conformational changes (e.g. side-chain rotations) by allowing overlap in the (rigid-body) docking

hard vs soft-rigid docking

- “Implicit” flexibility
- Solutions will require refinement to remove bumps
Dealing with flexibility: “soft docking”

- Implementation example in a grid-based method

Core grid points corresponding to a flexible side-chain are empty

=> no core overlap during docking

(source: biggs Krippahl)

Dealing with flexibility: docking from ensembles of conformations

- Instead of using a single starting structure use an ensemble corresponding to static snapshots of various conformations, e.g.
  - from NMR
  - from MD or other conformational sampling method

- Applicable both for rigid and flexible docking

Explicit flexibility in docking

- Only for explicit representation of systems, i.e. not for grid- or surface-based methods

- Increases computational costs
- Often only introduced in later refinement stages

Overview

- Introduction
- Information sources
- General aspects of docking
  - Representation of the system
  - Search methods
  - Dealing with flexibility
  - Scoring
- Information-driven docking with HADDOCK
- Incorporating biophysical data into docking
- Conclusions & perspectives
Scoring

• **The holy grail in docking!**

• Depends on the representation of the system and treatment of flexibility

• Depends on the type of complexes
  - e.g. antibody-antigen might behave differently than enzyme-inhibitors complexes

Score is often a combination of various (empirical) terms such as
  - Intermolecular van der Waals energy
  - Intermolecular electrostatic energy
  - Hydrogen bonding
  - Buried surface area
  - Desolvation energy
  - Entropy loss
  - Amino-acid interface propensities
  - Statistical potentials such as pairwise residue contact matrices
  - …

• Experimental filters sometimes applied *a posteriori* if data available (e.g. NMR chemical shift perturbations, mutagenesis,...)

In general, the more sophisticated the scoring function, the more computationally expensive it becomes!
Clustering protein complexes

- Docking methods often produce thousands of models.
- Scoring functions do not perfectly describe the energy landscape.
- Clustering groups similar structures together and allows better analysis.
- Similarity is defined by a specific measure (e.g. RMSD, interface RMSD, FCC)

HADDOCK: An integrative modeling platform

- Incorporates ambiguous and low-resolution data to aid the docking
- Capable of docking up to 6 molecules
- Symmetries can be leveraged
- Powerful algorithms to handle flexibility at the interface
- Final flexible refinement in explicit solvent
- One of the best performing software in CAPRI

Data-driven docking with HADDOCK

List of interface residues for protein A
List of interface residues for protein B

Effective distance \( d_{\text{bef}} \) calculated as

\[
\frac{1}{d_{\text{bef}}} = \left( \sum_{i=A}^{n_a} \sum_{j=B}^{n_b} \sum_{k=1}^{N_r} \frac{1}{d_{m_k}^{i,j}} \right)^{-1}
\]

Ambiguous Interaction Restraint:
a residue must make contact with any residue from the other list
Different fraction of restraints (typically 50%) randomly deleted for each docking trial to deal with inaccuracies and errors in the information used

(Nilges & Brunger 1991)
Searching the interaction space in HADDOCK

- Experimental and/or predicted information is combined with an empirical force field into an energy function whose minimum is searched for
  \[ V_{\text{potential}} = V_{\text{bonds}} + V_{\text{angles}} + V_{\text{torsion}} + V_{\text{non-bonded}} + V_{\exp} \]
- Search is performed by a combination of gradient driven energy minimization and molecular dynamics simulations

Classical mechanics

- Molecular dynamics: generates successive configurations of the system by integrating Newton’s second law
  \[ \frac{d^2}{dt^2} \vec{r}_i = \frac{\vec{F}_i}{m_i} \quad \text{with} \quad \vec{F}_i = -\frac{\partial V}{\partial \vec{r}_i} \]

Torsion angle dynamics

- dynamics time step dictated by bond stretching: waste of CPU time
- important motions are around torsions
- \( \sim 3 \) degrees of freedom per AA (vs \( 3N_{\text{atom}} \) for Cartesian dynamics)
- Available in DYANA, X-PLOR, CNS, X-PLOR-NIH

HADDOCK docking protocol

Succession of energy minimization and molecular dynamics protocols reminiscent of NMR structure calculations
Rigid-body Energy Minimization

Rigid-body protocol allows generation of several thousand of models in a short period of time.

Simultaneous docking of max. 6 molecules, resembling in vivo complex assembly (vs. sequential docking)

Typically, 10,000 conformations are sampled but only the best 1,000 are written to disk.

Rotational and translational optimization of the interacting partners, guided by the data-driven energy function.

Flexible simulated annealing in torsion angle space at the interface region

thorough optimization reproduces small conformational changes

Refinement in explicit solvent

Short molecular dynamics simulation in explicit solvent to refine residue-residue contacts, mainly electrostatics, at the interface.

Position restraints on backbone heavy atoms ensure conformation remains largely the same.

Explicit solvent models include TIP3P water and DMSO (membrane mimic).

Typically, all models of it1 are refined, i.e. there is no selection between it1 and itw.

Refinement in explicit solvent to optimize the contacts at the interface can be used in isolation to refine and score existing models.
Energetics & Scoring

- OPLS non-bonded parameters (Jorgensen, JACS 110, 1657 (1988))
- 8.5Å non-bonded cutoff, switching function, $\varepsilon = 10$
- Clustering of solutions
- Ranking of based on cluster-based HADDOCK score:

| Rigid: | Score = 0.01 $E_{\text{air}} + 0.01 E_{\text{vdW}} + 1.0 E_{\text{elec}} + 1.0 E_{\text{desolv}} - 0.01 \text{BSA}$ |
| Flexible: | Score = 0.1 $E_{\text{air}} + 1.0 E_{\text{vdW}} + 1.0 E_{\text{elec}} + 1.0 E_{\text{desolv}} - 0.01 \text{BSA}$ |
| Water: | Score = 0.1 $E_{\text{air}} + 1.0 E_{\text{vdW}} + 0.2 E_{\text{elec}} + 1.0 E_{\text{desolv}}$ |

- $E_{\text{air}}$: ambiguous interaction restraint energy
- $E_{\text{desolv}}$: desolvation energy using Atomic Solvation Parameters (Fernandez-Recio et al JMB 335, 843 (2004))
- BSA: buried surface area

Haddock web portal

- ~7500 registered users
- > 125000 served runs since June 2008
- > 33% on the GRID

De Vries et al. Nature Prot. 2010

Distribuion of resources

Currently (April 2015) ~ 111’000 CPU cores
41 sites

HADDOCK’s user base
The HADDOCK PDB structure gallery

>120 entries – Jan. 2015

Image collage from http://www.pdb.org

Overview

- Introduction
- Information sources
- General aspects of docking
- Information-driven docking with HADDOCK
- Incorporating biophysical data into docking
- Conclusions & perspectives

Data Incorporation

*a priori*: as Restraint
*a posteriori*: as Filter

http://www.cs.gmu.edu/~ashehu/?q=ProjectionGuidedExploration

NMR Example: CSP-driven docking

- Ub-cleaving enzyme
  - Josephin
- Which di-Ub linkage type is cleaved, K48 and/or K63 linkage?
- Collaboration with Annalisa Pastore (London, MRC)

Nicastro et al., Plos One, 2010
**NMR Example: CSP-driven docking**

**Input for docking:**
- Catalytic Triad
- 2 Binding-sites
  - CSP + Mutation
- FMD Protocol

---

**Overview**

- Introduction
- Information sources
- General aspects of docking
- Information-driven docking with HADDOCK
- Incorporating biophysical data into docking
  - Chemical shift perturbation data
  - MS-CCS as filters in docking
- Conclusions & perspectives

---

**Data Incorporation**

- a priori: as Restraint
- a posteriori: as Filter

---

*Figure 3. Cleavage of diUb chains by ataxin-3. A*

*Input for docking:*
- Lys48-linkage
- Lys63-linkage

---

*Joséphine et al., Plos One, 2010*
Integration of shape information

Ion Mobility Mass Spectrometry

- **Collision Cross Section (CCS):** rotationally averaged shape adopted by a given molecular ion under particular gas phase conditions

Integration of shape information

Ruotolo et al.,

Insight into cyanobacterial circadian timing: the KaiB-KaiC interaction

Circadian clock controlled by the Kai system consisting of three proteins: KaiA, KaiB and KaiC

Interactions define the phosphorylation status of KaiC and control the phase of the cycle

Information from MS:
- **From native MS:** Stochiometry of the KaiB-KaiC complex (6:1)
- **From HD exchange:** Binding interface and allosteric effects upon binding

Snijder et al. PNAS 111, 1379 (2014)

The KaiB-KaiC interaction: HDX

KaiB

R74

R22

K66

protected

deprotected

The KaiB-KaiC interaction: HDX
The KaiB-KaiC interaction: CCS

Collision cross section from MS allows to filter the HADDOCKing solutions

HADDOCK best scoring/most populated solution of CII

Overview

Introduction
Information sources
General aspects of docking
Information-driven docking with HADDOCK
Incorporating biophysical data into docking
- Chemical shift perturbation data
- MS-CCS as filters in docking
- Cryo-EM data
- Conclusions & perspectives

Cryo-EM data representation

- Large macromolecular complexes
  - Limited resolution (10–30Å)
  - Combined with high resolution structures

Cryo-EM data: high resolution modelling

- Rigid body fitting
  - Manual fitting (UCSF Chimera)
  - Automatic fitting software (CoLoRes, PowerFit)
  - Does not take into account the flexibility and energetics of the interface

- Flexible fitting
  - Requires an unambiguous fit of the subunits
  - The applicable resolution extend is debatable
  - Overfitting is an unresolved issue
  - Does not take into account other sources of data (mutagenesis, etc.)
HADDOCK and Cryo-EM: Tightly integrated

Rigid body fitting stage:
- Centroids are used for approximate placement
- Complex is refined directly against the map (X-ray routines)
- Many solutions are generated (10,000)

Scoring
- Physical and empirical based energy terms
- Local cross correlation between model and Cryo-EM data
- HADDOCK-score

Refinement
- Top 400 models are refined
- Simulated annealing and molecular dynamics in explicit water
- Additional cross correlation energy term

HADDOCK and Cryo-EM: Benchmarking

Benchmarked on 17 complexes from protein-protein docking benchmark 4.0
- Synthetic data generated at 10, 15 and 20 Å resolution

Cryo-EM data improves the quality of the models generated during rigid-body docking

Cryo-EM data improves the quantity of the models generated during rigid-body docking
**Effect of resolution on scoring**

- **10Å**
- **15Å**
- **20Å**

**HADDOCK and Cryo-EM: Benchmarking**

Flexible refinement brings the models closer to the native structure.

ΔRMSD (rigid – refined) histograms (positive is better)

**Test case: Modelling two proteins onto the 16S ribosome**

- 9.8Å resolution map (EMDB 1884)
- Corresponding PDB 2ykr (Guo et al., PNAS 2011)

- After refinement (400 models)
  - Six clusters
  - Best scoring: i-RMSD <2Å

**Real case: Integrative modelling of KsgA on the 16S ribosome**

Data available:
- 13.5Å resolution map (EMDB:2017)
- Ribosome crystal structure
- KsgA crystal structure
- Hydroxyl radical footprinting
- Mutagenesis data

Information:
- Cryo-EM shows the position of KsgA
- Helices 24, 27 and 45 of 16S rRNA are involved in the interaction
- Residues R221, R222, K223 of KsgA are involved in the interaction

(Zou et al., NSMB 2008; Boehringer et al., JBC 2012)
KsgA on the 16S ribosome: Current model

- Rigid body fit of KsgA in density (4adv)
- Residues R221, R222 and K223 show no favorable interactions
- Clashes!!! (yellow balls)

KsgA on the 16S ribosome: Current model

- Rigid body fit of KsgA in density (4adv)
- Residues R221, R222 and K223 show no favorable interactions
- Clashes!!! (yellow balls)

KsgA on the 16S ribosome: Can we do better?

HADDOCK-EM solutions reveal additional details of the interface

- No clashes
- R221, R222 and K223 form favorable H-bonds
- Reveals possible additional key residues: R147 and R248
Conclusions

- Cryo-EM data fully supported into HADDOCK
  - Implicitly via distance restraints to drive the docking
  - Explicitly for final optimization and scoring.
- Versatile implementation:
  - Map size-independent
  - Not all density needs to be accounted for
  - Compatible with all other complementary sources of information available in HADDOCK & symmetry
  - Can also be used with SAXS-derived shapes
- Integrative modelling can give new insights into interactions (e.g. KsgA – 16S)

More recent developments

HADDOCK is now capable of using cryo-EM data in combination with all other supported sources of information

PowerFit:
- Fast and sensitive rigid body fitting in lower-resolution densities
- Van Zundert & Bonvin, Structure 2015

DisVis:
- Quantification and visualization of the information content of distance restraints
- Van Zundert & Bonvin, Bioinformatics 2015

Overview

Introduction
Information sources
General aspects of docking
Information-driven docking with HADDOCK
Incorporating biophysical data into docking
Conclusions & perspectives

Conclusions

- (Information-driven) docking is useful to generate models of biomolecular complexes, even when little information is available
- While such models may not be fully accurate, they provide working hypothesis and can still be sufficient to explain and drive the molecular biology behind the system under study
- ... and with a little bit of effort they can be validated!
- Information-driven docking is complementary to classical structural methods