Data-driven docking for the study of biomolecular complexes

Outline

- Introduction
- Data-driven docking
- NMR-based HADDOCK application examples
- HADDOCK’s adventures in CAPRI
- Conclusions & perspectives

Study of biomolecular complexes

- Classical NMR & X-ray crystallography approaches can be time-consuming
- Problems arise with “bad behaving”, weak and/or transient complexes!
- Complementary computational methods are needed!

Understanding protein function requires to take the step from structure to interactions, the latter being much more numerous
Data-driven docking

- There is a wealth of (easily) available experimental data on biomolecular interactions.
- When classical structural studies fail, these are however often not used and the step to modelling (docking) is most of the time not taken.
- These data can be very useful to filter docking solutions or even to drive the docking and thus limit the conformational search problem.

For a review see: van Dijk et al. FEBS Journal 272, 293-312 (2005)

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

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

Detection
- - Mass spectrometry

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 15N or 13C HSQC

Experimental sources:
NMR orientational data (RDCs, relaxation)

Advantages/disadvantages
+ Atomic level
- Labeling needed

Detection
- NMR

Experimental sources:
NMR saturation transfer

Amide protons at interface are saturated
==> intensity decrease

Advantages/disadvantages
+ Residue/atomic level
+ No need for assignment if combined with a.a. selective labeling
- Labeling (including deuteration) needed

Other potential sources
- Paramagnetic probes in combination with NMR
- Interface predictions based on sequence conservation and various surface properties (e.g. evolutionary trace methods, neural networks...)
- 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
Outline

- Introduction
- Data-driven docking
- NMR-based HADDOCK application examples
- HADDOCK's adventures in CAPRI
- Conclusions & perspectives

Combining experimental 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

Data-driven docking: HADDOCK
"High Ambiguity-Driven DOCKing"

Interface definition:
- Active residues:
  involved in the interaction
  (e.g. from NMR data, mutagenesis, ...)
  and high solvent accessibility
- Passive residues:
  all solvent accessible neighbors

Ambiguous interaction restraints:

\[
d_{ij}^{\text{amb}} = \left( \sum_{k=1}^{n_{\text{amb}}} \sum_{l=1}^{n_{\text{amb}}} \frac{1}{d_{kl}} \right)^{1/2}
\]

Dominguez, Boelens & Bonvin (2003). JACS 125, 1731

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

Dealing with flexibility in HADDOCK

- Docking from ensembles of starting structures (e.g. from MD)

- "Soft" docking by scaling down intermolecular interaction

- Explicit flexibility introduced step-wise:
  1) First side-chains at interface
  2) Then both side-chains and backbone at interface
HADDOCK docking protocol

- Position proteins 150Å away from each other and apply random rotations
- Rigid body energy minimization: first only rotations, then rotations + translations
- Final refinement in explicit solvent
- Clustering and analysis
- Semi-flexible simulated annealing in torsion angle space:
  1. Rigid body dynamics
  2. SA with flexible side-chains at the interface
  3. SA with flexible backbone and side-chains at the interface

Outline

- Introduction
- Data-driven docking
- NMR-based HADDOCK application examples
  - Ubch5-Not4 (E2-E3 complex)
  - K48-linker di-ubiquitin
- HADDOCK's adventures in CAPRI
- Conclusions & perspectives

Ubiquitination pathway

- NOT4: E3 ubiquitin ligase
- RING Domain: Responsible for the E2 interaction

The Not4 Ring finger-Ubch5B complex

- NOT4 RING6 finger
- NOT4: E3 ubiquitin ligase
- RING Domain: Responsible for the E2 interaction
- Ubch5B: E2 ubiquitin conjugating enzyme
- Homology model from Ubc4 (90% identity) + MD

Source: http://www.hgu.mrc.ac.uk/Research/Gordon

(Note: NMR structure now available; Houben et al. JMB 2004)
The Not4 Ring finger-Ubch5B complex (HADDOCK results)

Mutagenesis

"HADDOCK-directed mutagenesis & two-hybrids experiments"

Altered specificity mutants!

LexA-UbcH5B

B42-CNOT4
D48K
E49K
D48K,E49K

Outline

• Introduction
• Data-driven docking
• NMR-based HADDOCK application examples
  • Ubch5-Not4 (E2-E3 complex)
  • K48-linker di-ubiquitin
• HADDOCK’s adventures in CAPRI
• Conclusions & perspectives

Best HADDOCK solution after inclusion of the mutagenesis data

Dominguez, Bonvin, Winkler, van Schaik, Timmers & Boelens, Structure 2004
Ubiquitination pathway

Source: http://www.hgu.mrc.ac.uk/Research/Gordon

K48-linked Ub2

Structural Properties of Polyubiquitin Chains in Solution
Ranjit Varadan, Olivier Weiker*, Cecile Picquet* and David Finkman**

Department of Chemistry and Biochemistry, Center of Biophysics, and University of Pittsburgh, University of Cambridge, UK. Cambridge, UK. Email: O. Winkler, Center of Biophysics, University of Pittsburgh, Pittsburgh, PA 15261, USA.

Crystal structure: 1AAR

RDC analysis with Pales (JACS (2000) 122:3791)

Q-factor = 0.44


Exp. Data: (Bruker 600 @298K)

• 1H-15N HSQC

• 15N R1 and R2, 15N(1H) NOE

• H-N RDCs: liquid crystalline

K48-linked Ub2

UB2: AIRs definition

Distal domain:
- 10 active residues
- 12 passive residues

Proximal domain:
- 12 active residues
- 11 passive residues

Total:
- 22 AIRs
- 4 isopeptide bond restraints
- 2x46 RDCs

Red: active
Green: passive
Orange: moving but not accessible
Yellow: neighbors but not accessible

ubiD: distal domain
ubiP: proximal domain

ubiD: distal domain
ubiP: proximal domain
**Outline**

- Introduction
- Data-driven docking
- NMR-based HADDOCK application examples
- HADDOCK's adventures in CAPRI
- Conclusions & perspectives

---

**Crystal vs Solution Structure**

<table>
<thead>
<tr>
<th></th>
<th>Crystal</th>
<th>Solution structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helix-angle (°)</td>
<td>128</td>
<td>154 (6)</td>
</tr>
<tr>
<td>RDC Q-factor</td>
<td>0.44</td>
<td>0.15 (0.01)</td>
</tr>
<tr>
<td>BSA (Å²)</td>
<td>1534</td>
<td>1749 (54)</td>
</tr>
</tbody>
</table>

RMSD: 1.7±0.3 Å

---

**Independent validation using diffusion anisotropy data**

<table>
<thead>
<tr>
<th></th>
<th>X-ray</th>
<th>Solution structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \chi^2/\text{df} )</td>
<td>4.8</td>
<td>3.3 (0.2)</td>
</tr>
</tbody>
</table>

---

**Crystal dimer \( Q = 0.44 \)**

**Solution structure \( Q = 0.15 \)**

RDC analysis with Pales (JACS (2000) 122:3791)

---

Van Dijk, Fushman & Bonvin Proteins (2005)
HADDOCK’s adventures in CAPRI

“Critical assessment of predicted interactions”
blind test for protein-protein docking
http://capri.ebi.ac.uk

- We participated to rounds 4 to 7 for a total of 10 targets
- For HADDOCK, we derived information to define AIRs mainly from literature and sequence information:
  - Known mutations
  - Epitope mapping
  - Conservation of exposed surface residues
  - Prediction from a neural network taking sequence alignments (conservation) and solvent accessibility as input (H.X. Zhou, Y. Shun (2001), Proteins, 44 (3), 336-343).

Capri scoring criteria

<table>
<thead>
<tr>
<th>Rank</th>
<th>( F_{\text{native}} )</th>
<th>I-RMSD [Å]</th>
<th>i-RMSD [Å]</th>
</tr>
</thead>
<tbody>
<tr>
<td>High ***</td>
<td>&gt; 0.5</td>
<td>( x &lt; 1.0 )</td>
<td>or ( x &lt; 1.0 )</td>
</tr>
<tr>
<td>Medium **</td>
<td>&gt; 0.3</td>
<td>( 1 &lt; x &lt; 5 )</td>
<td>or ( 1 &lt; x &lt; 2 )</td>
</tr>
<tr>
<td>Acceptable *</td>
<td>&gt; 0.1</td>
<td>( 5 &lt; x &lt; 10 )</td>
<td>or ( 2 &lt; x &lt; 4 )</td>
</tr>
<tr>
<td>Incorrect</td>
<td>&lt; 0.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- \( F_{\text{native}} \): fraction of native contacts (within 5Å)
- I-RMSD: rmsd on second protein after superposition on first
- i-RMSD: rmsd on interface residues (within 10Å)

Targets 18 and 19

T18: Only conservation data for one of the two molecules
  - Antibody: CDR
  - Antigen: known epitopes, but none in binding site!!!

Bad or incomplete data, bad result......
Still 40% of epitopes correct, but wrong rotation!

Target 15 **

- Complex between the ImmD immunity protein and the catalytic domain of Colicin D (bound-bound but with shaved surface side-chains)

Bien-victoire

<table>
<thead>
<tr>
<th>fr-native</th>
<th>I-RMSD</th>
<th>i-RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>T15_P75.6.B</td>
<td>0.893</td>
<td>1.331</td>
</tr>
<tr>
<td>T15_P75.3.B</td>
<td>0.893</td>
<td>0.632</td>
</tr>
<tr>
<td>T15_P75.5.B</td>
<td>0.875</td>
<td>1.393</td>
</tr>
<tr>
<td>T15_P67.7.B</td>
<td>0.875</td>
<td>0.547</td>
</tr>
<tr>
<td>T15_P75.1.B</td>
<td>0.857</td>
<td>2.145</td>
</tr>
<tr>
<td>T15_P75.2.B</td>
<td>0.768</td>
<td>3.289</td>
</tr>
<tr>
<td>T15_P75.7.B</td>
<td>0.679</td>
<td>5.508</td>
</tr>
<tr>
<td>T15_P57.8.A</td>
<td>0.500</td>
<td>6.044</td>
</tr>
<tr>
<td>T15_P84.1.B</td>
<td>0.464</td>
<td>5.382</td>
</tr>
<tr>
<td>T15_P75.8.B</td>
<td>0.429</td>
<td>10.636</td>
</tr>
</tbody>
</table>

---

Total #submissions: 100 (target cancelled)
7 acceptables 5 medium 4 high from 6 groups

HADDOCK best (#1 modell)
Target (IV74)

Little data ... still good results

(Graille et al. EMBO 2004)
Target 10: data used in docking

- Low pH trimeric form of the tick-borne encephalitis virus glycoprotein

Data used:
- Epitope mapping
- Conservation

90°

Target 10 **

- Fuzzy data... good results!

<table>
<thead>
<tr>
<th></th>
<th>fr-native</th>
<th>L-RMSD</th>
<th>l-RMSd</th>
</tr>
</thead>
<tbody>
<tr>
<td>T10_P17.10.B</td>
<td>0.302</td>
<td>2.885</td>
<td>1.882</td>
</tr>
<tr>
<td>T10_P04.4.B</td>
<td>0.191</td>
<td>9.273</td>
<td>5.478</td>
</tr>
<tr>
<td>T10_P04.8.B</td>
<td>0.185</td>
<td>10.267</td>
<td>5.840</td>
</tr>
<tr>
<td>T10_P16.2.C</td>
<td>0.150</td>
<td>8.456</td>
<td>4.513</td>
</tr>
<tr>
<td>T10_P04.6.B</td>
<td>0.150</td>
<td>16.792</td>
<td>9.947</td>
</tr>
<tr>
<td>T10_P14.2.C</td>
<td>0.124</td>
<td>13.337</td>
<td>8.754</td>
</tr>
<tr>
<td>T10_P04.7.C</td>
<td>0.120</td>
<td>14.669</td>
<td>8.590</td>
</tr>
<tr>
<td>T10_P14.5.B</td>
<td>0.116</td>
<td>13.112</td>
<td>8.647</td>
</tr>
</tbody>
</table>

HADDOCK best

Target

Total #submissions: 170
3 acceptables 1 medium 0 high
from 4 groups

Bressanelli et al. EMBO 2004

Target 11/12: data used in docking

- Cohesin-dockerin complex models
  (T11: dockerin is a homology model; T12: dockerin is the bound form)

Cohesin:
- Conservation
  - Mutagenesis
  - Conservation

Dockerin:

Target 11 **

- “Good” data... reasonable results (as good as homology model)

<table>
<thead>
<tr>
<th></th>
<th>fr-native</th>
<th>L-RMSD</th>
<th>l-RMSd</th>
</tr>
</thead>
<tbody>
<tr>
<td>T11_P12.6.B</td>
<td>0.418</td>
<td>6.111</td>
<td>1.173</td>
</tr>
<tr>
<td>T11_P11.9.B</td>
<td>0.418</td>
<td>5.807</td>
<td>1.087</td>
</tr>
<tr>
<td>T11_P17.10.B</td>
<td>0.400</td>
<td>6.044</td>
<td>1.989</td>
</tr>
<tr>
<td>T11_P16.3.B</td>
<td>0.345</td>
<td>5.998</td>
<td>1.206</td>
</tr>
<tr>
<td>T11_P12.5.B</td>
<td>0.345</td>
<td>6.032</td>
<td>1.183</td>
</tr>
<tr>
<td>T11_P12.3.B</td>
<td>0.345</td>
<td>6.526</td>
<td>1.222</td>
</tr>
<tr>
<td>T11_P11.1.B</td>
<td>0.345</td>
<td>6.237</td>
<td>1.126</td>
</tr>
<tr>
<td>T11_P03.7.B</td>
<td>0.345</td>
<td>6.179</td>
<td>1.741</td>
</tr>
<tr>
<td>T11_P14.5.B</td>
<td>0.327</td>
<td>7.522</td>
<td>1.685</td>
</tr>
</tbody>
</table>

HADDOCK best

Target

Total #submissions: 190
33 acceptables 11 medium 0 high
from 10 groups

Bressanelli et al. EMBO 2004

AB/6-05
**Target 12**

- "Good" data... "bad" results...!
- Dockerin has an internal C2 symmetry: native-like symmetrical solutions with 13% native contacts were submitted...
- A native solution with 22% native contact and i-RMSD of 2.66Å was not selected...

<table>
<thead>
<tr>
<th>Target</th>
<th>l-RMSD</th>
<th>i-RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>T12_P18.9.B</td>
<td>0.927</td>
<td>3.218</td>
</tr>
<tr>
<td>T12_P25.2.B</td>
<td>0.909</td>
<td>2.674</td>
</tr>
<tr>
<td>T12_P25.10.B</td>
<td>0.909</td>
<td>1.753</td>
</tr>
<tr>
<td>T12_P21.3.B</td>
<td>0.891</td>
<td>1.130</td>
</tr>
<tr>
<td>T12_P13.7.B</td>
<td>0.891</td>
<td>1.689</td>
</tr>
<tr>
<td>T12_P12.1.B</td>
<td>0.873</td>
<td>0.985</td>
</tr>
<tr>
<td>T12_P11.9.B</td>
<td>0.873</td>
<td>0.560</td>
</tr>
</tbody>
</table>

Total #Submissions: 214
16 acceptables 0 medium 21 high

**Target 13 ***

- Antibody-antigen complex (bound/unbound model)
- Good data (known epitopes) good results

<table>
<thead>
<tr>
<th>Target</th>
<th>l-RMSD</th>
<th>i-RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>T13_P25.1.HL</td>
<td>0.871</td>
<td>2.568</td>
</tr>
<tr>
<td>T13_P17.1.HL</td>
<td>0.800</td>
<td>3.782</td>
</tr>
<tr>
<td>T13_P02.5.HL</td>
<td>0.757</td>
<td>5.620</td>
</tr>
<tr>
<td>T13_P02.9.HL</td>
<td>0.714</td>
<td>5.307</td>
</tr>
<tr>
<td>T13_P10.10.HL</td>
<td>0.671</td>
<td>3.480</td>
</tr>
<tr>
<td>T13_P17.2.HL</td>
<td>0.614</td>
<td>10.041</td>
</tr>
<tr>
<td>T13_P19.2.HL</td>
<td>0.600</td>
<td>3.338</td>
</tr>
<tr>
<td>T13_P19.6.HL</td>
<td>0.557</td>
<td>8.216</td>
</tr>
<tr>
<td>T13_P21.9.HL</td>
<td>0.543</td>
<td>7.103</td>
</tr>
<tr>
<td>T13_P19.1.HL</td>
<td>0.486</td>
<td>7.075</td>
</tr>
</tbody>
</table>

Total #Submissions: 210
9 acceptables 6 medium 6 high

**Target 14 ***

- Complex between Protein Phosphatase 1-beta (PP1, unbound homology model) and the Myosin Phosphatase Targeting Subunit 1 (bound MYPT1)

<table>
<thead>
<tr>
<th>Target</th>
<th>l-RMSD</th>
<th>i-RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>T14_P94.1.B</td>
<td>0.611</td>
<td>0.581</td>
</tr>
<tr>
<td>T14_P75.9.B</td>
<td>0.611</td>
<td>0.929</td>
</tr>
<tr>
<td>T14_P75.10.B</td>
<td>0.611</td>
<td>0.929</td>
</tr>
<tr>
<td>T14_P75.1.B</td>
<td>0.611</td>
<td>0.929</td>
</tr>
<tr>
<td>T14_P82.2.B</td>
<td>0.599</td>
<td>1.246</td>
</tr>
<tr>
<td>T14_P84.1.B</td>
<td>0.586</td>
<td>2.340</td>
</tr>
<tr>
<td>T14_P82.1.B</td>
<td>0.586</td>
<td>1.648</td>
</tr>
<tr>
<td>T14_P75.2.B</td>
<td>0.580</td>
<td>1.126</td>
</tr>
<tr>
<td>T14_P75.3.B</td>
<td>0.573</td>
<td>1.627</td>
</tr>
</tbody>
</table>

Total #Submissions: 250
35 acceptables 22 medium 17 high
from 15 groups
**Target 20**

- Complex X (unbound-homology model)

*Reasonably good data ... BUT large conformational change*

<table>
<thead>
<tr>
<th>Target</th>
<th>fr-native 1-RMSD</th>
<th>i-RMSD</th>
<th>Total #submissions: 266</th>
</tr>
</thead>
<tbody>
<tr>
<td>T20_P50.8.A</td>
<td>0.360</td>
<td>12.092</td>
<td>3.129</td>
</tr>
<tr>
<td>T20_P50.4.A</td>
<td>0.342</td>
<td>7.748</td>
<td>2.340</td>
</tr>
<tr>
<td>T20_P59.1.A</td>
<td>0.298</td>
<td>21.196</td>
<td>4.830</td>
</tr>
<tr>
<td>T20_P50.7.A</td>
<td>0.289</td>
<td>23.313</td>
<td>4.628</td>
</tr>
<tr>
<td>T20_P28.9.A</td>
<td>0.289</td>
<td>41.023</td>
<td>12.723</td>
</tr>
<tr>
<td>T20_P28.7.A</td>
<td>0.228</td>
<td>22.737</td>
<td>6.015</td>
</tr>
<tr>
<td>T20_P28.5.A</td>
<td>0.228</td>
<td>22.060</td>
<td>6.125</td>
</tr>
<tr>
<td>T20_P28.1.A</td>
<td>0.199</td>
<td>19.068</td>
<td>5.149</td>
</tr>
<tr>
<td>T20_P28.8.A</td>
<td>0.219</td>
<td>19.683</td>
<td>4.623</td>
</tr>
<tr>
<td>T20_P28.4.A</td>
<td>0.158</td>
<td>26.337</td>
<td>5.190</td>
</tr>
<tr>
<td>T20_P28.2.A</td>
<td>0.184</td>
<td>26.337</td>
<td>5.190</td>
</tr>
</tbody>
</table>

No acceptable solution, but our best solution has > 60 and 70% of the correct epitope residues at the interface.

**Target 21**

- complex between Orc1 and Sir1 domains from yeast (unbound-unbound)

*Reasonably good data ... Good results*

<table>
<thead>
<tr>
<th>Target</th>
<th>fr-native 1-RMSD</th>
<th>i-RMSD</th>
<th>Total #submissions: 318</th>
</tr>
</thead>
<tbody>
<tr>
<td>T21_P34.M05</td>
<td>0.634</td>
<td>4.407</td>
<td>1.577</td>
</tr>
<tr>
<td>T21_P65.M07</td>
<td>0.585</td>
<td>7.593</td>
<td>2.409</td>
</tr>
<tr>
<td>T21_P65.M06</td>
<td>0.561</td>
<td>6.096</td>
<td>1.995</td>
</tr>
<tr>
<td>T21_P65.M08</td>
<td>0.537</td>
<td>5.262</td>
<td>1.941</td>
</tr>
<tr>
<td>T21_P59.M03</td>
<td>0.463</td>
<td>7.617</td>
<td>1.921</td>
</tr>
<tr>
<td>T21_P59.M04</td>
<td>0.390</td>
<td>6.367</td>
<td>2.798</td>
</tr>
<tr>
<td>T21_P59.M02</td>
<td>0.342</td>
<td>6.748</td>
<td>2.499</td>
</tr>
<tr>
<td>T21_P46.M03</td>
<td>0.342</td>
<td>25.462</td>
<td>5.124</td>
</tr>
<tr>
<td>T21_P33.M09</td>
<td>0.342</td>
<td>9.097</td>
<td>3.049</td>
</tr>
</tbody>
</table>

HADDOCK best vs target (1ZHI)

**Outline**

- Introduction
- Data-driven docking
- NMR-based HADDOCK application examples
- HADDOCK’s adventures in CAPRI
- Conclusions & perspectives
Conclusions & perspectives

• Data-driven docking is useful to generate models of biomolecular complexes, even when little information is available as demonstrated in CAPRI (2 high and 4 medium solutions out of 10 submitted targets).

• While such models may not be fully accurate, they 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!

• Data-driven docking is complementary to classical structural methods.

Acknowledgements

H ADDOC K
• Cyril Dominguez
• Aalt-Jan van Dijk
• Rolf Boelens
• Alexandre Bonvin

Ubiquitin dimer
• Aalt-Jan van Dijk (Utrecht)
• David Fushman (U. Maryland)

N ot 4-UbcH5B
• Sebastian Winkler
• Thomas Albert
• Marc Timmers
• Cyril Dominguez
• Hiroyuki Hanzawa
• Alexandre Bonvin
• Rolf Boelens

The CAPRI team:
• Aalt-Jan van Dijk
• Sjoerd de Vries
• Cyril Dominguez
• Alexandre Bonvin

Not4-UbcH5B
• Florida State University:
• Huan-Xiang Zhou
• Huling Chen

The End.

Thank you for your attention!

HADDOK online: http://www.nmr.chem.uu.nl/haddock

Dominguez, Boelens & Bonvin (2003), JACS 125, 1731