Modelling of biomolecular complexes by data-driven docking

Modelleren van biomoleculaire complexen door middel van data-gestuurde dokking

(met een samenvatting in het Nederlands)

Proefschrift ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. W.H. Gispen, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 16 oktober 2006 des middags te 14.30 uur

Door

Aalt Dirk Jan van Dijk

Geboren op 7 oktober 1979
te Utrecht, Nederland

A.D.J. van Dijk "Modelling of biomolecular complexes by data-driven docking"

Promotor: Prof. Dr. R. Boelens

Co-promotor: Dr. A.M.J.J. Bonvin

Beiden zijn verbonden aan het Bijvoet Centrum voor Biomoleculair Onderzoek,
Faculteit Betawetenschappen van de Universiteit Utrecht

Table of contents

Chapter 1.	General introduction: structural perspectives on protein-protein interactions.	5
Chapter 2.	Data-driven docking for the study of biomolecular complexes.	13
Chapter 3.	Various strategies of using Residual Dipolar Couplings in NMR-driven protein docking: application to Lys48-linked di-ubiquitin and validation against ^{15}N-relaxation data.	37
Chapter 4.	Combining NMR relaxation with chemical shift perturbation data to drive protein-protein docking.	59
Chapter 5.	Solvated docking: introducing water into the modelling of biomolecular complexes.	69
Chapter 6.	Data-driven docking: HADDOCK's adventures in CAPRI.	83
Chapter 7.	Modelling protein-protein complexes along the cytochrome c oxidase Cu-delivery pathway.	93
Appendix		109
Summary		129
Samenvatting		131
Dankwoord		133
List of publications		135

*Note: color figures are printed in the Appendix; in the text, these are marked with * before the figure caption. The Appendix also contains supporting information for chapter 3 and 5.*

A.D.J. van Dijk "Modelling of biomolecular complexes by data-driven docking"

Ontwerp omslag: Hanneke van Dijk-du Mortier en AV-dienst Departement Scheikunde, faculteit Betawetenschappen, Universiteit Utrecht

ISBN-10: 90-393-4354-3
ISBN-13: 978-90-393-4354-8

Chapter 1.
General introduction: structural perspectives on protein-protein interactions.

Interactome mapping

Proteins are the key players in virtually all biological processes. Since their amino acid sequence is encoded by DNA, deciphering of the genes which contain this DNA has been a central focus in the field of molecular biology. The combined efforts of molecular biologists culminated in the first drafts of the human genome in 2001 [1,2], almost 25 years after the first sequencing of a DNA molecule (of bacteriophage φX174) [3]. In the meantime, several hundreds of genomes have been sequenced [4]. An essential step in order to understand this plethora of data is to map the interactions between the proteins that are encoded by these sequences, since it is by interacting that those proteins fulfill their functions. After early efforts to chart protein-protein interaction networks focused on small-scale biological processes, proteome-scale maps of *S. cerevisiae*, *C. elegans* and *D. melanogaster* soon followed [5]. Recently, initial attempts to map the interactions within the human proteome appeared [6,7]. In Table 1 an overview of large scale protein-protein interaction mapping studies is provided.

The topic of this thesis is the modeling of the structures of protein-protein interactions. Determination of protein structures is motivated by the fact that a structure explains a multitude of biochemical and biophysical data. Correspondingly, structures of protein complexes are important for the understanding of protein-protein interaction data.

This introduction provides a background to this thesis by describing the broad picture of the "interactome", defined as the list of physical interactions mediated by all proteins of an organism. I will first discuss interactome mapping methods and applications of interactome data, followed by a discussion of the various methods used to determine or model structures of protein-protein complexes. At the end of this chapter we will turn towards 'interactome-oriented' approaches in structural biology. Note that chapter 2 contains a more specific introduction towards the computational technique used in the work described in this thesis, i.e. data-driven docking.

To date, most interaction maps have been obtained using the yeast two-hybrid system [8] (see Table 1). This is a genetic system where the interaction between two proteins is detected via the reconstruction of a transcription factor and the subsequent activation of reporter genes. A second approach which has been applied to characterize interactomes is affinity purification followed by mass spectrometric identification [9]. Other approaches exist [10], including fluorescence-based technologies [11], which potentially can be used for interactome mapping. Computational methods such as the analysis of domain fusion [12], conserved gene clusters [13] or phylogenetic profiling [14] are also used widely. Data obtained in experiments using these approaches have been integrated into a number of databases; Table 2 gives an overview of those.

Comparison of different datasets indicates only a small amount of overlap, and comparison to high-confidence interactions based on small-scale studies from literature indicates that interactions based on a single method have accuracy (percentage of predictions that is correct) and coverage (percentage of true interactions that is predicted) of around 10% or less [15] (note that these numbers are lower limits since the reference set is incomplete). Interactions supported by multiple methods have higher accuracy when compared to reference sets (up to ~ 50%) but coverage inevitably is lower. It is fair to say that these interactome datasets are first drafts which need to be further refined in the future. Indeed, in more recent papers presenting interactome data it is common to include quality indicators in order to assess the validity of the interactions and also to present a 'core' set of high-quality interactions. Quality can be assessed using for example GO annotations [16] (these are identifiers of cellular component, biological process and molecular function), based on the assumption that for interacting proteins these annotations will be most likely similar, or RNA

expression profiles [17], where the assumption is that functionally related genes tend to be expressed in a concerted fashion.

Table 1. Large scale interactome mapping studies.

Organism (#ORFs)[a]/Reference	# proteins	#interactions	Method[b]
Kaposi sarcoma-associated herpesvirus (89)			
[18]	50	123	Y2H
Varicella-roster herpesvirus (69)			
[18]	75	173	Y2H
Human (22000)			
[19]	N/A	25464	Literature mining
[20]	7748	31609	Literature mining
[6]	1549	2754	Y2H
[7]	1705	3186	Y2H
C. elegans (19099)			
[21]	2254	18183	Predictions
[22]	1541	2135	Y2H
S. cerevisiae (6200)			
[23]	3289	11334	Literature mining
[24]	2708	7123	AP/MS
[25]	2760	491	AP/MS
[26]	4681	34000	Literature mining
[27]	3617	1578	AP/MS
[28]	1167	232	AP/MS
[29]	1548	2358	Literature mining
[30]	1004	957	Y2H
D. melanogaster (13600)			
[31]	1711	2185	Y2H
[32]	488	1814	Y2H
[33]	7048	20405	Y2H
P. falciparum (5334)			
[34]	1415	1290	Predictions
[35]	1308	2846	Y2H
H. pylori (1590)			
[36]	741	1280	Y2H

[a]#ORFs: number of open reading frames.
[b]Y2H: yeast two hybrid; AP: affinity purification; MS: massa spectrometry.

Interactome data as described above can be analyzed at different levels. One level is the topology of the underlying network; for an overview of topological parameters of a few networks, see [37]. Concepts such as scale-free or small-world networks [38] can be related to general biological properties like robustness and adaptability [39,40]. Some network motifs are found much more often than would be expected randomly [41].

Secondly, interaction data can be useful in the annotation of the function of proteins [42,43]. It has been stated that 'network comparisons provide essential biological information beyond what is gleaned from the genome' [44]. An organism for which interactome-based function prediction was performed is *P. falciparum*, the causative agent of malaria [34].

Thirdly, and most generally, all cellular events like differentiation, metabolism etc. are essentially network processes. Interactome data provide testable hypotheses to explain and not

merely describe these events. In a recent review [45], biological questions that can be answered using computational strategies to analyze proteomic networks were discussed.

Table 2. Databases of protein-protein interactions.

Database	website	Organism	Datasource
BIND [46]	www.bind.ca	Various organisms	Large scale experiments
CYGD [47]	http://mips.gsf.de/genre/proj/yeast/	*S. cerevisiae*	Literature mining
DIMA [48]	http://mips.gsf.de/genre/proj/dima	Various organisms	Conserved domains
DIP [49]	http://dip.doe-mbi.ucla.edu/	Various organisms	Literature mining
HPRD [50]	www.hprd.org	Human	Literature mining
IntAct [51]	http://www.ebi.ac.uk/intact/	Various organisms	Literature mining
MINT [52]	http://mint.bio.uniroma2.it	Various organisms	Literature mining
MPact [53]	http://mips.gsf.de/genre/proj/mpact	*S. cerevisiae*	Literature mining
MPPI [54]	http://mips.gsf.de/proj/ppi/	Mammalian	Literature mining
OPHID [55]	http://ophid.utoronto.ca	Human	Predictions
Predictome [56]	http://predictome.bu.edu/	Various organisms	Predictions
STRING [57]	http://string.embl.de	Various organisms	Predictions

Experimental approaches towards structure determination of protein complexes

The two classical structure determination methods, X-ray crystallography and NMR are the main experimental approaches towards determination of protein-protein complexes [58-60]. However, both methods have inherent problems with the determination of structures of complexes. In crystallography the bottleneck is found in the crystallization process; while in NMR the most obvious problem is the protein size limit. Structural genomic projects [61], focusing on single chain structures, shifted the balance between monomeric and multimeric structures even more towards the sides of the singles, although efforts are underway to extend structural genomics towards protein - protein complexes [62,63].

New avenues towards structural characterization of (possibly large) macromolecular assemblies were paved when Electron Microscopy started to be applied [64,65]. This can also be combined with homology modeling, by fitting modeled structures into EM densities [66-68]. A variety of other methods can give information about interfaces in biomolecular complexes. The use of these in combination with computational docking (see below) will be described in chapter 2.

As a result of the application of various experimental methods, around three thousand structures of structurally different ‘interaction types’ (interactions with a different structural composition, comparable to a protein fold) have been solved [63]. Although this might sound impressive, this number represents only a small fraction of the whole interaction space. For example, using interaction data from different sources it was estimated that there are in total 10000 different interaction types [69]. Since about 1995, the number of interaction types has increased at the rate of roughly 200 – 300 per year, meaning that a full coverage of the interaction space would require another 25 years. Moreover, the number of 3000 solved structures mentioned above refers also to multidomain proteins and is only an upper bound for the real number of structures of protein complexes in the Protein Databank. In addition, other estimates for the number of interactions are higher than the number of 10.000 given above. For example, for yeast alone estimates range between 20000 and 30000 [15,70,71].

This clearly illustrates the need for computational approaches to obtain models for biomolecular complexes, since the number of interactions is huge and in addition for many transiently interacting systems it will be difficult if not impossible to obtain a structural model using traditional crystallography or NMR approaches.

Computational approaches towards structure determination of protein complexes

Computational approaches to characterize structures of complexes can be divided into two main classes: threading and docking. The former is comparable to threading approaches for single structure prediction. It scans the amino acid sequence of an unknown structure against a database of solved structures; a scoring function is used to assess the compatibility of the sequence to the structure. In the case of protein complexes, 'dimeric threading' is used: each target sequence is assigned to a monomeric template structure by threading; then, those templates that belong to the same physically interacting dimer template are selected [72,73].

Docking is a computational method which uses the known structures of the components of a complex to define the structure of the complex itself. Docking algorithms consist of a *sampling* stage, i.e. a way to generate structures of the complex, and a *scoring* method, i.e. a way to decide which of the generated structures are 'good'. The details of this will be discussed further in chapter 2.

Structural interactomics

Finally, I want to discuss examples of the usefulness of "structural interactomics" or "structure based systems biology" [74,75]. A straightforward approach is the large scale modelling of complexes. Aloy *et al.* modeled a large set of yeast complexes by using known structures of homologous complexes; the resulting models were validated by electron microscopy [76]. Multimeric threading has been applied to the yeast genome, resulting in over 7000 structural models for protein-protein interactions [72]. Also, ligand specificity can be predicted on a large scale dataset. For example, interaction sites were predicted on a proteome scale, using protein interaction data and the associated sequence information [77,78]. The information in a protein-protein interaction dataset was applied to discover sequence signatures; structural information was not used, but still many of the sequence signatures that were discovered appear in the surfaces of proteins [79], as would be expected if indeed these contain residues involved in interactions.

It is also interesting to note that after the initial success of structural genomics projects in elucidating protein structures, efforts are now underway to extend these approaches towards protein - protein complexes [62,63]. One example of this kind of approaches is the 3D Repertoire project (www.3drepertoire.org), which has the aim to resolve structures for all amenable protein complexes from budding yeast by experimental and computational approaches. Another example is the recently funded SPINE2 project (www.spineurope.org).

Another application might in principle be the modeling of predicted interactions in order to decide whether complex formation is indeed likely. However, for this to be performed one would have to reliably calculate the free energy, which is not possible at the moment.

Finally, an application of structural interactomics that is highly relevant to drug design is the search for inhibitors of specific protein-protein interactions [80-83]. Although this is a field that still is in its early stages, there is good promise that it will result in novel therapeutics that exploit the specific details of interacting surfaces [84].

Scope of thesis

It is clear from the discussion above that proteome-scale insight into structures of complexes is highly relevant, but experimental approaches alone are not sufficient to obtain this insight.

In addition, although advances have been made, computational approaches to generate models for protein-protein complexes are also far from perfect. Therefore, in this thesis we focus on data-driven docking, a computational approach that uses any known information on protein interfaces in order to generate models for protein-protein complexes. In chapter 2, the use of experimental information in docking approaches is reviewed, and the data-driven docking method HADDOCK is discussed. In the next two chapters, we give specific examples of the use of NMR data in docking: chemical shift perturbation data combined with residual dipolar couplings (chapter 3) or relaxation data (chapter 4). In chapter 5, we focus on the docking methodology: specifically, the possible structural role of water at protein-protein interfaces and its implications for docking are discussed. The last two chapters contain validation and application of our docking method. In chapter 6, we discuss HADDOCK's adventures in the blind docking experiment CAPRI. Finally, chapter 7 describes the application of data-driven docking towards elucidation of the structures of complexes along the cytochrome c oxidase Cu-delivery pathway.

References

1. Venter JC, Adams MD *et al.* The sequence of the human genome. Science (2001), **291**, 1304-+.
2. Lander ES, Linton LM *et al.* Initial sequencing and analysis of the human genome. Nature (2001), **409**, 860-921.
3. Sanger F, Air GM *et al.* Nucleotide sequence of bacteriophage [phi]X174 DNA. (1977), **265**, 695.
4. Benson DA, Karsch-Mizrachi I *et al.* GenBank. Nucleic Acids Res (2006), **34**, D16-D20.
5. Vidal M. Interactome modeling. Febs Lett (2005), **579**, 1834-1838.
6. Rual JF, Venkatesan K *et al.* Towards a proteome-scale map of the human protein-protein interaction network. Nature (2005), **437**, 1173-1178.
7. Stelzl U, Worm U *et al.* A human protein-protein interaction network: A resource for annotating the proteome. Cell (2005), **122**, 957-968.
8. Fields S, Song OK. A Novel Genetic System to Detect Protein Protein Interactions. Nature (1989), **340**, 245-246.
9. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature (2003), **422**, 198-207.
10. Piehler J. New methodologies for measuring protein interactions in vivo and in vitro. Curr Opin Struc Biol (2005), **15**, 4-14.
11. Yan YL, Marriott G. Analysis of protein interactions using fluorescence technologies. Drug Discov Today (2004), **9**, S27-S31.
12. Marcotte EM, Pellegrini M *et al.* Detecting protein function and protein-protein interactions from genome sequences. Science (1999), **285**, 751-753.
13. Overbeek R, Fonstein M *et al.* The use of gene clusters to infer functional coupling. P Natl Acad Sci USA (1999), **96**, 2896-2901.
14. Pellegrini M, Marcotte EM *et al.* Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. P Natl Acad Sci USA (1999), **96**, 4285-4288.
15. von Mering C, Krause R *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. Nature (2002), **417**, 399-403.
16. Harris MA, Clark J *et al.* The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res (2004), **32**, D258-D261.
17. Deane CM, Salwinski L *et al.* Protein interactions - Two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics (2002), **1**, 349-356.
18. Uetz P, Dong YA *et al.* Herpesviral protein networks and their interaction with the human proteome. Science (2006), **311**, 239-242.
19. Gandhi TKB, Zhong J *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet (2006), **38**, 285-293.
20. Ramani AK, Bunescu RC *et al.* Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. Genome Biol (2005), **6**, -.
21. Zhong WW, Sternberg PW. Genome-wide prediction of C-elegans genetic interactions. Science (2006), **311**, 1481-1484.
22. Li SM, Armstrong CM *et al.* A map of the interactome network of the metazoan C-elegans. Science (2004), **303**, 540-543.
23. Reguly T, Breitkreutz A *et al.* Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. Journal of Biology (2006), **5**, 11.

24. Krogan NJ, Cagney G *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature (2006), **440**, 637-643.
25. Gavin AC, Aloy P *et al.* Proteome survey reveals modularity of the yeast cell machinery. Nature (2006), **440**, 631-636.
26. Lee I, Date SV *et al.* A probabilistic functional network of yeast genes. Science (2004), **306**, 1555-1558.
27. Ho Y, Gruhler A *et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature (2002), **415**, 180-183.
28. Gavin AC, Bosche M *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature (2002), **415**, 141-147.
29. Schwikowski B, Uetz P *et al.* A network of protein-protein interactions in yeast. Nat Biotechnol (2000), **18**, 1257-1261.
30. Uetz P, Giot L *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature (2000), **403**, 623-627.
31. Formstecher E, Aresta S *et al.* Protein interaction mapping: A Drosophila case study. Genome Res (2005), **15**, 376-384.
32. Stanyon CA, Liu GZ *et al.* A Drosophila protein-interaction map centered on cell-cycle regulators. Genome Biol (2004), **5**, -.
33. Giot L, Bader JS *et al.* A protein interaction map of Drosophila melanogaster. Science (2003), **302**, 1727-1736.
34. Date SV, Stoeckert CJ. Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale. Genome Res (2006), **16**, 542-549.
35. LaCount DJ, Vignali M *et al.* A protein interaction network of the malaria parasite Plasmodium falciparum. Nature (2005), **438**, 103-107.
36. Rain JC, Selig L *et al.* The protein-protein interaction map of Helicobacter pylori. Nature (2001), **409**, 211-215.
37. Suthram S, Sittler T *et al.* The Plasmodium protein network diverges from those of other eukaryotes. Nature (2005), **438**, 108-112.
38. Song CM, Havlin S *et al.* Self-similarity of complex networks. Nature (2005), **433**, 392-395.
39. Jeong H, Mason SP *et al.* Lethality and centrality in protein networks. Nature (2001), **411**, 41-42.
40. Li D, Li JQ *et al.* Protein interaction networks of Saccharomyces cerevisiae, Caenorhabditis elegans and Drosophila melanogaster: Large-scale organization and robustness. Proteomics (2006), **6**, 456-461.
41. Yeger-Lotem E, Sattath S *et al.* Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. P Natl Acad Sci USA (2004), **101**, 5934-5939.
42. Sun SW, Zhao Y *et al.* Faster and more accurate global protein function assignment from protein interaction networks using the MFGO algorithm. Febs Lett (2006), **580**, 1891-1896.
43. McDermott J, Bumgarner R *et al.* Functional annotation from predicted protein interaction networks. Bioinformatics (2005), **21**, 3217-3226.
44. Sharan R, Suthram S *et al.* Conserved patterns of protein interaction in multiple species. P Natl Acad Sci USA (2005), **102**, 1974-1979.
45. Janes KA, Lauffenburger DA. A biological approach to computational models of proteomic networks. Curr Opin Chem Biol (2006), **10**, 73-80.
46. Bader GD, Betel D *et al.* BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res (2003), **31**, 248-250.
47. Guldener U, Munsterkotter M *et al.* CYGD: the comprehensive Yeast Genome Database. Nucleic Acids Res (2005), **33**, D364-D368.
48. Pagel P, Oesterheld M *et al.* The DIMA web resource - exploring the protein domain network. Bioinformatics (2006), **22**, 997-998.
49. Salwinski L, Miller CS *et al.* The Database of Interacting Proteins: 2004 update. Nucleic Acids Res (2004), **32**, D449-D451.
50. Peri S, Navarro JD *et al.* Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res (2004), **32**, D497-D501.
51. Hermjakob H, Montecchi-Palazzi L *et al.* IntAct: an open source molecular interaction database. Nucleic Acids Res (2004), **32**, D452-D455.
52. Zanzoni A, Montecchi-Palazzi L *et al.* MINT: a Molecular INTeraction database. Febs Lett (2002), **513**, 135-140.
53. Guldener U, Munsterkotter M *et al.* MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res (2006), **34**, D436-D441.
54. Pagel P, Kovac S *et al.* The MIPS mammalian protein-protein interaction database. Bioinformatics (2005), **21**, 832-834.
55. Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics (2005), **21**, 2076-2082.

56. Mellor JC, Yanai I *et al.* Predictome: a database of putative functional links between proteins. Nucleic Acids Res (2002), **30**, 306-309.
57. von Mering C, Huynen M *et al.* STRING: a database of predicted functional associations between proteins. Nucleic Acids Res (2003), **31**, 258-261.
58. Zuiderweg ER. Mapping protein-protein interactions in solution by NMR spectroscopy. Biochemistry (2002), **41**, 1-7.
59. Dauter Z. Current state and prospects of macromolecular crystallography. Acta Crystallogr D (2006), **62**, 1-11.
60. Bonvin AMJJ, Boelens R *et al.* NMR analysis of protein interactions. Curr Opin Chem Biol (2005), **9**, 501-508.
61. Chandonia JM, Brenner SE. The impact of structural genomics: Expectations and outcomes. Science (2006), **311**, 347-351.
62. Strong M, Sawaya MR *et al.* Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis P Natl Acad Sci USA (2006), **103**, 8060-8065.
63. Bravo J, Aloy P. Target selection for complex structural genomics Curr Opin Struc Biol (2006), **16**, 385-392.
64. Frank J. Single-particle imaging of macromolecules by cryo-electron microscopy. Annu Rev Bioph Biom (2002), **31**, 303-319.
65. Orlova EV, Saibil HR. Structure determination of macromolecular assemblies by single-particle analysis of cryo-electron micrographs. Curr Opin Struc Biol (2004), **14**, 584-590.
66. Gao HX, Sengupta J *et al.* Study of the structural dynamics of the E-coli 70S ribosome using real-space refinement. Cell (2003), **113**, 789-801.
67. Topf M, Baker ML *et al.* Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. J Struct Biol (2005), **149**, 191-203.
68. Topf M, Sali A. Combining electron microscopy and comparative protein structure modeling. Curr Opin Struc Biol (2005), **15**, 578-585.
69. Aloy P, Russell RB. Ten thousand interactions for the molecular biologist. Nat Biotechnol (2004), **22**, 1317-1321.
70. Grigoriev A. On the number of protein-protein interactions in the yeast proteome. Nucleic Acids Res (2003), **31**, 4157-4161.
71. Bader GD, Hogue CWV. Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol (2002), **20**, 991-997.
72. Lu L, Arakaki AK *et al.* Multimeric threading-based prediction of protein-protein interactions on a genomic scale: Application to the Saccharomyces cerevisiae proteome. Genome Res (2003), **13**, 1146-1154.
73. Grimm V, Zhang Y *et al.* Benchmarking of dimeric threading and structure refinement. Proteins (2006), **63**, 457-465.
74. Aloy P, Russell RB. Structure-based systems biology: a zoom lens for the cell. Febs Lett (2005), **579**, 1854-1858.
75. Aloy P, Russell RB. Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Bio (2006), **7**, 188-197.
76. Aloy P, Bottcher B *et al.* Structure-based assembly of protein complexes in yeast. Science (2004), **303**, 2026-2029.
77. Li HQ, Li JY *et al.* Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. Bioinformatics (2006), **22**, 989-996.
78. Li HQ, Li JY. Discovery of stable and significant binding motif pairs from PDB complexes and protein interaction datasets. Bioinformatics (2005), **21**, 314-324.
79. Fang JW, Haasl RJ *et al.* Discover protein sequence signatures from protein-protein interaction data. Bmc Bioinformatics (2005), **6**, 277.
80. Vidal M, Brachmann RK *et al.* Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. P Natl Acad Sci USA (1996), **93**, 10315-10320.
81. Huang J, Schreiber SL. A yeast genetic system for selecting small molecule inhibitors of protein-protein interactions in nanodroplets. P Natl Acad Sci USA (1997), **94**, 13396-13401.
82. Golemis EA, Tew KD *et al.* Protein interaction targeted drug discovery: Evaluating critical issues. Biotechniques (2002), **32**, 636-+.
83. Arkin MR, Wells JA. Small-molecule inhibitors of protein-protein interactions: Progressing towards the dream. Nat Rev Drug Discov (2004), **3**, 301-317.
84. Fletcher S, Hamilton AD. Protein surface recognition and proteomimetics: mimics of protein surface structure and function. Curr Opin Chem Biol (2005), **9**, 632-638.

Chapter 2.
Data-driven docking for the study of biomolecular complexes.

Aalt D.J. van Dijk, Rolf Boelens and Alexandre M.J.J. Bonvin,
FEBS Journal **272** (2005), 293 – 312.

Abstract

With the amount of genetic information available, a lot of attention has focused on systems biology, in particular biomolecular interactions. Considering the huge number of such interactions, and their often weak and transient nature, conventional experimental methods such as X-ray crystallography and NMR spectroscopy are not sufficient to gain structural insight into those. A wealth of biochemical and/or biophysical data can, however, readily be obtained for biomolecular complexes. Combining these data with docking (the process of modeling the 3D structure of a complex from its known constituents) should provide valuable structural information and complement the classical structural methods. In this review we discuss and illustrate the various sources of data that can be used to map interactions and their combination with docking methods to generate structural models of the complexes. Finally a perspective on the future of this kind of approach is given.

Introduction

With the available amount of genetic information, a lot of attention is focused on systems biology. Here a central question is: how do the various biomolecular units work together to fulfill their tasks? In order to answer this question, structural information on complexes is needed. Biochemical and biophysical experiments are widely used to gain insight into biomolecular interactions. The information generated in this way can in principle be used to model the structure of the complex under study. Taking the step from data to modeling (docking) is however not common practice. Docking approaches allow to generate models of a biomolecular complex using as starting information the known structure of its constituents. Combining experimental data with docking makes sense considering that the number of single proteins, domains thereof or other biomolecules whose three dimensional (3D) structures have been solved is much larger than the number of solved structures of complexes and is steadily increasing as a result of the world wide structural genomics initiatives. The advantages of docking approaches upon conventional structural techniques are the speed and the possibility to study complexes that could only otherwise be studied with considerable effort (or not at all). One particular class of complexes for which this is the case are weak or transient, short-lived complexes; this is all the more interesting as these are often of the utmost biological importance. Other examples are the biologically highly relevant complexes of membrane or membrane-associated proteins, which are also notoriously difficult to study by NMR spectroscopy or X-ray crystallography.

Conventional crystallographic and NMR structural biology techniques have proven their value and will continue to do so. There are, however, problems associated with these techniques that are not likely to be completely overcome, especially when dealing with complexes. For crystallography, the main bottleneck is the crystallization, which can be a daunting task. For NMR, large complexes cause severe line broadening, which, at present, sets the upper limit for NMR to molecular sizes below 100 kDa. Moreover, to solve a structure by NMR in a conventional way, complete chemical shifts assignment and collection of structural restraints such as NOEs are challenging tasks, especially for large systems like complexes.

In this review, we wish to highlight the use of biochemical and biophysical data in docking approaches not only because of the general interest in docking as explained above, but also because it is still rather common practice to experimentally map interfaces without taking the next step to generate a structural model of the complex. We review only part of the docking field, namely the approaches relying on the use of additional biochemical and/or biophysical data. Generally, docking approaches that do not use any kind of experimental data

have difficulties in generating consistently reliable structures of complexes. Nevertheless, clear progress has been achieved in the field of '*ab-initio* docking', as reviewed in [1-4], and illustrated by the CAPRI (Critical Assessment of PRedicted Interactions) experiment [5], a "blind" docking competition in which participants have a limited time to predict the structure of a complex given only the structures of the constituents. Our discussion will be limited to biomolecular complexes, without reviewing protein – small ligand complexes; however, much of what is presented here will also be valid for that class of complexes. For a review on 'guided docking' to study protein-ligand complexes, see reference [6].

The review is organized as follows. We will first discuss the various kinds of biochemical and biophysical data that can be combined with docking. For each of those, examples will be given, and their strengths and weaknesses for use in docking will be discussed. We will then describe the basics of current docking methodologies and highlight our newly developed data-driven docking method HADDOCK [7]. We will end with conclusions and give a broader perspective on what could be the future of data-supported docking.

Sources of experimental data to define interfaces

Data from biochemical and/or biophysical experiments that provide information on residues located at the interface of a complex are potential sources to be used in docking. Critical issues are the level of detail that can be obtained (e.g. is the information residue-specific or not?) and the reliability of the data. Here we discuss, with those issues in mind, the techniques that have been used to obtain interface information for docking. In Figure 1 we present an overview of the most common methods. For a selected set of examples, we will also discuss how these data relate to the experimental high-resolution structure solved by conventional methods (see Table 4). Other experimental methods such as small angle X-ray scattering (SAXS) or electron microscopy and tomography can also provide valuable information about the 'shape' and organization of biomolecular complexes. As these are a rather different kind of approaches, we will not review them here, but only briefly mention their potential in our conclusions and perspectives. A general review about structural perspectives on protein-protein interactions can be found in reference [8].

Mutagenesis

When using mutagenesis to derive information for docking, one considers as candidates only the residues that are on the surface of the partner proteins. The general idea then is that mutation of an interface residue will influence the interaction, whereas for non-interface residues the mutation will have no effect. A variety of methods can be used to find out whether complex formation is affected by mutations, such as Surface Plasmon Resonance (SPR) [9], Mass Spectrometry (MS), yeast two-hybrid systems [10] and phage display libraries [11]. Target residues for mutagenesis can be selected based upon knowledge such as conservation (see below), but it is also possible to do a in-depth systematic scanning as in alanine scanning mutagenesis studies [12, 13]. An online database with results from alanine scanning mutagenesis data has been established, called ASEdb (http://www.asedb.org) [14]. These methods indicate which residues are in the interface, but do not give information about the contacts that are made across the interface. More detailed information can be obtained using so-called double mutant cycles [15]. Here one creates a series of mutants for both proteins. By measuring the K_d values for combinations of mutants, one can assess whether the influence of mutation X in protein A on the complex formation depends on mutation Y in protein B. If this is the case, the mutations are coupled, and one infers that the residues are close in space, i.e. that they are in contact or close proximity across the interface.

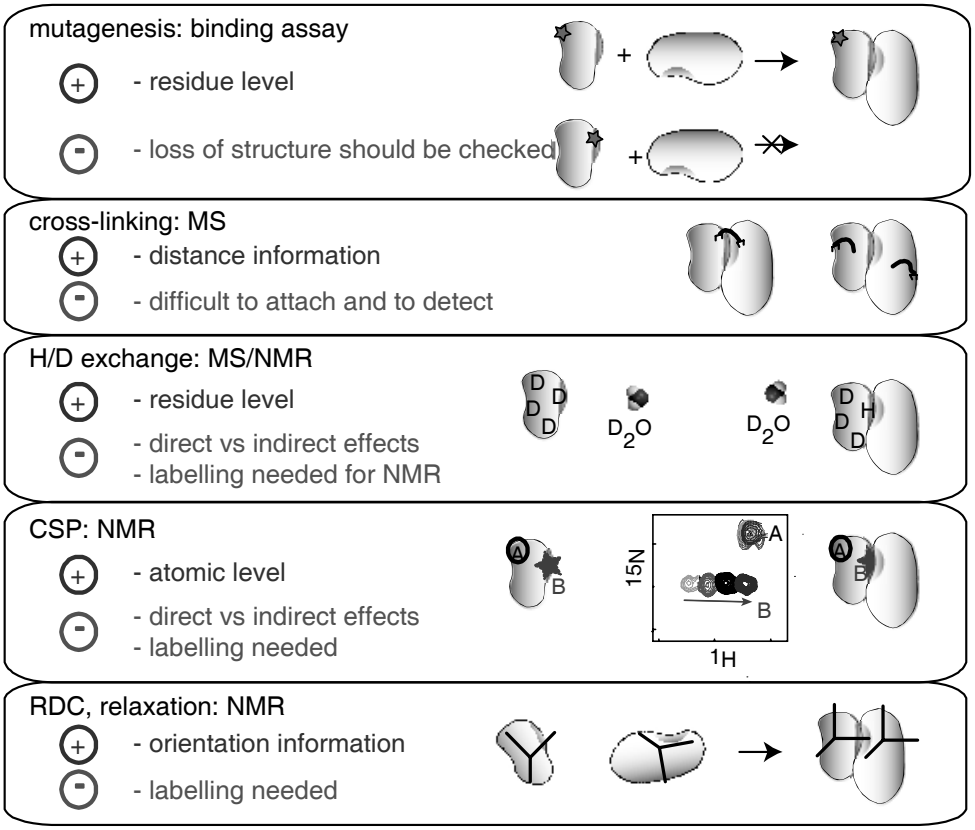


***Figure 1.** Illustration of the various data sources used in combination with docking. *Left*: advantages (+) and disadvantages (-); *right*: pictorial representation of the data source: the green and red shapes represent the two components of the complex. Mutagenesis: the blue star indicates a mutated residue; cross-linking: the black line indicates a cross-link; H/D exchange: 'D' and 'H' indicate residues where exchange can and cannot take place, respectively; CSP (chemical shift perturbation): HSQC spectrum showing one peak that does not shift and one peak that shifts on complex formation (the corresponding residues are indicated on the protein shapes); RDC, relaxation: the axis system indicates the tensor which provides orientational information.

A general warrant when using mutagenesis data is that it is unsound to assume that residues for which no effect is seen upon mutation, do not participate in an important interaction, unless it can be demonstrated that water, or nearby side chains, do not effectively substitute for the deleted atoms [13]. Another point is that one should, in principle, always check whether the mutants do not affect the 3D structure of the free components themselves, i.e. whether the native structures are preserved or not. Mutagenesis approaches, when carried out extensively, are able to generate a rather detailed map of the interface of a biomolecular complex. In Table 1 we give an overview of complexes for which mutagenesis data have been used in docking.

Mass spectrometry

There has been increasing interest in mass spectrometry as a tool in structural biology in general, and also specifically to obtain information about biomolecular complexes [16, 17]. One approach that can be used is H/D exchange. Here the rate of exchange gives information about the accessibility of the residue in question; rate differences between free and bound

forms indicate that a given residue is protected upon complex formation and thus probably involved in the interaction [18, 19].

Table 1. Examples of complexes docked using mutagenesis data.[a]

Complex	Information used	Ref.
Mutagenesis		
- FAK FAT domain – paxillin LD2 peptide	GST domain fusion	[89]
- TF/fVIIa/fXa	Charge altering mutations	[152]
- $R_{II\alpha}$ – C_α subunits of PKA	Neutron scattering, mutagenesis	[110]
- SDF-1α – heparin	SPR	[153]
- RCC1 – Ran	SPR	[51]
- Glycophorin A dimer	[b]	[45]
- Phospholamban pentamer	[b]	[44, 154, 155]
- Staphylokinase – microplasmin	Phage display	[156]
- Gα - G$\beta\gamma$ - receptor	G protein activation assay	[157]
- 30S ribosomal subunit – colicin E3	immunoblotting	[70, 71]
- EmrE dimer	Cysteine mutagenesis, cross-linking	[78]
- Hsc70 - auxilin	Rescue-mutant pair, CSP	[158]
- Kv1.3 K+ channel – six different scorpion toxins	comparison of electrostatic energy with binding affinity	[63]
- Integrin αIib TM domain homodimer	CAT-ELISA	[47]
- C1q – C-reactive protein / IgG	[b]	[49]
- antibody fragment - □ bungarotoxin	CDR on antibody; epitope mapping	[159]
- Malonyl-CoA – COT/CPT	Enzyme activity assay, immunoblotting	[160]
- gp120 – CD4	[b]	[7]
- Protein-DNA complexes of 434 cro and lac headpiece	Ethylation interference	[34]
- LexA DBD – DNA	Ethylation interference	[72]
- LexA – DNA	cross-linking	[161]
- Repressor – protein - DNA	DNA footprinting	[162]
- Fis - DNA	Chemical interference, nuclease DNA cleavage site	[163]
- EnvZ dimer	Cysteine substitutions and disulfide cross-linking detection	[164]
- subunit c oligomer of H^+-transporting ATP synthase	Cysteine substitutions and disulfide cross-linking detection	[165]
- Yeast cofactor A - β-tubulin	Two-hybrid assay	[166]
- FOG-$ZF3_{KRA}$ – TACC3	Two-hybrid assay; NMR CSP	[90]
Double mutant cycles		
- BgK – Kv1.1	Electrophysiological experiments, dose-response curve	[74]
- Agitoxin – *shaker* K^+ channel	[b]	[75]
- IFN-alpha2 – ifnar2	Reflectometric interference spectroscopy	[77]
- Alpha-cobratoxin – alpha7 receptor	Binding competition	[76]

[a] GST, glutathione S-transferase; SPR, Surface Plasmon Resonance; CSP, Chemical Shift Perturbation.
[b] Data were taken from the literature without giving any experimental details.

Another possibility is cross-linking, where residues close in space are detected by first covalently linking two molecules by the use of a cross-linking reagent, and then subjecting the resulting material to peptide mass fingerprinting or other protein identification methods [20]. Although these methods are promising, the cross-linking reaction is problematic and the information is often not easy to interpret. The detection of cross-linked residues is especially non trivial. To date MS data have not often yet been combined with docking approaches (Table 2).

Table 2. Examples of complexes docked using Mass Spectrometry data.

Complex	Information used	Ref.
- Calmodulin – melittin	cross-linking	[85]
- Aminoacylase-1 dimer	proteolysis, cross-linking	[111]
- PKA – C and R subunit	H/D exchange	[50]
- C1r $(\gamma\text{-B})_2$	cross-linking	[167]
- IL-6 homodimer	cross-linking	[112]

NMR

Conventional NMR methods have been used for more than a decade to study biomolecular complexes. In the classical approach, one first has to perform a resonance assignment that is as complete as possible, and then collect structural restraints such as NOEs, which can be detected between protons that are close in space (<5Å), and residual dipolar couplings that provide orientational information. Using such restraints, one can accurately define the structure of a biomolecule or a biomolecular complex. In addition to its conventional use in structure determination, NMR is very well suited to map interfaces of biomolecular complexes with so called chemical shift perturbation (CSP) experiments [21]. Here, easily obtainable heteronuclear single quantum coherences (HSQC) spectra of one (^{15}N-labeled) partner in the complex are recorded in the absence and presence of increasing amounts of the partner protein ("titration experiments"). Changes in chemical shifts of one molecule on addition of a second molecule allow assessment of which residues of the labeled molecule are perturbed by the formation of the complex. One then repeats this procedure with the second molecule labeled. Under the assumption that the perturbed residues correspond to the interacting residues, a detailed map of the interface is obtained.

Two other NMR techniques that are able to give similar information are H/D-exchange and cross-saturation or saturation transfer (SAT) [22]. As in MS, NMR can also easily be used to perform H/D exchange experiments; again, differences in exchange rates when comparing uncomplexed and complexed forms point to protected residues that are assumed to be at the interface. In cross-saturation experiments, the observed protein is perdeuterated and ^{15}N-labeled with its amide deuterons exchanged back to protons, while the other "donating" partner protein is unlabeled. Saturation of the unlabeled protein leads by cross-relaxation mechanisms to signal attenuation (again typically monitored by ^{15}N-HSQC spectra) of those residues in the labeled protein that are at close proximity. The labeling scheme can be reversed to map the other interface. Deuteration is a requisite here. Cross-saturation experiments are believed to give a more reliable picture of the interface than CSP data which can suffer from "false positives" because of conformational changes.

Other relatively easily obtainable NMR parameters are residual dipolar couplings (RDCs) [23]. These provide information about the orientation of the components with respect to each other, and can be used in addition to CSP data in docking approaches. Comparable information can be extracted from relaxation experiments in the case of diffusion anisotropy [24].

Table 3. Examples of complexes docked using NMR data.[a]

Complex	Information used	Ref.
Protein – protein		
- Cyt c – cyt f	CSP	[56]
- Cyt c – cyt c peroxidase	CSP	[54]
- Plastocyanin – cyt f	PC, CSP	[80, 81]
- myoglobin – cyt b5	CSP, ^{15}N relaxation	[57]
- Ubiquitin – YUH1	CSP	[38]
- Ubiquitin – hHR23A UBA1, UBA2	CSP	[93]
- hHR23a (four linked domains)	CSP, RDC	[168]
- Ubiquitin – p47 UBA domain	CSP	[96]
- di-ubiquitin	CSP, RDC	[169, 170]
- UbcH5B – CNOT4	CSP, mutagenesis	[88]
- mms2 – ubc13 – ubiquitin – ubiquitin	CSP	[59]
- EIN-HPr[b], IIA(Glc)-HPr[b], IIA(Mtl)-HPr[b]	CSP, RDC	[84]
- Bem1 PB1- Cdc24 PB1	CSP, mutagenesis	[95]
- RPA70A – Rad51N	CSP, mutagenesis	[94]
- CAD-ICAD[b]	SAT, RDC	[82]
- EIN-HPr[b]	CSP, RDC	[67]
- EIN-HPr[b], E2A-HPr[b]	CSP	[7]
- Atx1 – Ccc2 domain	CSP	[92]
- HR1b – Rac1	CSP	[171]
- FcεRIα - IgE Cε2	CSP	[172]
- FcεRI – peptide	CSP, mutagenesis, NOE	[66]
- LpxA – acyl carrier protein	CSP, RDC, mutagenesis	[91]
Protein – carbohydrates		
- Tri,hexa saccharide – antibody	SAT	[173]
- (glycosylated) PDTRP – antibody SM3	SAT	[174]
- fibronectin (13,14)F3 – heparin	CSP	[62]
Protein – nucleic acids		
- NS1A(1-73) – 16bp dsRNA	CSP	[40]
- UvrC CTD – junction DNA	CSP	[39]
- XPA-MBD – 9bp ssDNA	CSP	[175]
- Rom – RNA kissing hairpin	CSP	[41]
- Pf3 ssDBP – ssDNA	CSP	[83]
- CylR2 – 22bp DNA	CSP	[73]

[a] CSP: chemical shift perturbation; PC: pseudocontact shifts; SAT: saturation transfer.
[b] These complexes were also solved using the classical NOE-based approach.

An NMR parameter that can also be useful is the pseudocontact shift. It results from residual electron-nuclei dipolar interactions in molecules [21]. The use of paramagnetic tags attached to a protein can induce this phenomenon [25, 26]. As pseudocontact shifts contain long-range information, they can be very useful in docking approaches. It is also possible to use paramagnetic ions as probes, as they induce broadening of the NMR signals for the residues they contact. In a complex, the interface residues will be protected from such effects, allowing a reliable detection of the interface [27]. An overview of complexes for which NMR data have been used in docking approaches is given in Table 3.

Reliability issues

It should be clear that there is a wealth of experimental data, not all of them having been discussed here, that can be used to define interface residues. The question of the reliability of this information is of course very important. In Table 4 we give an overview of some complexes for which the experimental data have been compared explicitly with the (at that time available) corresponding 3D structures. In Figure 2, as an example, experimental data for the antibody D1.3 – antibody E5.2 complex is mapped onto the surfaces of the two proteins.

Table 4. Comparison of experimental information defining interfaces with the experimental X-ray or NMR structures[a]

Complex	Information used
Mutagenesis data	
- Barnase – barstar	DMC: coupling energy decreases as distance increases [176]
- antibody D1.3 – antibody E5.2	DMC: of 13 identified, 9 in interface and 4 not in interface showing significant coupling, but lower than the contacting residues [177]
- Cyt c – peroxidase	mutations: sites coincide with X-ray defined sites; DMC: couplings for residues that are more than 10Å apart, concluded to be due to small rearrangements [178]
- Cyt c2 – RC	DMC: coupling approximately inversely proportional to distance [179]
Mass Spectrometry data	
- DnaA domain 4 – DnaA box	cross linking data correctly locate interaction site to a six residue peptide fragment identified previously by X-ray/NMR [180]
- Ribosome	Comparison of > 2500 experimental distance restraints (cross-linking, footprinting and cleavage data) with X-ray structure showing good agreement [144]
NMR data	
- Lysozyme – antibody	H/D: of 15 perturbed: 5 on epitope, 5 at edge, 5 far away [181]
- OMTKY3 – Ctr	CSP fully consistent with X-ray [182]
- rNTF2 – FxFG-containing Nsp1-P30	high affinity X-ray site seen by NMR; NMR also finds low affinity site → NMR better able to identify weak interactions [183]
- Zf1-3 (TFIIIA) – 15 bp DNA	CSP data do not correspond exactly to the interface, but arise from a number of effects [184]
- CAD – ICAD	NOE and SAT defined interface is quite consistent with X-ray; CSP defined interface is a bit different [82]
- Nova1 – RNA	Cross-saturation defined residues match closely the X-ray interface; CSP data define the same residues and a few additional ones [185]
- RNAse E S1 homodimer	CSP used to assess validity of crystallography dimer; data match the contacting residues seen in the crystal [186]

[a] CSP: chemical shift perturbation; DMC: double mutant cycles; SAT: saturation transfer

Although these are only a few examples, the general trend indicates that the experimental sources discussed above provide quite reliable information on interface residues. Sometimes they can result from small rearrangements and secondary effects, but as long as these "false positives" are not too numerous, they can be dealt with in computational approaches (see below). If conformational changes are too large, however, docking approaches are probably bound to fail. It is not simple to predict a priori from the data if such effects should be

expected. Sometimes, clustering of predicted interface residues on the surface can give a good indication that the mapped interface is very likely to be correct one.

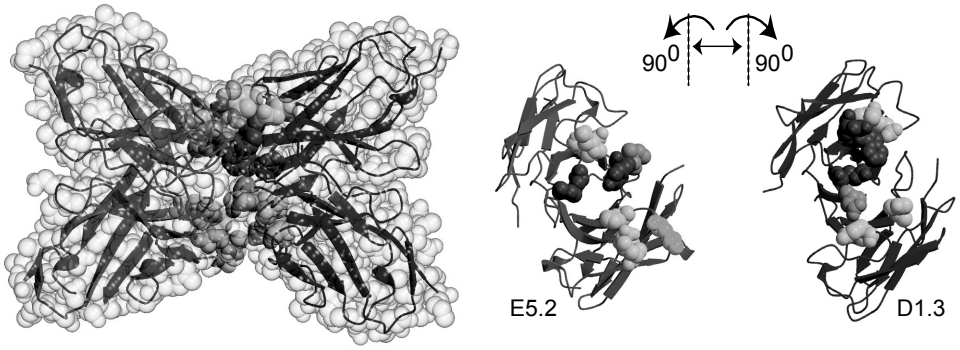


***Figure 2.** Mapping of the mutagenesis data [177] onto the structure of the antibody D1.3 – antibody E5.2 complex [187] (pdb entry 1dvf). *Left*: structure of the complex; *right*: interaction surface of E5.2 (left) and D1.3 (right) color coded according to the measured ΔΔG value [177] in mutagenesis experiments. Red: ΔΔG > 4.0 kcal/mol; orange: ΔΔG 2.1 – 4.0 kcal/mol; yellow: ΔΔG 1.1 – 2.0 kcal/mol; green: ΔΔG < 1.0 kcal/mol. Figures are prepared using Molscript [188] and Raster3D [189].

Computational docking approaches using experimental data

In the docking literature one often finds the distinction between 'bound' and 'unbound' docking: the former refers to docking using the structures of the single proteins as they are present in the complex, and the latter to docking using the structures of the free proteins. As only the latter is of biological relevance, here "docking" will refer to "unbound docking" (although in some cases a method is, as a first, easier step, tested in bound docking).

As defined in the introduction, docking methods generate a model of a complex based on the known 3D structures of its free components. To do this in a computer, two things are needed: a way to generate structures of the complex, i.e. a *sampling method*, and a way to decide which of the generated structures are 'good', i.e. a *scoring method*. The output typically consists of a large number of solutions, some of which get a high rank and are accordingly considered to correspond to the 'real' structure, whereas others get a lower rank and are discarded.

Docking methods vary in the way sampling and scoring are implemented, and also in the representation of the molecules in the calculations. An important choice to be made is whether the proteins are kept rigid or whether flexibility is needed. Flexibility can be introduced in various ways, e.g. by using an ensemble of rigid structures (experimental or generated for example by MD-methods) corresponding to static snapshots of possible conformational changes, by allowing some interpenetration of the docked molecules (sometimes called 'soft' rigid body docking, as opposed to 'hard' rigid body docking, where no overlap is allowed at all), or by allowing explicit side-chain and/or backbone flexibility during the docking. The type of *sampling* depends on the way in which the molecules are represented. When a grid representation of the molecules is used, rigid body docking can be done by calculating correlations (like e.g. surface complementarity) using Fast Fourier Transform methods [28-33]. When the protein is explicitly represented using an atomic model, one can use various sampling methods such as Monte Carlo [34-36] and molecular dynamics methods [7] or genetic algorithms [36] in combination with simulated annealing schemes. The *scoring* is typically based on some kind of force field [37], which assigns an energy to atom-atom (or residue-residue) pairs, and subsequently adds all those together to get

the energy of a given configuration. Often, terms such as buried surface area and desolvation energy are added. Force fields can have a physical basis or can be knowledge-based (derived by counting how often a given pair occurs in a database of experimental structures). Using biochemical and/or biophysical data in docking approaches has advantages both for the sampling and the scoring stages. During the sampling, more 'relevant' configurations are produced while in the scoring, the ranking of true-positives (i.e. correct solutions) can be improved compared to *ab-initio* docking where typically tens to hundreds of false positives are scored at the top. An important difference between various methods is whether the experimental data are only introduced in the scoring (i.e. to filter the solutions that have been generated), or whether they are also used during sampling. In the following we will discuss a number of methods that have been proposed, first the procedures that only use experimental data for scoring, and next those that incorporate experimental data in the sampling itself. In Figure 3 a graphical representation is given of the choices to make in the various docking approaches with respect to the incorporation of experimental data and the treatment of flexibility.

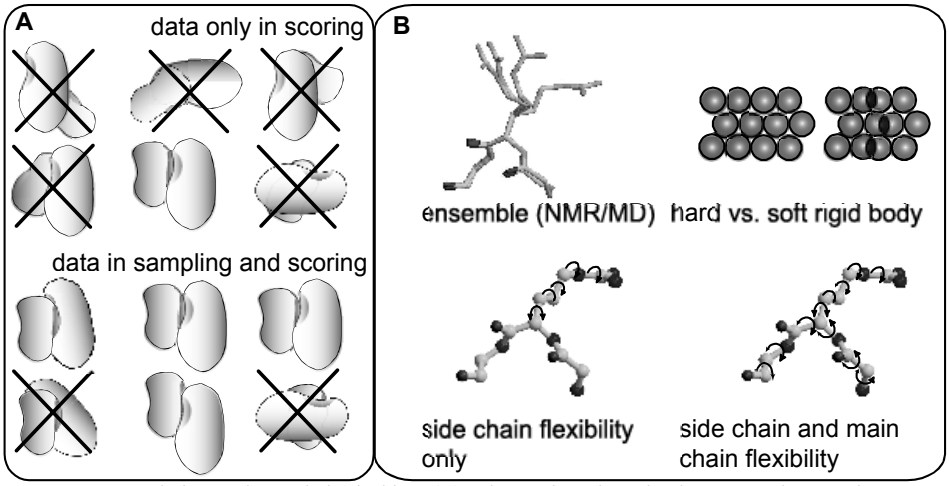


***Figure 3.** Some choices to be made in docking. (A) When to introduce the data? Here the complex structures resulting from a hypothetical docking method are shown, and the scoring is represented in a simplified way as discarding the complexes which do not satisfy the experimental restraints (indicated by the black crosses); (B) How to deal with flexibility: using an ensemble of starting structures; by soft rigid body docking; and explicitly during the docking by allowing side chain and/or main chain flexibility.

Although computer-based approaches should be preferred in terms of reproducibility, it is also possible to "manually" build models of complexes based on experimental information. There are quite a few examples where this has been done [38-42], some of which have been compared with pure *ab-initio* docking results [43]. We should note that each docking approach has its own advantages and disadvantages, and at this moment the 'docking problem' is still unsolved: no single docking method will always give the right answer. The docking field is still in active development and various approaches to the problem are pursued, as will be discussed below.

Docking methods using experimental data only in the scoring stage

A large variety of docking methods exists and have been used before applying a filter based on experimental data. One approach consists of a systematic grid search for all possible orientations (three translations and six rotations). This is only feasible for small systems and

simplified models, as otherwise scoring all possible configurations becomes intractable. Such a method has been used for probing transmembrane helix multimers, e.g. the dimeric transmembrane region of glycophorin A and the phospholamban pentamer. The low energy structures resulting from the grid search were filtered using mutagenesis data [44-47].

When studying larger systems, and especially if one wants to introduce sophisticated amounts of flexibility in the docking, exhaustive grid searches become unrealistic. A fast method to perform grid calculations based on spherical Fourier correlations is implemented in the program Hex [48]. It has been combined with mutagenesis data [49]. Fast Fourier Transform (FFT) methods have often been used in docking. For example, the docking program DOT [29] was used in combination with MS H/D data to filter solutions [50]. Other examples of FFT-based methods are the soft docking program GRAMM [30] which has been used in combination with mutagenesis data [51] and FTDOCK [28], which was originally tested on several complexes using experimental data (e.g. active site information in the case of enzyme-inhibitor complexes) and was recently combined with NMR data (CSP and RDCs) to filter solutions [52]. Another grid-approach, which uses Boolean type operations and was optimized heuristically for speed, is the docking program BiGGER [53]. This program allows for soft rigid body docking (in [54] hard and soft docking were compared). BiGGER is often used in combination with NMR CSP data [55-59].

There are several docking approaches that do not use a grid but rather an explicit search in the configurational space, e.g. DOCK [60, 61], Autodock [36], which was used in combination with CSP data [62], and other methods based on Brownian Dynamics simulations followed by Molecular Dynamic refinement of the initial models [63]. NMR CSP data have also been used in a more quantitative way for filtering docking solutions, by back-calculating chemical shift changes from the models with programs like Shifts [64] or Shiftx [65] and comparing them with the experimental values [66]. This approach has also been combined with RDCs [67]. The above methods have been successfully applied to model various biomolecular complexes (Tables 1-3).

Docking methods using experimental data to drive the docking

The advantage of using the data in the sampling stage of docking is that 'correct' or 'near-correct' configurations should be enriched, compared to approaches in which the data is only used in the scoring stage, provided of course that the experimental information is correct. This becomes especially important when the number of configurations is too large to be adequately sampled, as is often the case when flexibility is introduced.

As will be clear from the following discussion, there are different possibilities to incorporate the experimental data during the sampling stage. This partly depends on the kind of data that is used (e.g. the level of detail and the amount of inherent ambiguity) and the sampling method. 'Geometric' methods might limit the number of orientations selected for docking rather than adding experimental terms to an energy function. The search space is thus reduced based on the available experimental data. The subsequent docking and scoring stages then proceed as in *ab initio* docking [68]. Other approaches use anchor points based on experimental data, e.g. Treedock [69], or incorporate the experimental data by up weighting given residues in FFT-based rigid body docking approaches ('weighted geometric docking') [32, 70, 71]. Another popular possibility is to use some kind of distance restraints. This means that an additional energy term is created, which is high if residues which, according to the data, should be at the interface, i.e. close to each other, are far away in the proposed complex, and contrarily, low if they are near.

Ethylation interference and mutagenesis data have been used as experimental input for protein-DNA docking in the early data-driven Monte-Carlo docking program Monty [34, 72, 73], which allows side-chain flexibility and DNA deformations. Double mutant cycle data,

giving information about residue-residue contacts, have been incorporated as distance restraints in various applications [74-77]. A comparable approach was used to incorporate cross-linking data for a dimer of a four-transmembrane helix protein [78]: here a total of 10 distance restraints could be defined with rather small error bounds due to the rigid nature of the linker. There are several examples of the combination of NMR information with rigid body docking. Rigid body docking in X-plor [79] has been used to model the dynamic complex between plastocyanin and cytochrome f based on upper bound distance restraints derived from pseudo-contact shifts and CSP data, and lower bound distance restraints for residues assumed not to be in the interface [80, 81]. Saturation transfer (SAT) and RDC restraints have been combined with energy minimization to model the CAD-ICAD complex (complex between the CAD domain of caspase activated deoxyribonuclease and the CAD domain of its inhibitor) [82]. The nucleoprotein superhelix-DNA complex was modeled using CSP restraints in a grid search [83].

Some experimental data are highly ambiguous and only provide information about interface residues, but not about the specific contacts they make. Docking approaches should thus be capable of incorporating such ambiguity. Typical examples here would be the CSP data obtained from NMR titration experiments or mutagenesis data. With this in mind, we developed an information-driven semiflexible docking approach called HADDOCK [7] in which any kind of information about interface residues can be incorporated as a highly ambiguous interaction restraint (AIR) (see below). Related approaches have been described in [84] where NMR CSP data and RDCs were used, and in [85] for cross-linking information detected by MS.

HADDOCK

The method

As is clear from the discussion above, there is a wealth of experimental sources that can provide information about interfaces of biomolecular complexes. These data are generally not used, however. Our docking approach HADDOCK, an acronym for High Ambiguity Driven DOCKing [7], makes use of such information to drive the docking while allowing for various degrees of flexibility. The information is encoded in ambiguous interaction restraints (AIRs) similar to the ambiguous restraints commonly used in NMR structure determination [86]. The ambiguity here refers to the way in which the restraints are defined: between any residue which, based on experimental data, is believed to be an interface residue (called active residue), and all such residues (plus surface neighbors, called passive residues) on the partner molecule. An AIR is defined as an ambiguous intermolecular distance (d_{iAB}) with a maximum value of typically 2 Å between any atom m of an active residue i of protein A (m_{iA}) and any atom n of both active and passive residues k (N_{res} in total) of protein B (n_{kB}) (and inversely for protein B). The effective distance d_{iAB}^{eff} for each restraint is calculated using the equation:

$$d_{iAB}^{eff} = \left(\sum_{m_{iA}=1}^{Natoms} \sum_{k=1}^{NresB} \sum_{n_{kB}=1}^{Natoms} \frac{1}{d_{m_{iA}n_{kB}}^{6}} \right)^{-\frac{1}{6}}$$

where N_{atoms} indicates all atoms of a given residue and N_{res} the sum of active and passive residues for a given molecule. The definition of passive residues ensures that residues which are at the interface but are not detected (e.g. no CSP when using NMR, or no change in binding upon mutation) are still able to satisfy the AIR restraints, i.e. contact active residues of the partner molecule. The $1/r^6$ summation [87] is used to mimic the attractive part of a Lennard-Jones potential and ensures that the AIRs are satisfied as soon as any two atoms of the two proteins are in contact. The AIRs are incorporated as an additional energy term to the

energy function that one tries to minimize during the sampling. The docking proceeds in three stages during which increasing amounts of flexibility are introduced. In the first stage, the molecules are considered as rigid bodies and a large number of solutions are generated. In the second stage, a limited amount of flexibility is introduced first in the side chains and subsequently in both side-chains and backbone of pre-defined flexible segments encompassing the active and passive residues. Finally, the solutions are refined in explicit solvent. The final structures are clustered and scored using a combination of energy terms (mainly intermolecular van der Waals and electrostatic energies and restraint energies); for details see [7, 88]. Note that fully flexible models can also be defined, for example for the docking of an unstructured peptide onto a protein.

Applications

Several groups have used HADDOCK to generate models of biomolecular complexes in combination with different sources of information such as mutagenesis [89-91] or NMR CSP data [88, 89, 91-96]. A common problem resulting from the highly ambiguous nature of the interaction restraints is that symmetrical solutions are often obtained corresponding for example to a 180° rotation of one molecule with respect to the other. In cases where energy considerations cannot distinguish between the symmetrical solutions, additional information should ideally be supplemented. This was the case for the UbcH5-Not4 complex [88] (Figure 4a). To solve the symmetry problem, the HADDOCK models were used for structure-directed mutagenesis. Reverse mutants could be produced in which two residues of opposite charges across the interface were swapped, restoring thereby the binding. This provided unique, unambiguous information to select the correct solution.

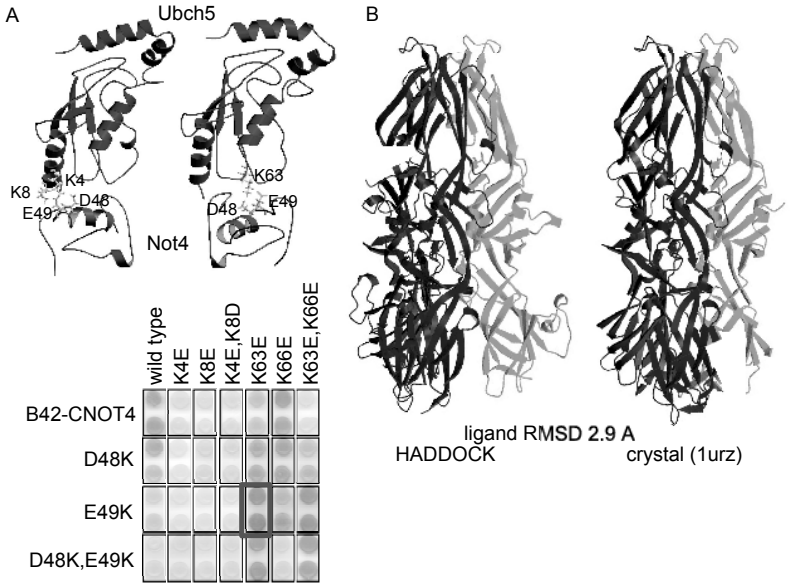


***Figure 4.** Two examples of structures calculated using HADDOCK. (A) the Ubch5/Not4 complex (pdb entry 1ur6) [88]. In a first docking run using only NMR CSP data, two models were obtained (top left and top right). Based on these, mutagenesis experiments were performed to discriminate between the two models: the charge-reversing double mutant E49K,K63E did restore the complex (red box), while the double mutants including K4E or K8E did not restore complex formation. Only the left solution is consistent with this information. (B) TBE virus envelope glycoprotein E trimer (CAPRI target 10), for which epitope, conservation and protection from enzymatic digestion data were introduced in HADDOCK, resulting in a docking model (left) within 2.9Å ligand-RMSD from the crystal structure [190] (pdb entry 1urz, right). The three subunits are color-coded; note that two segments (residue 148-159 and 204-209) are missing in the crystal structure.

In the case of the transient complex between the yeast copper chaperone Atx1 and the first soluble domain of the copper transporting ATPase Ccp2, a copper ion was explicitly introduced in the docking calculations based on NMR CSP data and found to move from Atx1 to Ccp2, consistent with the physiological direction of transfer [92]. The copper transfer intermediate was a result of the flexible docking protocol since no restraints were introduced to force the copper ion to move. This example indicates that flexible data-driven docking can be used to investigate not only 'static' structures but also more 'dynamic' aspects of biomolecular complexes. When available, classical NMR data such as NOEs can also be incorporated in HADDOCK as was the case to generate the solution structure of a non-specific protein-DNA complex [97].

Recently, we participated to the fourth and fifth round of the 'blind docking competition' CAPRI. As CAPRI is not especially meant for data-supported docking, we had to search literature and databases and use sequence conservation criteria (predicted *via* a neural network [98]) to define AIRs. Using HADDOCK, we were able to generate structures that are close to the experimentally defined structures even with low-resolution, "fuzzy" data such as epitope mapping and protection from enzymatic digestion. As an example, we successfully predicted the trimeric form of the TBE virus envelope glycoprotein E within 2.9Å ligand-RMSD (Figure 4b) (the ligand RMSD is defined as the RMSD calculated on one component after superposition of the other components). Our participation to the CAPRI experiment has, however, taught us that in some cases our docking methods, as well as others, can fail.

Conclusions and perspectives

The combination of biochemical and biophysical data with docking has many different applications. Docking models can obviously be used to select residues to be targeted for example for mutagenesis. One interesting point is that it becomes possible, when flexibility is explicitly introduced, to investigate structural changes at the interface upon complex formation, or even dynamic events as shown above for the copper transfer complex. Here we discuss what the future of this kind of approaches might be.

Perspectives on data used in docking

One interesting development is the use of conservation data to define interface residues (reviewed in [99]). Several methods have been developed for this purpose; examples are the use of a neural network [98, 100], the determination of invariant polar residues [101], three-dimensional cluster analysis [102], the use of phylogenetic trees [103], the Evolutionary Trace (ET) method [104, 105] and the Promate approach where conservation is combined with general interface characteristics [106]. Information from predicted interfaces has been used to model several complexes, for example, the Hsp90-p23 [107] and Gαβγ trimer-receptor complexes [42] based on predictions obtained with the ET method, and the complex between the α1 and β2 subunits of hemoglobin and the FtsA homodimer [43] based on conservation data and correlated mutations [46]. With the increasing amount of genomic data available, this kind of analysis can be expected to become more and more important. In addition, protein interaction networks can be compared using PathBLAST [108]; homologies based on this could provide additional information. Similarly, homology modeling, which has been improving over the years [109], in addition to being used to generate starting structures, could be combined with docking approaches, as illustrated with mutagenesis and neutron scattering data [110] and with MS data [111, 112]. An interesting example of the combination of homology modeling and docking is the Multiprospector multimeric threading approach [113] which has been applied to the *S. cerevisiae* proteome [114]: it threads the sequences of the

single chains of a target complex; if a template is found that is part of a complex, both chains of the target are rethreaded, now also incorporating an interfacial energy term.

Two experimental techniques which are very promising in combination with docking are cryo-electron microscopy or tomography and SAXS. Both techniques provide 'shape' information into which the structures of known constituents of a complex can be fit. Cryo-electron microscopy has been used for a large number of yeast complexes [115] and for the 80S ribosome from *S. cerevisase* [116]. For further discussion see reference [8]. SAXS data have been applied in docking to a variety of systems [117-124]. Specific examples are the twinfilin-capping protein complex [125] for which models of the single components were fit to the SAXS data and compared with mutagenesis data, and the FixJ response regulator where the rotation angle between the two domains was probed [126].

Another technique that can potentially be used is fluorescence. Interface information could be obtained for example for the complex of HscA with IscU LPPVK motif containing peptides [127]: the ability of Trp residues at the N or C terminus of the peptides to quench the fluorescence of labeled HscA was measured and this allowed to define the substrate binding orientation. In another example, docking simulations of HLA-1 dimers and complexes of those with CD8 and TCR were compared with Fluorescence Resonance Energy Transfer (FRET) data [128]. The use of FRET to study protein-DNA interactions has been reviewed [129]. Infrared spectroscopy might also become useful. For example, it was possible to define the tilt and relative orientation of transmembrane helices in the pentameric phospholamban [130] and the tetrameric M2 protein complex [131] based on infrared data.

With respect to the techniques discussed in section 2, at least for MS and NMR improvements can be expected. An example of a new MS approach for mapping interfaces is the modification of solvent-accessible side chains by hydroxyl radicals from millisecond exposure of aqueous solutions to X-rays; the modification sites can be identified by mass spectrometry and differences between complexed and uncomplexed forms indicate the location of the binding interface [132, 133]. In NMR, new approaches are emerging that might overcome the assignment problem. Comparison of experimental and back-calculated unassigned 1D ^{1}H spectra of a complex has been proposed as a mean to filter docking solutions; the feasibility of this approach was demonstrated for four complexes [134]. Other methods that do not require chemical shift assignments but rely on the combination of amino-acid specific labeling with saturation transfer or titration experiments have been reported as well [135, 136]. Provided that selective labeling can be efficiently performed, such methods should clearly speed up interface mapping by NMR.

Considering that information driven-docking will be much faster than conventional structural methods, it makes sense to invest some time and effort in making sure that the experimental data are reliable and really reveal interface residues. Therefore, whatever experimental technique is preferred, it is worth combining information from various sources.

Perspectives on docking methods

Not only from the data side, but also from the methodological point of view improvements are needed and can be expected. It will be possible one day to perform reliable *ab initio* docking, in which case no data will be needed at all, but this is probably not within our reach for the coming years. Still, active developments in the *ab iinitio* docking field will definitively benefit data driven docking approaches. Next to the need for proper scoring schemes, another important aspect is the handling of flexibility during docking. Although several methods exist that perform reasonably well in this respect, many still only use rigid body (soft) docking. Potential improvements might include a more widespread use of energy-driven sampling methods such as molecular dynamics, before docking to generate ensembles of starting structures, during docking to allow for induced conformational changes and/or after docking

to refine the (rigid body) solutions. Other advanced computational methods are emerging aiming at identifying parts of a molecule that are likely to be flexible and undergo conformational changes upon complex formation [137, 138]. Another kind of flexibility which, in our opinion without a good reason, has not had much attention, is that complexes themselves might be dynamical. As the forces that hold together the non-covalently linked complexes are, in most cases, weaker than those that are involved in covalent interactions, one would expect mobility to play a bigger role here. This will be particularly true in the case of weak and transient complexes. Methods should be developed which take this into account.

Perspectives on experimental systems amenable to data-driven docking

Finally, the range of systems which are studied using docking approaches can also be extended. Although it might not be strictly speaking docking, it is interesting to note that the kind of methods that we discussed here in the context of biomolecular complexes can also be applied to generate structures of single proteins by docking structural elements. This was done using cross-linking data to refine a homology model of FGF-2 [139] and with distance restraints for the lactose permease which consists of 12 transmembrane helices [140]. In another example, dipolar EPR distances, disulfide mapping distances and electron cryo-microscopy data were used in a special kind of exhaustive search using a graph-theory algorithm to generate models of rhodopsin [141]. Docking-like approaches are particularly interesting to model transmembrane helical proteins, as these typically contain considerable helical content already in their unfolded state; this means that docking approaches can be applied using helical segments as structural entities, as described for example in reference [142]. A general review about helix-helix interactions in the folding of membrane protein can be found in reference [143].

At the other extreme, data have become available for many giant multi subunit complexes such as the ribosome [144] or the regulatory complex of the Drosophila 26 S proteasome [145], but docking approaches have not often been used for them. A combinatorial approach such as CombDock [146] may be useful here, but HADDOCK or other docking methods can also easily be extended to deal with multiple subunits (as shown for the trimer example above), although, for large assemblies, computational requirements might become a limiting factor. Another kind of biological system for which data are becoming available now are protein-lipid assemblies. Using EPR, the orientation of phospholipase A_2 [147, 148] with respect to the surface of phospholipid vesicles was studied. For the C2 domain of protein kinase A, fluorescence and EPR data were used to elucidate the surface of the protein that contacts the membrane and to generate a model for the protein attached to a membrane [149]. NMR spin label data were also used to provide the depth and the angle of micelle insertion of the FYVE domain of early endosome antigen I [150]. Finally, one interesting type of systems to which increasing attention is given are proteins that, in their monomeric form, are unstructured and only fold during complex formation. A docking approach was used to study the complex of the (pre-folded) actin with the (only folding upon binding) thymosin beta4, using a combination of NMR data, mutation data and cross-linking data as restraints in the docking [151].

In conclusion, we have shown that docking methods can provide valuable biological insight, when combined with a limited amount of experimental data. Such a combination will, without doubt, become more widely used in the near future.

Acknowledgments

Financial support from the Netherlands Organization for Scientific Research (N.W.O.) through a “Jonge Chemici” grant to AMJJB (grant number 700.50.512) is acknowledged. We thank Cyril Dominguez and Sjoerd de Vries (Utrecht University) for helpful discussions.

References

1. Halperin, I., Ma, B. Y., Wolfson, H. & Nussinov, R. (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409-443.
2. Wodak, S. J. & Janin, J. (2003) Structural basis of macromolecular recognition. *Advances in Protein Chemistry* **61**, 9-73.
3. Brooijmans, N. & Kuntz, I. D. (2003) Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 335-373.
4. Vajda, S. & Camacho, C. J. (2004) Protein-protein docking: is the glass half-full or half-empty? *Trends Biotechnol.* **22**, 110-116.
5. Janin, J., Henrick, K., Moult, J., Ten Eyck, L., Sternberg, M. J. E., Vajda, S., Vasker, I. & Wodak, S. J. (2003) CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins* **52**, 2-9.
6. Fradera, X. & Mestres, J. (2004) Guided docking approaches to structure-based design and screening. *Curr. Top. Med. Chem.* **4**, 687-700.
7. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. (2003) HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731-1737.
8. Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, M. & Sali, A. (2004) A structural perspective on protein-protein interactions. *Curr. Opin. Struct. Biol.* **14**, 313-324.
9. McDonnell, J. M. (2001) Surface plasmon resonance: towards an understanding of the mechanisms of biological molecular recognition. *Curr. Opin. Chem. Biol.* **5**, 572-577.
10. Vidal, M., Brachmann, R. K., Fattaey, A., Harlow, E. & Boeke, J. D. (1996) Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. *Proc. Natl. Acad. Sci. USA* **93**, 10315-10320.
11. Sidhu, S. S., Fairbrother, W. J. & Deshayes, K. (2003) Exploring protein-protein interactions with phage display. *ChemBioChem* **4**, 14-25.
12. Clackson, T. & Wells, J. A. (1995) A Hot-Spot of Binding-Energy in a Hormone-Receptor Interface. *Science* **267**, 383-386.
13. DeLano, W. L. (2002) Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.* **12**, 14-20.
14. Thorn, K. S. & Bogan, A. A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* **17**, 284-285.
15. Carter, P. J., Winter, G., Wilkinson, A. J. & Fersht, A. R. (1984) The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus). *Cell* **38**, 835-840.
16. Hanson, C. L. & Robinson, C. V. (2004) Protein-nucleic acid interactions and the expanding role of mass spectrometry. *J. Biol. Chem.* **279**, 24907-24910.
17. Hernandez, H. & Robinson, C. V. (2001) Dynamic protein complexes: Insights from mass spectrometry. *J. Biol. Chem.* **276**, 46685-46688.
18. Lanman, J. & Prevelige, P. E. (2004) High-sensitivity mass spectrometry for imaging subunit interactions: hydrogen/deuterium exchange. *Curr. Opin. Struct. Biol.* **14**, 181-188.
19. Garcia, R. A., Pantazatos, D. & Villarreal, F. J. (2004) Hydrogen/deuterium exchange mass spectrometry for investigating protein-ligand interactions. *Assay and Drug Development Technologies* **2**, 81-91.
20. Back, J. W., de Jong, L., Muijsers, A. O. & de Koster, C. G. (2003) Chemical cross-linking and mass spectrometry for protein structural modeling. *J. Mol. Biol.* **331**, 303-313.
21. Zuiderweg, E. R. (2002) Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry* **41**, 1-7.
22. Takahashi, H., Nakanishi, T., Kami, K., Arata, Y. & Shimada, I. (2000) A novel NMR method for determining the interfaces of large protein-protein complexes. *Nat. Struct. Biol.* **7**, 220-223.
23. Bax, A. (2003) Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci.* **12**, 1-16.
24. Fushman, D., Varadan, R., Assfalg, M. & Walker, O. (2004) Determining domain orientation in macromolecules by using spin-relaxation and residual dipolar coupling measurements. *Prog. Nucl. Magn. Reson. Spectrosc.* **44**, 189-214.
25. Gaponenko, V., Altieri, A. S., Li, J. & Byrd, R. A. (2002) Breaking symmetry in the structure determination of (large) symmetric protein dimers. *J. Biomol. NMR* **24**, 143-148.
26. Gaponenko, V., Sarma, S. P., Altieri, A. S., Horita, D. A., Li, J. & Byrd, R. A. (2004) Improving the accuracy of NMR structures of large proteins using pseudocontact shifts as long-range restraints. *J. Biomol. NMR* **28**, 205-212.
27. Arumugam, S. & Van Doren, S. R. (2003) Global orientation of bound MMP-3 and N-TIMP-1 in solution via residual dipolar couplings. *Biochemistry* **42**, 7950-7958.

28. Gabb, H. A., Jackson, R. M. & Sternberg, M. J. E. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272**, 106-120.
29. Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I. & Ten Eyck, L. F. (2001) Protein docking using continuum electrostatics and geometric fit. *Protein Eng.* **14**, 105-113.
30. Vakser, I. A. (1995) Protein Docking for Low-Resolution Structures. *Protein Eng.* **8**, 371-377.
31. Meyer, M., Wilson, P. & Schomburg, D. (1996) Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking. *J. Mol. Biol.* **264**, 199-210.
32. Ben-Zeev, E. & Eisenstein, M. (2003) Weighted geometric docking: Incorporating external information in the rotation-translation scan. *Proteins* **52**, 24-27.
33. Chen, R., Li, L. & Weng, Z. P. (2003) ZDOCK: An initial-stage protein-docking algorithm. *Proteins* **52**, 80-87.
34. Knegtel, R. M. A., Boelens, R. & Kaptein, R. (1994) Monte Carlo docking of protein-DNA complexes: incorporation of DNA flexibility and experimental data. *Protein Eng.* **7**, 761-767.
35. Abagyan, R., Totrov, M. & Kuznetsov, D. (1994) Icm - a New Method for Protein Modeling and Design - Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comp. Chem.* **15**, 488-506.
36. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K. & Olson, A. J. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.* **19**, 1639-1662.
37. Mackerell, A. D. (2004) Empirical force fields for biological macromolecules: Overview and issues. *J. Comp. Chem.* **25**, 1584-1604.
38. Rajesh, S., Sakamoto, T., Iwamoto-Sugai, M., Shibata, T., Kohno, T. & Ito, Y. (1999) Ubiquitin binding interface mapping on yeast ubiquitin hydrolase by NMR. *Biochemistry* **38**, 9242-53.
39. Singh, S., Folkers, G. E., Bonvin, A. M. J. J., Boelens, R., Wechselberger, R., Niztayev, A. & Kaptein, R. (2002) Solution structure and DNA-binding properties of the C-terminal domain of UvrC from E.coli. *EMBO J.* **21**, 6257-6266.
40. Chien, C. Y., Xu, Y. J., Xiao, R., Aramini, J. M., Sahasrabudhe, P. V., Krug, R. M. & Montelione, G. T. (2004) Biophysical characterization of the complex between double-stranded RNA and the N-terminal domain of the NS1 protein from influenza A virus: Evidence for a novel RNA-binding mode. *Biochemistry* **43**, 1950-1962.
41. Comolli, L. R., Pelton, J. G. & Tinoco, I. (1998) Mapping of a protein-RNA kissing hairpin interface: Rom and Tar-Tar. *Nuc. Acids Res.* **26**, 4688-4695.
42. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996) Evolutionarily conserved G(alpha beta gamma) binding surfaces support a model of the G protein-receptor complex. *Proc. Natl. Acad. Sci. USA* **93**, 7507-7511.
43. Carettoni, D., Gomez-Puertas, P., Yim, L., Mingorance, J., Massidda, O., Vicente, M., Valencia, A., Domenici, E. & Anderluzzi, D. (2003) Phage-display and correlated mutations identify an essential region of subdomain 1C involved in homodimerization of Escherichia coli FtsA. *Proteins* **50**, 192-206.
44. Adams, P. D., Arkin, I. T., Engelman, D. M. & Brunger, A. T. (1995) Computational Searching and Mutagenesis Suggest a Structure for the Pentameric Transmembrane Domain of Phospholamban. *Nat. Struct. Biol.* **2**, 154-162.
45. Adams, P. D., Engelman, D. M. & Brunger, A. T. (1996) Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching. *Proteins* **26**, 257-261.
46. Pazos, F., HelmerCitterich, M., Ausiello, G. & Valencia, A. (1997) Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511-523.
47. Li, R. H., Gorelik, R., Nanda, V., Law, P. B., Lear, J. D., DeGrado, W. F. & Bennett, J. S. (2004) Dimerization of the transmembrane domain of integrin alpha(IIb) subunit in cell membranes. *J. Biol. Chem.* **279**, 26666-26673.
48. Ritchie, D. W. & Kemp, G. J. L. (2000) Protein docking using spherical polar Fourier correlations. *Proteins* **39**, 178-194.
49. Gaboriaud, C., Juanhuix, J., Gruez, A., Lacroix, M., Darnault, C., Pignol, D., Verger, D., Fontecilla-Camps, J. C. & Arlaud, G. J. (2003) The crystal structure of the globular head of complement protein C1q provides a basis for its versatile recognition properties. *J. Biol. Chem.* **278**, 46974-46982.
50. Anand, G. S., Law, D., Mandell, J. G., Snead, A. N., Tsigelny, I., Taylor, S. S., Ten Eyck, L. F. & Komives, E. A. (2003) Identification of the protein kinase A regulatory R-I alpha-catalytic subunit interface by amide H/H-2 exchange and protein docking. *Proc. Natl. Acad. Sci. USA* **100**, 13264-13269.
51. Azuma, Y., Renault, L., Garcia-Ranea, J. A., Valencia, A., Nishimoto, T. & Wittinghofer, A. (1999) Model of the Ran-RCC1 interaction using biochemical and docking experiments. *J. Mol. Biol.* **289**, 1119-1130.
52. Dobrodumov, A. & Gronenborn, A. M. (2003) Filtering and selection of structural models: Combining docking and NMR. *Proteins* **53**, 18-32.
53. Palma, P. N., Krippahl, L., Wampler, J. E. & Moura, J. J. G. (2000) BiGGER: A new (soft) docking algorithm for predicting protein interactions. *Proteins* **39**, 372-384.

54. Pettigrew, G. W., Pauleta, S. R., Goodhew, C. F., Cooper, A., Nutley, M., Jumel, K., Harding, S. E., Costa, C., Krippahl, L., Moura, I. & Moura, J. (2003) Electron transfer complexes of cytochrome c peroxidase from Paracoccus denitrificans containing more than one cytochrome. *Biochemistry* **42**, 11968-11981.
55. Morelli, X. J., Palma, P. N., Guerlesquin, F. & Rigby, A. C. (2001) A novel approach for assessing macromolecular complexes combining soft-docking calculations with NMR data. *Protein Sci.* **10**, 2131-2137.
56. Crowley, P. B., Rabe, K. S., Worrall, J. A. R., Canters, G. W. & Ubbink, M. (2002) The ternary complex of cytochrome f and cytochrome c: Identification of a second binding site and competition for plastocyanin binding. *ChemBioChem* **3**, 526-533.
57. Worrall, J. A. R., Liu, Y. J., Crowley, P. B., Nocek, J. M., Hoffman, B. M. & Ubbink, M. (2002) Myoglobin and cytochrome b(5): A nuclear magnetic resonance study of a highly dynamic protein complex. *Biochemistry* **41**, 11721-11730.
58. Morelli, X., Dolla, A., Czjzek, M., Palma, P. N., Blasco, F., Krippahl, L., Moura, J. J. G. & Guerlesquin, F. (2000) Heteronuclear NMR and soft docking: An experimental approach for a structural model of the cytochrome c(553)-ferredoxin complex. *Biochemistry* **39**, 2530-2537.
59. McKenna, S., Moraes, T., Pastushok, L., Ptak, C., Xiao, W., Spyracopoulos, L. & Ellison, M. J. (2003) An NMR-based model of the ubiquitin-bound human ubiquitin conjugation complex Mms2 center dot Ubc13 - The structural basis for lysine 63 chain catalysis. *J. Biol. Chem.* **278**, 13151-13158.
60. Meng, E. C., Gschwend, D. A., Blaney, J. M. & Kuntz, I. D. (1993) Orientational Sampling and Rigid-Body Minimization in Molecular Docking. *Proteins* **17**, 266-278.
61. Cuff, L., Ulrich, R. G. & Olson, M. A. (2003) Prediction of the multimeric assembly of staphylococcal enterotoxin A with cell-surface protein receptors. *J. Mol. Graph. Mod.* **21**, 473-486.
62. Sachchidanand, Lequin, O., Staunton, D., Mulloy, B., Forster, M. J., Yoshida, K. & Campbell, I. D. (2002) Mapping the heparin-binding site on the (13-14)F3 fragment of fibronectin. *J. Biol. Chem.* **277**, 50629-50635.
63. Yu, K., Fu, W., Liu, H., Luo, X., Chen, K. X., Ding, J., Shen, J. & Jiang, H. (2004) Computational Simulations of Interactions of Scorpion Toxins with the Voltage-Gated Potassium Ion Channel. *Biophys. J.* **86**, 3542-3555.
64. Xu, X. P. & Case, D. A. (2001) Automated prediction of N-15, C-13(alpha), C-13(beta) and C-13 ' chemical shifts in proteins using a density functional database. *J. Biomol. NMR* **21**, 321-333.
65. Neal, S., Nip, A. M., Zhang, H. Y. & Wishart, D. S. (2003) Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts. *J. Biomol. NMR* **26**, 215-240.
66. Stamos, J., Eigenbrot, C., Nakamura, G. R., Reynolds, M. E., Yin, J., Lowman, H. B., Fairbrother, W. J. & Starovasnik, M. A. (2004) Convergent Recognition of the IgE Binding Site on the High-Affinity IgE Receptor. *Structure* **12**, 1289-1301.
67. McCoy, M. A. & Wyss, D. F. (2002) Structures of protein-protein complexes are docked using only NMR restraints from residual dipolar coupling and chemical shift perturbations. *J. Am. Chem. Soc.* **124**, 2104-2105.
68. Schneidman-Duhovny, D., Inbar, Y., Polak, V., Shatsky, M., Halperin, I., Benyamini, H., Barzilai, A., Dror, O., Haspel, N., Nussinov, R. & Wolfson, H. J. (2003) Taking geometry to its edge: Fast unbound rigid (and hinge-bent) docking. *Proteins* **52**, 107-112.
69. Fahmy, A. & Wagner, G. (2002) TreeDock: A tool for protein docking based on minimizing van der Waals energies. *J. Am. Chem. Soc.* **124**, 1241-1250.
70. Ben-Zeev, E., Zarivach, R., Shoham, M., Yonath, A. & Eisenstein, M. (2003) Prediction of the structure of the complex between the 30S ribosomal subunit and colicin E3 via weighted-geometric docking. *J. Biomol. Struct. Dyn.* **20**, 669-675.
71. Zarivach, R., Ben-Zeev, E., Wu, N., Auerbach, T., Bashan, A., Jakes, K., Dickman, K., Kosmidis, A., Schluenzen, F., Yonath, A., Eisenstein, M. & Shoham, M. (2002) On the interaction of colicin E3 with the ribosome. *Biochimie* **84**, 447-54.
72. Knegtel, R. M. A., Fogh, R. H., Ottleben, G., Ruterjans, H., Dumoulin, P., Schnarr, M., Boelens, R. & Kaptein, R. (1995) A Model for the Lexa Repressor DNA Complex. *Proteins* **21**, 226-236.
73. Rumpel, S., Razeto, A., Pillar, C. M., Vijayan, V., Taylor, A., Giller, K., Gilmore, M. S., Becker, S. & Zweckstetter, M. (2004) Structure and DNA-binding properties of the cytolysin regulator CylR2 from Enterococcus faecalis. *EMBO J.* **23**, 3632-3642.
74. Gilquin, B., Racape, J., Wrisch, A., Visan, V., Lecoq, A., Grissmer, S., Menez, A. & Gasparini, S. (2002) Structure of the BgK-Kv1.1 complex based on distance restraints identified by double mutant cycles - Molecular basis for convergent evolution of Kv1 channel blockers. *J. Biol. Chem.* **277**, 37406-37413.
75. Eriksson, M. A. L. & Roux, B. (2002) Modeling the structure of Agitoxin in complex with the Shaker K+ channel: A computational approach based on experimental distance restraints extracted from thermodynamic mutant cycles. *Biophys. J.* **83**, 2595-2609.
76. Fruchart-Gaillard, C., Gilquin, B., Antil-Delbeke, S., Le Novere, N., Tamiya, T., Corringer, P. J., Changeux, J. P., Menez, A. & Servent, D. (2002) Experimentally based model of a complex between a snake toxin and the alpha 7 nicotinic receptor. *Proc. Natl. Acad. Sci. USA* **99**, 3216-3221.

77. Roisman, L. C., Piehler, J., Trosset, J. Y., Scheraga, H. A. & Schreiber, G. (2001) Structure of the interferon-receptor complex determined by distance constraints from double-mutant cycles and flexible docking. *Proc. Natl. Acad. Sci. USA* **98**, 13231-13236.
78. Gottschalk, K.-E., Soskine, M., Schuldiner, S. & Kessler, H. (2004) A Structural Model of EmrE, a Multi-Drug Transporter from Escherichia coli. *Biophys. J.* **86**, 3335-3348.
79. Brunger, A. T. (1992) *X-PLOR 3.1 manual*, Yale University Press, New Haven, CT, USA.
80. Ubbink, M., Ejdeback, M., Karlsson, B. G. & Bendall, D. S. (1998) The structure of the complex of plastocyanin and cytochrome f, determined by paramagnetic NMR and restrained rigid-body molecular dynamics. *Structure* **6**, 323-335.
81. Crowley, P. B., Otting, G., Schlarb-Ridley, B. G., Canters, G. W. & Ubbink, M. (2001) Hydrophobic interactions in a cyanobacterial plastocyanin-cytochrome f complex. *J. Am. Chem. Soc.* **123**, 10444-10453.
82. Matsuda, T., Ikegami, T., Nakajima, N., Yamazaki, T. & Nakamura, H. (2004) Model building of a protein-protein complexed structure using saturation transfer and residual dipolar coupling without paired intermolecular NOE. *J. Biomol. NMR* **29**, 325-338.
83. Folmer, R. H. A., Nilges, M., Papavoine, C. H. M., Harmsen, B. J. M., Konings, R. N. H. & Hilbers, C. W. (1997) Refined structure, DNA binding studies, and dynamics of the bacteriophage Pf3 encoded single-stranded DNA binding protein. *Biochemistry* **36**, 9120-9135.
84. Clore, G. M. & Schwieters, C. D. (2003) Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from H-1(N)/N-15 chemical shift mapping and backbone N-15-H-1 residual dipolar couplings using conjoined rigid body/torsion angle dynamics. *J. Am. Chem. Soc.* **125**, 2902-2912.
85. Schulz, D. M., Ihling, C., Clore, G. M. & Sinz, A. (2004) Mapping the topology and determination of a low-resolution three-dimensional structure of the calmodulin-melittin complex by chemical cross-linking and high-resolution FTICRMS: Direct demonstration of multiple binding modes. *Biochemistry* **43**, 4703-4715.
86. Nilges, M. & O'Donoghue, S. I. (1998) Ambiguous NOEs and automated NOE assignment. *Prog. Nucl. Magn. Reson. Spectrosc.* **32**, 107-139.
87. Nilges, M. (1993) A Calculation Strategy for the Structure Determination of Symmetrical Dimers by H-1-Nmr. *Proteins* **17**, 297-309.
88. Dominguez, C., Bonvin, A. M. J. J., Winkler, G. S., van Schaik, F. M. A., Timmers, H. T. M. & Boelens, R. (2004) Structural model of the UbcH5B/CNOT4 complex revealed by combining NMR, mutagenesis, and docking approaches. *Structure* **12**, 633-644.
89. Gao, G., Prutzman, K. C., King, M. L., Scheswohl, D. M., DeRose, E. F., London, R. E., Schaller, M. D. & Campbell, S. L. (2004) NMR Solution Structure of the Focal Adhesion Targeting Domain of Focal Adhesion Kinase in Complex with a Paxillin LD Peptide: EVIDENCE FOR A TWO-SITE BINDING MODEL. *J. Biol. Chem.* **279**, 8441-8451.
90. Simpson, R. J. Y., Lee, S. H. Y., Bartle, N., Sum, E. Y., Visvader, J. E., Matthews, J. M., Mackay, J. P. & Crossley, M. (2004) A classic zinc finger from friend of GATA mediates an interaction with the coiled-coil of transforming acidic coiled-coil 3. *J. Biol. Chem.* **279**, 39789-39797.
91. Jain, N. U., Wyckoff, T. J. O., Raetz, C. R. H. & Prestegard, J. H. (2004) Rapid Analysis of Large Protein-Protein Complexes Using NMR-derived Orientational Constraints: The 95 kDa Complex of LpxA with Acyl Carrier Protein. *J. Mol. Biol.* **343**, 1379-1389.
92. Arnesano, F., Banci, L., Bertini, I. & Bonvin, A. M. J. J. (2004) A Docking Approach to the Study of Copper Trafficking Proteins: Interaction between Metallochaperones and Soluble Domains of Copper ATPases. *Structure* **12**, 669-676.
93. Mueller, T. D., Kamionka, M. & Feigon, J. (2004) Specificity of the Interaction between Ubiquitin-associated Domains and Ubiquitin. *J. Biol. Chem.* **279**, 11926-11936.
94. Stauffer, M. E. & Chazin, W. J. (2004) Physical interaction between replication protein A and Rad51 promotes exchange on single-stranded DNA. *J. Biol. Chem.* **279**, 25638-25645.
95. van Drogen-Petit, A., Zwahlen, C., Peter, M. & Bonvin, A. M. (2004) Insight into molecular interactions between two PB1 domains. *J. Mol. Biol.* **336**, 1195-210.
96. Yuan, X. M., Simpson, P., Mckeown, C., Kondo, H., Uchiyama, K., Wallis, R., Dreveny, I., Keetch, C., Zhang, X. D., Robinson, C., Freemont, P. & Matthews, S. (2004) Structure, dynamics and interactions of p47, a major adaptor of the AAA ATPase, p97. *EMBO J.* **23**, 1463-1473.
97. Kalodimos, C. G., Biris, N., Bonvin, A. M. J. J., Levandoski, M. M., Guennuegues, M., Boelens, R. & Kaptein, R. (2004) Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science* **305**, 386-389.
98. Zhou, H. X. & Shan, Y. B. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* **44**, 336-343.
99. Lichtarge, O. & Sowa, M. E. (2002) Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 21-27.

100. Fariselli, P., Pazos, F., Valencia, A. & Casadio, R. (2002) Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.* **269**, 1356-1361.
101. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. E. (2001) Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**, 395-408.
102. Landgraf, R., Xenarios, I. & Eisenberg, D. (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487-1502.
103. Armon, A., Graur, D. & Ben-Tal, N. (2001) ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447-463.
104. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342-358.
105. Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E. & Lichtarge, O. (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139-154.
106. Neuvirth, H., Raz, R. & Schreiber, G. (2004) ProMate: A structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.* **338**, 181-199.
107. Zhu, S. & Tytgat, J. (2004) Evolutionary epitopes of Hsp90 and p23: implications for their interaction. *FASEB J.* **18**, 940-947.
108. Kelley, B. P., Yuan, B. B., Lewitter, F., Sharan, R., Stockwell, B. R. & Ideker, T. (2004) PathBLAST: a tool for alignment of protein interaction networks. *Nuc. Acids Res.* **32**, W83-W88.
109. Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. (2003) Assessment of progress over the CASP experiments. *Proteins* **53**, 585-595.
110. Tung, C. S., Walsh, D. A. & Trewhella, J. (2002) A structural model of the catalytic subunit-regulatory subunit dimeric complex of the cAMP-dependent protein kinase. *J. Biol. Chem.* **277**, 12423-12431.
111. D'Ambrosio, C., Talamo, F., Vitale, R. M., Amodeo, P., Tell, G., Ferrara, L. & Scaloni, A. (2003) Probing the dimeric structure of porcine aminoacylase 1 by mass spectrometric and modeling procedures. *Biochemistry* **42**, 4430-4443.
112. Taverner, T., Hall, N. E., O'Hair, R. A. J. & Simpson, R. J. (2002) Characterization of an antagonist interleukin-6 dimer by stable isotope labeling, cross-linking, and mass spectrometry. *J. Biol. Chem.* **277**, 46487-46492.
113. Lu, L., Lu, H. & Skolnick, J. (2002) MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* **49**, 350-364.
114. Lu, L., Arakaki, A. K., Lu, H. & Skolnick, J. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: Application to the Saccharomyces cerevisiae proteome. *Gen. Res.* **13**, 1146-1154.
115. Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A. C., Bork, P., Superti-Furga, G., Serrano, L. & Russell, R. B. (2004) Structure-based assembly of protein complexes in yeast. *Science* **303**, 2026-2029.
116. Spahn, C. M. T., Beckmann, R., Eswar, N., Penczek, P. A., Sali, A., Blobel, G. & Frank, J. (2001) Structure of the 80S ribosome from Saccharomyces cerevisiae - tRNA-ribosome and subunit-subunit interactions. *Cell* **107**, 373-386.
117. Dainese, E., Svergun, D., Beltramini, M., Di Muro, P. & Salvato, B. (2000) Low-resolution structure of the proteolytic fragments of the Rapana venosa hemocyanin in solution. *Arch. Biochem. Biophys.* **373**, 154-162.
118. de Azevedo, W. F., dos Santos, G. C., dos Santos, D. M., Olivieri, J. R., Canduri, F., Silva, R. G., Basso, L. A., Renard, G., da Fonseca, I. O., Mendes, M. A., Palma, M. S. & Santos, D. S. (2003) Docking and small angle X-ray scattering studies of purine nucleoside phosphorylase. *Biochem. Biophys. Res. Comm.* **309**, 923-928.
119. Grossmann, J. G., Sharff, A. J., O'Hare, P. & Luisi, B. (2001) Molecular shapes of transcription factors TFIIB and VP16 in solution: Implications for recognition. *Biochemistry* **40**, 6267-6274.
120. Svergun, D. I., Aldag, I., Sieck, T., Altendorf, K., Koch, M. H. J., Kane, D. J., Kozin, M. B. & Gruber, G. (1998) A model of the quaternary structure of the Escherichia coli F-1 ATPase from X-ray solution scattering and evidence for structural changes in the delta subunit during ATP hydrolysis. *Biophys. J.* **75**, 2212-2219.
121. Callaghan, A. J., Grossmann, J. G., Redko, Y. U., Ilag, L. L., Moncrieffe, M. C., Symmons, M. F., Robinson, C. V., McDowall, K. J. & Luisi, B. F. (2003) Quaternary structure and catalytic activity of the Escherichia coli ribonuclease E amino-terminal catalytic domain. *Biochemistry* **42**, 13848-13855.
122. Marquez, J. A., Smith, C. I. E., Petoukhov, M. V., Lo Surdo, P., Mattsson, P. T., Knekt, M., Westlund, A., Scheffzek, K., Saraste, M. & Svergun, D. I. (2003) Conformation of full-length Bruton tyrosine kinase (Btk) from synchrotron X-ray solution scattering. *EMBO J.* **22**, 4616-4624.
123. Auguin, D., Barthe, P., Royer, C., Stern, M. H., Noguchi, M., Arold, S. T. & Roumestand, C. (2004) Structural basis for the co-activation of protein kinase B by T-cell leukemia-1 (TCL1) family proto-oncoproteins. *J. Biol. Chem.* **279**, 35890-35902.

124. Sun, Z., Reid, K. B. M. & Perkins, S. J. (2004) The Dimeric and Trimeric Solution Structures of the Multidomain Complement Protein Properdin by X-ray Scattering, Analytical Ultracentrifugation and Constrained Modelling. *J. Mol. Biol.* **343**, 1327-1343.
125. Falck, S., Paavilainen, V. O., Wear, M. A., Grossmann, J. G., Cooper, J. A. & Lappalainen, P. (2004) Biological role and structural mechanism of twinfilin-capping protein. *EMBO J.* **23**, 3010-9.
126. Birck, C., Malfois, M., Svergun, D. & Samama, J. P. (2002) Insights into signal transduction revealed by the low resolution structure of the FixJ response regulator. *J. Mol. Biol.* **321**, 447-457.
127. Tapley, T. L. & Vickery, L. E. (2004) Preferential substrate binding orientation by the molecular chaperone HscA. *J. Biol. Chem.* **279**, 28435-28442.
128. Gaspar, R., Bagossi, P., Bene, L., Matko, J., Szollosi, J., Tozser, J., Fesus, L., Waldmann, T. A. & Damjanovich, S. (2001) Clustering of class IHLA oligomers with CD8 and TCR: Three-dimensional models based on fluorescence resonance energy transfer and crystallographic data. *J. Immunol.* **166**, 5078-5086.
129. Hillisch, A., Lorenz, M. & Diekmann, S. (2001) Recent advances in FRET: distance determination in protein-DNA complexes. *Curr. Opin. Struct. Biol.* **11**, 201-207.
130. Torres, J., Adams, P. D. & Arkin, I. T. (2000) Use of a new label C-13=O-18 in the determination of a structural model of phospholamban in a lipid bilayer. Spatial restraints resolve the ambiguity arising from interpretations of mutagenesis data. *J. Mol. Biol.* **300**, 677-685.
131. Kukol, A., Adams, P. D., Rice, L. M., Brunger, A. T. & Arkin, I. T. (1999) Experimentally based orientational refinement of membrane protein models: A structure for the Influenza A M2 H+ channel. *J. Mol. Biol.* **286**, 951-962.
132. Guan, J. Q., Almo, S. C., Reisler, E. & Chance, M. R. (2003) Structural reorganization of proteins revealed by radiolysis and mass spectrometry: G-actin solution structure is divalent cation dependent. *Biochemistry* **42**, 11992-12000.
133. Guan, J. Q., Almo, S. C. & Chance, M. R. (2004) Synchrotron radiolysis and mass spectrometry: A new approach to research on the actin cytoskeleton. *Acc. Chem. Res.* **37**, 221-229.
134. Kohlbacher, O., Burchardt, A., Moll, A., Hildebrandt, A., Bayer, P. & Lenhof, H. P. (2001) Structure prediction of protein complexes by an NMR-based protein docking algorithm. *J. Biomol. NMR* **20**, 15-21.
135. Hajduk, P. J., Mack, J. C., Olejniczak, E. T., Park, C., Dandliker, P. J. & Beutel, B. A. (2004) SOS-NMR: A saturation transfer NMR-based method for determining the structures of protein-ligand complexes. *J. Am. Chem. Soc.* **126**, 2390-2398.
136. Parker, M. J., Aulton-Jones, M., Hounslow, A. M. & Craven, C. J. (2004) A combinatorial selective labeling method for the assignment of backbone amide NMR resonances. *J. Am. Chem. Soc.* **126**, 5020-5021.
137. Zacharias, M. (2004) Rapid protein-ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: Binding of FK506 to FKBP. *Proteins-Structure Function and Bioinformatics* **54**, 759-767.
138. Kovacs, J. A., Chacon, P. & Abagyan, R. (2004) Predictions of protein flexibility: First-order measures. *Proteins-Structure Function and Bioinformatics* **56**, 661-668.
139. Young, M. M., Tang, N., Hempel, J. C., Oshiro, C. M., Taylor, E. W., Kuntz, I. D., Gibson, B. W. & Dollinger, G. (2000) High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci. USA* **97**, 5802-5806.
140. Sorgen, P. L., Hu, Y. L., Guan, L., Kaback, H. R. & Girvin, M. E. (2002) An approach to membrane protein structure without crystals. *Proc. Natl. Acad. Sci. USA* **99**, 14037-14040.
141. Faulon, J. L., Sale, K. & Young, M. (2003) Exploring the conformational space of membrane protein folds matching distance constraints. *Protein Sci.* **12**, 1750-1761.
142. Sale, K., Faulon, J.-L., Gray, G. A., Schoeniger, J. S. & Young, M. M. (2004) Optimal bundling of transmembrane helices using sparse distance constraints. *Protein Sci.* **13**, 2613-2627.
143. DeGrado, W. F., Gratkowski, H. & Lear, J. D. (2003) How do helix-helix interactions help determine the folds of membrane proteins? Perspectives from the study of homo-oligomeric helical bundles. *Protein Sci.* **12**, 647-665.
144. Whirl-Carrillo, M., Gabashvili, I. S., Bada, M., Banatao, D. R. & Altman, R. B. (2002) Mining biochemical information: Lessons taught by the ribosome. *RNA* **8**, 279-289.
145. Kurucz, E., Ando, I., Sumegi, M., Holzl, H., Kapelari, B., Baumeister, W. & Udvardy, A. (2002) Assembly of the Drosophila 26 S proteasome is accompanied by extensive subunit rearrangements. *Biochem. J.* **365**, 527-536.
146. Inbar, Y., Benyamini, H., Nussinov, R. & Wolfson, H. J. (2003) Protein structure prediction via combinatorial assembly of sub-structural units. *Bioinformatics* **19**, 158i-168.
147. Ball, A., Nielsen, R., Gelb, M. H. & Robinson, B. H. (1999) Interfacial membrane docking of cytosolic phospholipase A(2) C2 domain using electrostatic potential-modulated spin relaxation magnetic resonance. *Proc. Natl. Acad. Sci. USA* **96**, 6637-6642.

148. Lin, Y., Nielsen, R., Murray, D., Hubbell, W. L., Mailer, C., Robinson, B. H. & Gelb, M. H. (1998) Docking phospholipase A(2) on membranes using electrostatic potential-modulated spin relaxation magnetic resonance. *Science* **279**, 1925-1929.
149. Kohout, S. C., Corbalan-Garcia, S., Gomez-Fernandez, J. C. & Falke, J. J. (2003) C2 domain of protein kinase C alpha: Elucidation of the membrane docking surface by site-directed fluorescence and spin labeling. *Biochemistry* **42**, 1254-1265.
150. Kutateladze, T. G., Capelluto, D. G. S., Ferguson, C. G., Cheever, M. L., Kutateladze, A. G., Prestwich, G. D. & Overduin, M. (2004) Multivalent mechanism of membrane insertion by the FYVE domain. *J. Biol. Chem.* **279**, 3050-3057.
151. Domanski, M., Hertzog, M., Coutant, J., Gutsche-Perelroizen, I., Bontems, F., Carlier, M. F., Guittet, E. & van Heijenoort, C. (2004) Coupling of folding and binding of thymosin beta 4 upon interaction with monomeric actin monitored by nuclear magnetic resonance. *J. Biol. Chem.* **279**, 23637-23645.
152. Norledge, B. V., Petrovan, R. J., Ruf, W. & Olson, A. J. (2003) The tissue factor/factor VIIa/factor Xa complex: a model built by docking. *Proteins* **53**, 640-8.
153. Sadir, R., Baleux, F., Grosdidier, A., Imberty, A. & Lortat-Jacob, H. (2001) Characterization of the stromal cell-derived factor-1alpha-heparin. *J. Biol. Chem.* **276**, 8288-96.
154. Karim, C. B., Stamm, J. D., Karim, J., Jones, L. R. & Thomas, D. D. (1998) Cysteine reactivity and oligomeric structures of phospholamban and its mutants. *Biochemistry* **37**, 12074-12081.
155. Herzyk, P. & Hubbard, R. E. (1998) Using experimental information to produce a model of the transmembrane domain of the ion channel phospholamban. *Biophys. J.* **74**, 1203-1214.
156. Jespers, L., Lijnen, H. R., Vanwetswinkel, S., Van Hoef, B., Brepoels, K., Collen, D. & De Maeyer, M. (1999) Guiding a docking mode by phage display: selection of correlated mutations. *J. Mol. Biol.* **290**, 471-9.
157. Onrust, R., Herzmark, P., Chi, P., Garcia, P. D., Lichtarge, O., Kingsley, C. & Bourne, H. R. (1997) Receptor and beta gamma binding sites in the alpha subunit of the retinal G protein transducin. *Science* **275**, 381-384.
158. Gruschus, J. M., Greene, L. E., Eisenberg, E. & Ferretti, J. A. (2004) Experimentally biased model structure of the Hsc70/auxilin complex: Substrate transfer and interdomain structural change. *Protein Sci.* **13**, 2029-2044.
159. Bracci, L., Pini, A., Bernini, A., Lelli, B., Ricci, C., Scarselli, M., Niccolai, N. & Neri, P. (2003) Biochemical filtering of a protein-protein docking simulation identifies the structure of a complex between a recombinant antibody fragment and alpha-bungarotoxin. *Biochem. J.* **371**, 423-427.
160. Morillas, M., Gomez-Puertas, P., Rubi, B., Clotet, J., Arino, J., Valencia, A., Hegardt, F. G., Serra, D. & Asins, G. (2002) Structural model of a malonyl-CoA-binding site of carnitine octanoyltransferase and carnitine palmitoyltransferase I - Mutational analysis of a malonyl-CoA affinity domain. *J. Biol. Chem.* **277**, 11473-11480.
161. Dumoulin, P., Ebright, R. H., Knegtel, R., Kaptein, R., Granger-Schnarr, M. & Schnarr, M. (1996) Structure of the LexA repressor-DNA complex probed by affinity cleavage and affinity photo-cross-linking. *Biochemistry* **35**, 4279-4286.
162. Aloy, P., Moont, G., Gabb, H. A., Querol, E., Aviles, F. X. & Sternberg, M. J. E. (1998) Modelling repressor proteins docking to DNA. *Proteins* **33**, 535-549.
163. Tzou, W. S. & Hwang, M. J. (1999) Modeling helix-turn-helix protein-induced DNA bending with knowledge-based distance restraints. *Biophys. J.* **77**, 1191-1205.
164. Cai, S. J., Khorchid, A., Ikura, M. & Inouye, M. (2003) Probing catalytically essential domain orientation in histidine kinase EnvZ by targeted disulfide crosslinking. *J. Mol. Biol.* **328**, 409-418.
165. Dmitriev, O. Y., Jones, P. C. & Fillingame, R. H. (1999) Structure of the subunit c oligomer in the F1F0 ATP synthase: Model derived from solution structure of the monomer and cross-linking in the native enzyme. *Proc. Natl. Acad. Sci. USA* **96**, 7785-7790.
166. You, L., Gillilan, R. & Huffaker, T. C. (2004) Model for the yeast cofactor A-beta-tubulin complex based on computational docking and mutagensis. *J. Mol. Biol.* **341**, 1343-1354.
167. Lacroix, M., Rossi, V., Gaboriaud, C., Chevallier, S., Jaquinod, M., Thielens, N. M., Gagnon, J. & Arlaud, G. J. (1997) Structure and assembly of the catalytic region of human complement protease C1r: A three-dimensional model based on chemical cross-linking and homology modeling. *Biochemistry* **36**, 6270-6282.
168. Walters, K. J., Lech, P. J., Goh, A. M., Wang, Q. & Howley, P. M. (2003) DNA-repair protein hHR23a alters its protein structure upon binding. *Proc. Natl. Acad. Sci. USA* **100**, 12694-9.
169. Varadan, R., Walker, O., Pickart, C. & Fushman, D. (2002) Structural properties of polyubiquitin chains in solution. *J. Mol. Biol.* **324**, 637-647.
170. Varadan, R., Assfalg, N., Haririnia, A., Raasi, S., Pickart, C. & Fushman, D. (2004) Solution conformation of Lys(63)-linked di-ubiquitin chain provides clues to functional diversity of polyubiquitin signaling. *J. Biol. Chem.* **279**, 7055-7063.
171. Owen, D., Lowe, P. N., Nietlispach, D., Brosnan, C. E., Chirgadze, D. Y., Parker, P. J., Blundell, T. L. & Mott, H. R. (2003) Molecular Dissection of the Interaction between the Small G Proteins Rac1 and RhoA and Protein Kinase C-related Kinase 1 (PRK1). *J. Biol. Chem.* **278**, 50578-50587.

172. McDonnell, J. M., Calvert, R., Beavil, R. L., Beavil, A. J., Henry, A. J., Sutton, B. J., Gould, H. J. & Cowburn, D. (2001) The structure of the IgE C epsilon 2 domain and its role in stabilizing the complex with its high-affinity receptor Fc epsilon Rl alpha. *Nat. Struct. Biol.* **8**, 437-441.
173. Johnson, M. A. & Pinto, B. M. (2002) Saturation transfer difference 1D-TOCSY experiments to map the topography of oligosaccharides recognized by a monoclonal antibody directed against the cell-wall polysaccharide of Group A Streptococcus. *J. Am. Chem. Soc.* **124**, 15368-15374.
174. Moller, H., Serttas, N., Paulsen, H., Burchell, J. M., Taylor-Papadimitriou, J. & Meyer, B. (2002) NMR-based determination of the binding epitope and conformational analysis of MUC-1 glycopeptides and peptides bound to the breast cancer-selective monoclonal antibody SM3. *Eur. J. Biochem.* **269**, 1444-1455.
175. Buchko, G. W., Tung, C. S., McAteer, K., Isern, N. G., Spicer, L. D. & Kennedy, M. A. (2001) DNA-XPA interactions: a P-31 NMR and molecular modeling study of dCCAATAACC association with the minimal DNA-binding domain (M98-F219) of the nucleotide excision repair protein XPA. *Nuc. Acids Res.* **29**, 2635-2643.
176. Schreiber, G. & Fersht, A. R. (1995) Energetics of Protein-Protein Interactions - Analysis of the Barnase-Barstar Interface by Single Mutations and Double Mutant Cycles. *J. Mol. Biol.* **248**, 478-486.
177. Goldman, E. R., DallAcqua, W., Braden, B. C. & Mariuzza, R. A. (1997) Analysis of binding interactions in an idiotope-antiidiotope protein-protein complex by double mutant cycles. *Biochemistry* **36**, 49-56.
178. Pielak, G. J. & Wang, X. (2001) Interactions between yeast iso-1-cytochrome c and its peroxidase. *Biochemistry* **40**, 422-8.
179. Tetreault, M., Cusanovich, M., Meyer, T., Axelrod, H. & Okamura, M. Y. (2002) Double mutant studies identify electrostatic interactions that are important for docking cytochrome c(2) onto the bacterial reaction center. *Biochemistry* **41**, 5807-5815.
180. Kersten, B., Possling, A., Blaesing, F., Mirgorodskaya, E., Gobom, J. & Seitz, H. (2004) Protein microarray technology and ultraviolet crosslinking combined with mass spectrometry for the analysis of protein-DNA interactions. *Anal. Biochem.* **331**, 303-313.
181. Benjamin, D. C., Williams, D. C., Smithgill, S. J. & Rule, G. S. (1992) Long-Range Changes in a Protein Antigen Due to Antigen-Antibody Interaction. *Biochemistry* **31**, 9539-9545.
182. Song, J. & Markley, J. L. (2001) NMR chemical shift mapping of the binding site of a protein proteinase. *J. Mol. Recognit.* **14**, 166-71.
183. Morrison, J., Yang, J. C., Stewart, M. & Neuhaus, D. (2003) Solution NMR study of the interaction between NTF2 and nucleoporin FxFG. *J. Mol. Biol.* **333**, 587-603.
184. Foster, M. P., Wuttke, D. S., Clemens, K. R., Jahnke, W., Radhakrishnan, I., Tennant, L., Reymond, M., Chung, J. & Wright, P. E. (1998) Chemical shift as a probe of molecular interfaces: NMR studies of DNA binding by the three amino-terminal zinc finger domains from transcription factor IIIA. *J. Biomol. NMR* **12**, 51-71.
185. Ramos, A., Kelly, G., Hollingworth, D., Pastore, A. & Frenkiel, T. (2000) Mapping the interfaces of protein-nucleic acid complexes using cross-saturation. *J. Am. Chem. Soc.* **122**, 11311-11314.
186. Schubert, M., Edge, R. E., Lario, P., Cook, M. A., Strynadka, N. C. J., Mackie, G. A. & McIntosh, L. P. (2004) Structural characterization of the RNase E S1 domain and identification of its oligonucleotide-binding and dimerization interfaces. *J. Mol. Biol.* **341**, 37-54.
187. Fields, B. A., Goldbaum, F. A., Ysern, X., Poljak, R. J. & Mariuzza, R. A. (1995) Molecular-Basis of Antigen Mimicry by an Anti-Idiotope. *Nature* **374**, 739-742.
188. Kraulis, P. J. (1991) MOLSCRIPT: A program to produce both detailled and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946-950.
189. Merrit, E. A. & Murphy, M. E. P. (1994) Raster3D version 2.0: A program for photorealistic molecular graphics. *Acta Cryst.* **D50**, 869-873.
190. Bressanelli, S., Stiasny, K., Allison, S. L., Stura, E. A., Duquerroy, S., Lescar, J., Heinz, F. X. & Rey, F. A. (2004) Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. *EMBO J.* **23**, 728-738.

Chapter 3.
Various strategies of using Residual Dipolar Couplings in NMR-driven protein docking: application to Lys48-linked di-ubiquitin and validation against ^{15}N-relaxation data.

Aalt D.J. van Dijk, David Fushman and Alexandre M.J.J. Bonvin, *Proteins: Structure, Function and Bioinformatics* **60** (2005), 367 - 381.

Abstract

When classical, NOE-based approaches fail, it is possible, given high-resolution structures of the free molecules, to model the structure of a complex in solution based solely on chemical shift perturbation (CSP) data in combination with orientational restraints from residual dipolar couplings (RDCs) when available. RDCs can be incorporated into the docking following various strategies: as direct restraints and/or as intermolecular intervector projection angle restraints [Meiler et al., *J. Biomol. NMR* **2000**, *16*, 245]. The advantage of the latter for docking is that they directly define the relative orientation of the molecules. A combined protocol in which residual dipolar couplings are first introduced as intervector projection angle restraints and at a later stage as direct restraints is shown here to give the best performance. This approach, implemented in our information-driven docking approach HADDOCK [Dominguez *et al.*, *J. Am. Chem. Soc.* **2003**, *125*, 1731], is used to determine the solution structure of the Lys48-linked di-ubiquitin, for which chemical shift mapping, RDCs, and ^{15}N-relaxation data have been previously obtained [Varadan *et al.*, *J. Mol. Biol.* **2002**, *324*, 637]. The resulting structures, derived from chemical shift perturbation and RDC data, are cross-validated using ^{15}N-relaxation data. The solution structure differs from the crystal structure by a 20° rotation of the two ubiquitin units relative to each other.

Introduction

In this post-genomic era, biochemical research focuses more and more on proteomics. Recent advances in high-throughput methods have provided a first glimpse of the overall structure of protein-protein interaction networks in biological systems (for a review, see ref [1] and references therein) and promise to contribute increasingly to our understanding of how proteins cooperate.

Most proteins achieve their function by interacting with other proteins and forming an active complex. The structure determination of such complexes by X-ray and/or nuclear magnetic resonance (NMR) still represents a challenging task. Because of that, *ab initio* protein-protein docking is becoming increasingly popular (for a review, see ref. [2]). Recently, the CAPRI (Critical Assessment of PRedicted Interactions) experiment [3] provided a blind test of the performance of different docking approaches (see http://capri.ebi.ac.uk). Although it is clear that considerable progress has been made in the last years, it is also obvious that docking based only on shape complementarity and/or a general force field, without any auxiliary information, is still very challenging. Reasonable solutions can however be obtained when a small amount of experimental information can be included. We recently developed an information-driven method called HADDOCK (High Ambiguity Driven protein-protein DOCKing) to dock proteins using biochemical and/or biophysical data, such as mutagenesis data or the easily available chemical shift perturbation (CSP) data from NMR titration experiments [4]. Such information allows one to map the residues at the interface of the constituents of the complex and can be used to define so-called ambiguous interaction restraints (AIRs) to drive the docking. Here the ambiguity refers to the fact that residues can be identified to be located at the interface, but that there is no explicit knowledge on the specific pairwise interactions that residues from one component form with residues from the other component. In defining the ambiguous interaction restraints one distinguishes between two types of residues: active residues that have been experimentally identified and passive residues that correspond to surface neighbors of active residues. AIRs are defined between each active residue of one component and all passive and all active residues of the other component. The use of passive residues, which introduces some fuzziness in the definition of the interface, is required to account for the fact that not all interface residues can typically be

experimentally identified. For example, not all interacting residues in a complex show significant chemical shift perturbations in NMR titration experiments. The reverse is also true: sometimes residues that are not part of the interface do show significant chemical shift perturbations as a result of propagated effects such as for example a weakening of hydrogen bond strength by a slight β-sheet opening or small conformational changes. The use of a solvent accessibility criterion allows one to filter out some of these "false-positive" residues.

Classically, NMR-structure determination is done in a lengthy process of collecting and assigning NOEs, representing short-range inter-proton distances. During the last few years, residual dipolar couplings (RDCs) and ^{15}N relaxation anisotropy have been shown to be a very useful source of long-range structural information [5-10]. In contrast to NOEs, RDCs and relaxation rates provide orientational information, which is of a global character. When it comes to complexes one way to use this orientational information is to extract from the measured RDCs or relaxation rates for each component the alignment or diffusion tensor of the whole molecule. When the eigenvalues of the tensor of interest are equal for all components of the complex, this strongly indicates that the individual proteins orient or tumble together as a single rigid entity. Then, by rotating the individual components such that their alignment or diffusion tensors become collinear, the orientation of the individual molecules with respect to each other can be obtained [11-14]. Some ambiguity, however, typically remains because of the intrinsic degeneracy of RDC and relaxation data and the lack of information about the relative positioning of the molecules. When, in addition, a limited amount of intermolecular NOE-derived distances can be used, the structure of the complex can be solved. This approach has been demonstrated for a variety of systems such as RNA [15], protein-RNA [16] and protein-protein complexes [17].

More recently approaches have emerged, which do not use NOEs but rely on chemical shift mapping to determine the interface with the advantage that also weak and transient complexes can be studied in this way. Our HADDOCK approach, which we described above, makes use of such information. Chemical shift perturbation information can also be combined with RDCs to refine the relative orientation of the components. It has been previously shown that by combining restraints from chemical shift perturbations with orientational restraints from RDCs, the structure of a complex could be determined quite accurately [18]; the chemical shift perturbation restraints were taken into account by minimizing the difference between experimental and simulated chemical shift perturbations, using the program SHIFTS [19]. In another study, the chemical shift mapping data were transformed into a set of highly ambiguous intermolecular distance restraints as proposed in HADDOCK, and combined with orientational restraints from RDCs [20]. RDC and chemical shift perturbation data have also been used *a posteriori* to filter solutions obtained from *ab initio* docking [21] with the program FTDOCK [22].

RDCs are typically used directly as restraints in structure calculation methods by defining an external alignment tensor and back-calculating the RDC-values from the angle between the internuclear vector of interest and the alignment tensor. In the case of complexes, the orientation of each component is optimized with respect to a single alignment tensor. In this way, the restraints only indirectly define the relative orientation of the components. It has been shown that, by combining RDCs in a pairwise manner, it is possible to define intervector projection angle restraints that become independent of an external alignment tensor. The latter is thus no longer needed in structure calculations [23]. These angular restraints have the advantage that they can be introduced in the initial steps of structure calculation starting from random conformations. Such an approach has, until now, never been used for docking. From a computational point of view, this should have some definite advantages. By combining RDCs from the individual components, intermolecular projection angle restraints can be generated that directly define the relative orientation of the components and can be distinguished from the intramolecular components. The use of such restraints should provide an easy way to use

data from different alignment media together or to combine RDC restraints with e.g. relaxation anisotropy restraints [24].

In this article we show that the structure of complexes can be efficiently determined from a combination of chemical shift perturbation data and residual dipolar couplings using a combined approach in which the latter are first introduced as intermolecular intervector projection angle restraints and, at a later refinement stage, as direct restraints. The resulting models are independently validated using ^{15}N-relaxation data. This allows one to assess the possible influence of conformational averaging on the calculated structures since the timescales for the relaxation data differ from that for RDCs. This approach is applied to determine the solution structure of the Gly76-Lys48 linked di-ubiquitin chain (Ub_2) for which NMR chemical shift perturbations, residual dipolar couplings and diffusion anisotropy data have been obtained previously [25]. The ubiquitin-proteasomal pathway is the major mechanism of protein degradation in eukaryotic cells, and Lys48-linked tetra-ubiquitin (Ub_4) is the minimum signal required for efficient targeting of proteins to proteasomes [26-28]. Knowledge of the structure of polyubiquitin could provide useful insights into the mechanisms of its interaction with various chain recognition factors of the proteasome. At present, there is one crystal structure of a Lys48-linked Ub_2 (accession number: 1AAR) [29] and two crystal structures of Lys48-Ub_4 (accession numbers: 1F9J [30] and 1TBE [31]) in the Protein Data Bank [32], all obtained at acidic pH (~4.5). These structures differ from each other in the relative orientation and contacts between Ub units. In order to characterize the solution conformation of polyubiquitin, Lys48-linked Ub_2 and Ub_4 and Lys63-linked Ub_2 were recently investigated using a combination of NMR techniques, including chemical shift mapping, ^{15}N relaxation and RDC measurements [25,33]. It was found that Lys48-linked chains exhibit a pH-dependent conformational switch from an open (at pH 4.5) to a closed form (pH 6.8 and higher). In the closed conformation observed under near-physiological conditions in solution, the relative orientation of the two monomers with respect to each other was found to be distinct from that in the crystal structure. However, no interdomain NOEs could be reliably detected, probably due to conformational exchange broadening of the amide signals from the interface residues. Therefore, no solution structure of Ub_2 was generated, because the relaxation- and RDC-derived orientational restraints alone do not provide information on the relative positioning of the two domains. Here we show, using the previously collected NMR data [25] at pH 6.8, that the combination of intermolecular intervector projection angle restraints derived from RDC data and AIRs derived from chemical shift perturbation data can be used to determine the first solution structure of Lys48-linked Ub_2. The ^{15}N relaxation data are used for cross-validation of our structure. Finally, the solution structure of Ub_2 generated in this way is compared to the crystal structure obtained at pH 4.5.

Materials and Methods

Docking protocol

The original HADDOCK protocol is described in ref. [4]. It uses ambiguous interaction restraints (AIRs) defined from chemical shift perturbation data obtained in NMR titration experiments. The protocol consists of three consecutive stages (for details, see ref. [4]):

(i) randomization of orientations followed by rigid body energy minimization (EM);
(ii) semi-flexible simulated annealing in torsion angle space (TAD-SA), which consists of (ii-a) a rigid body Molecular Dynamics search and first simulated annealing, (ii-b) a second semi-flexible simulated annealing during which side chains at the interface are free to move, and (ii-c) a third semi-flexible simulated annealing during which both side chains and backbone at the interface are free to move; and
(iii) final refinement in Cartesian space with explicit solvent.

The docking was performed starting from the 10 conformations in the NMR ensemble (PDB entry 1D3Z) and from the two Ub monomers taken from the crystal structure of Ub_2 (PDB entry 1AAR) following the standard HADDOCK protocol. These 12 starting conformations result in 144 pairwise combinations for docking. In the initial rigid body docking phase, 1440 structures were generated (each combination being used ten times) and the best 200 in terms of total intermolecular energy were further submitted to the semi-flexible simulated annealing and final water refinement. After water refinement, the resulting structures were clustered based on RMSD and the lowest energy structures from the lowest energy cluster were selected; clustering was done with a 1.5 Å cutoff using the backbone RMSD of both proteins calculated after positional least square fitting on the first protein only.

Use of RDC restraints for docking

The measured RDCs are given by:

$$D^i(\beta^i\alpha^i) = 0.5D_0[A_a(3\cos^2\beta^i - 1) + \tfrac{3}{2}A_r(\cos 2\alpha^i \sin^2\beta^i)] \qquad \text{(equation 1)}$$

Here A_a is the axially symmetric part of the alignment tensor, equal to $[A_{zz} - 1/2(A_{xx} + A_{yy})]$ and A_r is the rhombic component of the alignment tensor, equal to $(A_{xx}-A_{yy})$, where A_{xx}, A_{yy} and A_{zz} are the x, y and z-component of the alignment tensor, respectively; α^i and β^i are the azimuthal and polar angles of the vector for which the RDC is reported, in the frame of the alignment tensor; $D_0 = -\left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_i\gamma_j h}{2\pi^2 r_{NH}^3}$ = 21.7 kHz is the strength of the (static) dipolar coupling in the NH-pair, r_{NH} is the length of the NH-vector, μ_0 is the magnetic permeability of vacuum, γ_i is the gyromagnetic ratio of spin *i*, and *h* is Planck's constant.

In order to use RDCs for structure calculation, the alignment tensor components need first to be determined. This was done by back-calculating RDCs based on the known structure of the single domains (10 NMR structures and 2 monomers from the crystal structure) and optimizing the tensor parameters, using the software Pales[34] with the option –bestFit (which uses singular value decomposition). This fitting procedure was performed separately for the RDC data measured for the distal and proximal domains leading to 24 sets of tensor parameters. Only RDCs from residues located in secondary structure elements were used (50 in total). We then selected the tensor parameters from the structure and RDC data of the domain with the lowest Q-factor.

Table 1. Force constants used during different stages of the docking protocol.

Stage[a]	k_{sani} (kcal mol^{-1} Hz^{-2})	$k_{vean,border}$ (kcal mol^{-1})	$k_{vean,centre}$ (kcal mol^{-1})	k_{air} (kcal mol^{-1} A^{-2})
i) Rigid body EM	0.0 - 0.04	0.0 - 4.0	0.0 - 1.0	1 - 10
ii-a) SA	0.01 - 0.02	1.0 – 10	0.25 - 2.5	10 - 50
ii-b) SA	0.02 - 0.2	10 – 40	2.5 -10	50
ii-c) SA	0.2	40	10	50
iii) Water refinement	0.2	40	10	50

a) i,ii and iii refer to the rigid body, simulated annealing (SA) and water refinement stages of the protocol, respectively. K_{sani}, force constant for direct RDC restraints. $k_{vean,border}$ and $k_{vean,center}$, are the force constants for RDC restraints expressed as intervector projection angles (VEAN) for the border and central parts of the VEAN potential function, respectively (see Eq. 2). k_{air}, force constant for Ambiguous Interaction Restraints. During the SANI run, $k_{vean,border}$ and $k_{vean,center}$ were set to 0 and during the VEAN run, k_{sani} was set to 0; during the VEAN-SANI run, k_{sani} was set to 0 during stage i and ii, and $k_{vean,border}$ and $k_{vean,center}$ were set to 0 during stage iii.

Experimental RDCs were introduced in the standard HADDOCK protocol in three different ways: (a) directly, using a floating alignment tensor (SANI [35] energy term in CNS [36]) ;

(b) as intervector (NH-NH) projection angle restraints (VEAN [23] energy term in CNS); and (c), as a combination of (a) and (b). In the last case, VEAN was used in stage i) and ii) of the docking protocol (rigid body docking and semi-flexible SA) and SANI was used in stage iii), the final water refinement. The values of the force constants in the protocol are given in Table I. The use of RDCs as direct restraints in structure calculations is described in ref. [35]. The use of RDCs as intermolecular intervector projection angle restraints has been described in ref [23]; however, since it is less commonly used, we will describe it again in the following.

Residual dipolar couplings as intermolecular intervector projection angle restraints
The energy function [23] in this case is given by:

$$E = \begin{cases} k_{border}(\varphi^{ij} - \varphi^{ij}_{ext1})^2 & 0 \le \varphi^{ij} \le \varphi^{ij}_{ext1} \\ 0 & \varphi^{ij}_{ext1} \le \varphi^{ij} \le \varphi^{ij}_{ext2} \\ k_{centre}\cos^2(\pi(\frac{\varphi^{ij} - \varphi^{ij}_{ext2}}{\varphi^{ij}_{ext3} - \varphi^{ij}_{ext2}} - \frac{1}{2})) & \varphi^{ij}_{ext2} \le \varphi^{ij} \le \varphi^{ij}_{ext3} \\ 0 & \varphi^{ij}_{ext3} \le \varphi^{ij} \le \varphi^{ij}_{ext4} \\ k_{border}(\varphi^{ij} - \varphi^{ij}_{ext4})^2 & \varphi^{ij}_{ext4} \le \varphi^{ij} \le 180^0 \end{cases} \quad \text{(equation 2)}$$

Here φ^{ij} denotes the value of the projection angle between the two vectors to which the restraint is applied, in our case, NH-vectors; its boundary values are given by (note that the original version of these equations in ref. [23] is not correct; the corrected version is given in ref. [37]):

$$\cos\varphi^{ij}_{ext} = \sqrt{\frac{2(3D^i - 2D_{zz} + D_{xx} + D_{yy})}{3(\frac{3}{2}(D_{xx} - D_{yy})\cos\alpha^i - 2D_{zz} + D_{xx} + D_{yy})}}\sqrt{\frac{2(3D^j - 2D_{zz} + D_{xx} + D_{yy})}{3(\frac{3}{2}(D_{xx} - D_{yy})\cos\alpha^j - 2D_{zz} + D_{xx} + D_{yy})}}\cos(\alpha^j \pm \alpha^i)$$

$$\pm\sqrt{1 - \frac{2(3D^i - 2D_{zz} + D_{xx} + D_{yy})}{3(\frac{3}{2}(D_{xx} - D_{yy})\cos\alpha^i - 2D_{zz} + D_{xx} + D_{yy})}}\sqrt{1 - \frac{2(3D^j - 2D_{zz} + D_{xx} + D_{yy})}{3(\frac{3}{2}(D_{xx} - D_{yy})\cos\alpha^j - 2D_{zz} + D_{xx} + D_{yy})}} \quad \text{(equation 3)}$$

with
$$\alpha^i_{\min} = \tfrac{1}{2}\arccos\left(\frac{6D^i + 2D_{zz} - D_{xx} - D_{yy}}{3(D_{xx} - D_{yy})}\right)$$
$$\alpha^i_{\max} = \pi - \tfrac{1}{2}\arccos\left(\frac{6D^i + 2D_{zz} - D_{xx} - D_{yy}}{3(D_{xx} - D_{yy})}\right) \quad \text{if} \left|\frac{6D^i + 2D_{zz} - D_{xx} - D_{yy}}{3(D_{xx} - D_{yy})}\right| \le 1$$

and else: $\alpha^i_{\min} = 0, \alpha^i_{\max} = \pi$.

Here, D^i and D^j are the values of the respective RDCs of the two NH-vectors; D_{xx}, D_{yy} and D_{zz} are given by $D_0 \cdot A_{xx}$, $D_0 \cdot A_{yy}$ and $D_0 \cdot A_{zz}$, respectively. Generally, two minima are allowed, due to the well-known degeneracy of the dependency of the RDC on α^i and β^i, and correspondingly, there are four boundary values for φ^{ij}, denoted φ^{ij}_{ext1} to φ^{ij}_{ext4} in equation 2. Two force constants must be defined, one for the border potential function (k_{border}) and one for the central part between the two minima (k_{centre}).

The generation of intermolecular intervector projection angle restraints is implemented in a slightly modified version of a python script (dipolar_segid.py) kindly provided by Drs. Helen Mott and Wayne Boucher (Cambridge University). This script is distributed with the HADDOCK package (see www.nmr.chem.uu.nl/haddock).

NMR data

The NMR data used in this study include RDC values for backbone amides, ^{15}N relaxation rates R_1 and R_2 and steady-state ^{1}H-^{15}N NOEs. Ub_2 chains were assembled from unlabeled and U-^{15}N labeled Ub units, such that only one of the two Ub domains was ^{15}N labeled in each NMR sample. The NMR studies were done at pH 6.8. The details of chain assembly and experimental NMR procedures have been presented elsewhere [25]. The overall average levels of R_1 in the distal Ub appeared to be slightly (6.5%) elevated compared to the proximal domain while the R_2 values showed an opposite tendency. This indicates a slight difference in the apparent overall rotational correlation time that could be attributed to interdomain mobility in Ub_2. It is also possible that small variations in protein concentration between NMR samples, where the distal or the proximal domain in Ub_2 was isotope labeled, contribute (due to aggregation effects) to the observed differences in the relaxation rates between the domains. Ub_2 showed some tendency to aggregate at NMR concentrations (~1 mM), as inferred from the observed concentration dependence of spin relaxation rates. Note that varying protein concentration can affect the viscosity of the sample (hence the relaxation rates); however, at these low ubiquitin concentrations, this effect is expected to be negligible. Although the relaxation measurements for these studies were performed at Ub_2 concentrations as low as 250–500 μM, some amount of aggregation could still be present. Note that the subnanosecond local backbone dynamics in both Ub domains are very similar to each other and to those in the monomeric Ub [38]. In order to be able to fit the experimental data for both domains simultaneously, relaxation rates for the distal domain were uniformly rescaled as follows: $R_2 \rightarrow R_2*1.065$, $R_1 \rightarrow R_1/1.065$. This scaling factor represents an average value of the ratio of R_2s between the proximal and the distal domains and of the inverse ratio of R_1s for these domains. Analysis of the diffusion tensor data showed that this procedure mostly affects the principal values but not the orientation of the diffusion tensor. This is further supported by the agreement between the orientation of the diffusion tensors derived from these data and calculated from hydrodynamic properties of the Ub_2 structures (Supplementary Material).

The analysis of the experimental data using computer program ROTDIF [38,39] was performed including core residues, identified as those belonging to Ub secondary structure, excluding six NH groups (residues 7,23,25,48,70,71) in the distal and seven (residues 23,25,48,56,68,70,71) in the proximal Ub that show conformational exchange broadening or could not be reliably evaluated from the spectra due to signal overlap. We have also performed a similar analysis including most of the NH vectors except those mentioned above and those in the β1/β2 loop and in the C-terminus. The results of these analyses were very similar in terms of the derived diffusion tensor characteristics. Only the "core"-residue results are presented here. The ROTDIF program determines the overall rotational diffusion tensor of a molecule from the ratio $\rho = \left(\frac{2R_2'}{R_1'} - 1 \right)^{-1}$ of the rates of ^{15}N longitudinal and transverse relaxation, R_1' and R_2'; the prime indicates that these rates are modified by subtracting the contributions from high-frequency components of the spectral density [40,41]. Note that this ratio, ρ, is independent, to the first approximation, of the site-specific values of the ^{15}N CSA tensor and of the backbone order parameters. The tensor derivation is based on the minimization of the target function

$$\chi^2 = \sum_i \left(\frac{\rho_i^{\exp} - \rho_i^{calc}}{\sigma_i} \right)^2 \qquad \text{(equation 4)}$$

where the superscripts 'exp' and 'calc' indicate the experimentally measured and the calculated values of ρ for the i-th NH vector, σ_i is the experimental error in ρ_i^{exp}, and ρ_i^{calc} is calculated for a given model of the overall tumbling as described in details in ref. [38]. All

data were treated using an axially symmetric and a fully anisotropic model for the rotational diffusion tensor.

Structural coordinates

The structural coordinates for the ubiquitin monomer used as starting structure in the docking were taken from the Brookhaven Protein Data Bank [32] entry 1D3Z [42], an NMR structure of the monomer, of which the 10 models were used, and entry 1AAR [29], the crystal structure of Ub_2, which was split in 2 parts, both being used as starting structures for a Ub monomer.

The ensemble of the 10 solution structures of Ub_2 described in this study have been deposited in the PDB under accession code 2BGF, together with the various NMR restraints used for the docking.

Analysis of intermolecular contacts

Intermolecular contacts (hydrogen bonds and non-bonded contacts) were analyzed with DIMPLOT which is part of the LIGPLOT software [43] using the default settings (3.9 Å heavy-atoms distance cut-off for non-bonded contacts; 2.7 Å and 3.35 Å proton-acceptor and donor-acceptor distance cut-offs respectively with minimum 90^0 angles (D-H-A, H-A-AA, D-A-AA) for hydrogen bonds. A contact is defined to be present in the solution structure if it is found in at least four of the 10 best structures.

Results

To compare the performance of using RDCs as direct restraints or as intermolecular intervector projection angle restraints, the NMR solution structure of the Gly76-Lys48 linked Ub_2 was first calculated in three different manners with HADDOCK [4]: i) using solely chemical shift perturbation data, ii) using chemical shift perturbation data in combination with direct RDC restraints and iii) using chemical shift perturbation data in combination with intermolecular intervector projection angle restraints derived from RDC data (see Material and Methods). These three runs will be denoted in the following as i) CSP, ii) SANI and iii) VEAN, respectively (SANI and VEAN refer to the energy terms in CNS [36]). Analysis of the results led us to define a protocol termed VEAN-SANI, in which both approaches are combined, using first the intervector projection angle restraints to drive the docking and then the direct RDC restraints for the final explicit solvent refinement. In cases where ambiguity is present in the experimental data, different orientations can be obtained by docking which should then be distinguished based on energetic considerations as shown previously [4]. To assess the performance and convergence of the various protocols presented here we therefore compare the population of the lowest energy clusters.

Restraints definition

The restraints that were used are listed in Table 2. The ubiquitin monomers, which are linked to one another via a Gly76-Lys48 isopeptide bond, are designated distal (D: containing Gly76) and proximal (P: containing Lys48 and a free C-terminus), respectively. For both distal and proximal domains, the active and passive residues used in the docking were defined using the procedure described previously [4]: the residues with combined ^1H and ^{15}N CSP above average (0.033 ppm) whose backbone or side-chain showed more than 50% relative solvent accessible surface area were defined as active residues. Their solvent accessible surface neighbors were defined as passive residues. From the list of active and passive residues for both monomers, 21 ambiguous interaction restraints (AIRs) were defined with an upper distance bound of 2Å.

Table 2. Data used in the docking.

Distal domain[a]	
Active residues[b]	L8, T9, A46, G47, K48, Q49, E51, H68, R72, L73
Passive residues[b]	K6, G10, K11, T12, D39, D52, G53, R54, L71, R74, G75, G76
RDCs	46 NH RDCs
Flexible segments[c]	F4-T14; P37-L56; T66-L71
Fully flexible segments[d]	R72-G76
Proximal domain[a]	
Active residues[b]	L8, T9, G47, K48, E51, H68, V70, R72, L73, R74, G76
Passive residues[b]	K6, G10, K11, T12, D39, A46, Q49, D52, G53, R54, L71, G75
RDCs	46 NH RDCs
Flexible segments[c]	F4-T14; P37-Q41; F45-L56; T66-L71
Fully flexible segments[d]	K48; R72-G76
Intervector projection angle restraints[e]	*Number of restraints*
Intermolecular	981
Intramolecular	972
Isopeptide bond (G76-K48)[f]	*Unambiguous restraint distance (Å)*
O-NZ	2.25 +/- 0.05
C-NZ	1.35 +/- 0.05
C-CE	2.45 +/- 0.05
CA-NZ	2.45 +/- 0.05

[a] The ubiquitin monomers, which are linked to one another via a Gly76-Lys48 isopeptide bond, are designated distal (D: containing Gly76) and proximal (P: containing Lys48 and a free C-terminus), respectively.
[b] Our docking protocol uses active residues (interface residues defined based on experimental CSP data (see text)) and passive residues (surface neighbors of passive residues). Ambiguous interaction restraints, AIRs (21), are defined from each active residue of one domain to all active and passive residues of the other domain, using a 2Å upper distance bound.
[c]Flexible segments are the parts of the molecules that are free to move during the SA stage ii-b and ii-c (see text). These are typically defined as all active and passive residues plus two sequential neighbors.
[d]Fully flexible segments are the parts of the molecules that are free to move during all SA stages (see text).
[e]One docking run was performed using only Chemical Shift Perturbation data (CSP). RDCs were used either as direct restraints (SANI) or as intervector projection angle restraints (VEAN), in both cases combined with CSP. The final protocol consists of a combination of VEAN and SANI together with CSP (VEAN-SANI); see text for details.
[f]The distal and proximal domains are connected via an isopeptide bond between Gly 76 carbonyl C and Lys 48 NZ; in the docking this is represented as a set of distance restraints based on typical distances for this bond in the crystallographic structure.

Two sets of each 46 NH-RDCs (for the distal and proximal domains) were available. As starting structures, we used the 10 models from the NMR structure of ubiquitin (1D3Z) and the two units of the Ub_2 X-ray structure (1AAR). For each of those 12 starting structures the principal components of the alignment tensor were calculated using Pales [34] for each RDC set independently. For this, only residues located in secondary structure elements were considered (50 for the two units together). The goodness of fit was assessed with the RDC Q-factor [42]. Model 8 from the NMR ensemble, fitted to the RDC-data for the distal domain, gave the lowest Q-factor (Q=0.21; R=0.96); the resulting principal component values (A_{xx}= 32.3 10^{-5}, A_{yy}= 38.3 10^{-5}, A_{zz}=-70.6 10^{-5}; A_a=-105.9 10^{-5}, and A_r=-6.02 10^{-5}) were used subsequently in the structure calculations. The RDC data were introduced either as direct restraints (92 in total) with explicit inclusion of an alignment tensor (SANI) or as intervector

projection angle restraints (VEAN). The latter were defined by taking pairwise combinations of the total set of 92 residual dipolar couplings and selecting only the intervector projection angle restraints restricting at least 25% of the conformational space. In this way, 981 inter-domain and 972 intra-domain restraints were defined. Furthermore, 4 unambiguous restraints were introduced to represent the covalent isopeptide bond from Gly76 of the distal domain to Lys48 of the proximal domain. This approach was chosen instead of defining a covalent bond to allow for randomization of the starting orientations and to facilitate the docking.

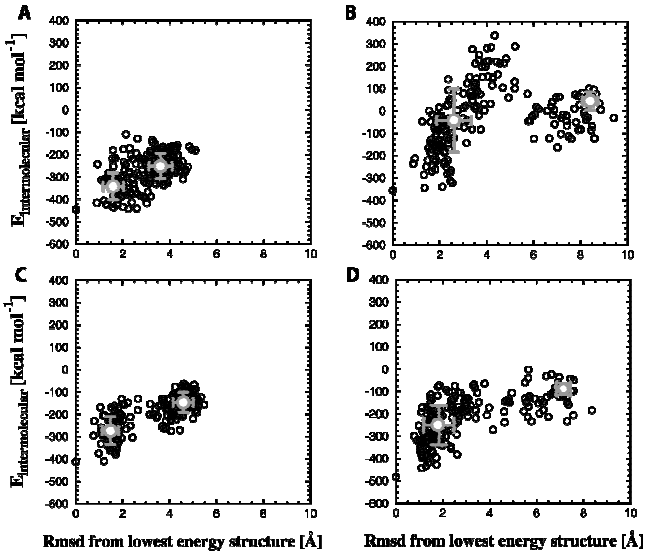


Figure 1. Intermolecular energies (E_{vdw}+E_{elec}+$E_{restraints}$) versus backbone RMSD from the lowest energy structure. Values for single structures are indicated by black dots and cluster averages and standard deviations in gray; (a) CSP; (b) VEAN; (c) SANI; and (d), final protocol, VEAN-SANI. For a description of the different docking runs, see text and footnote e of Table 2.

Docking from CSP

Figure 1 shows the intermolecular energies of the water-refined solutions as a function of the RMSD from the lowest energy structure of the respective runs. The 200 solutions after final refinement in explicit water were clustered using a 1.5 Å pair-wise RMSD cutoff and clusters were ranked using the average intermolecular energies calculated for the 10 lowest energy structures. In this way, different clusters sizes do not affect the ranking. After clustering, the lowest energy structure(s) of the lowest energy cluster were taken as the best docking solution. The lowest energy cluster for the CSP-docking has 39 members. As shown in Table 3, the Q-factor for this run is relatively high (0.64 +/- 0.07), indicating that those structures do not satisfy the RDC restraints. The CSP-structures are quite similar to the crystal structure, with an average backbone RMSD from it of 1.0 Å (Table 3). The crystal structure itself has a Q-factor of 0.44, which indicates that it is also not consistent with the experimental RDCs.

Docking from CSP and intermolecular intervector projection angle restraints (VEAN)

As is clear from Figure 1b, the VEAN run has a very good convergence; the lowest energy cluster is highly populated with 107 members. Note that the energies in Figure 1 contain the restraint energies (AIR, VEAN, SANI), which explains why they differ between the various runs. The corresponding VEAN-structures have a better Q-factor (0.35 +/- 0.03) than the CSP-only structures and the crystal structure. As can be seen in Table 3, the number of direct RDC (which were not used) or intervector projection angle restraint violations is lower than for the crystal or the CSP-only docked structure. The structures differ from the CSP-structures (backbone RMSD from average CSP structure 1.5 +/- 0.3 Å) and from the crystal structure

(backbone RMSD 2.4 +/- 0.5 Å). These differences are larger than the spread within the ensemble of 10 structures (see Table 4).

Table 3. Consistency of docked Ub_2 structures with RDC-data.

	Crystal (1AAR)	Solution structures[a]			
		CSP	SANI	VEAN	VEAN-SANI
Q-factor[b]	0.44	0.64 (0.07)	0.14 (0.01)	0.35 (0.03)	0.15 (0.01)
RDC viol[c] > 1 Hz	78	77 (3)	38 (5)	63 (3)	41 (4)
RDC viol[c] > 3 Hz	48	51 (6)	4 (1)	31 (3)	6 (2)
VEAN inter viol[d] > 10^0	148	154 (13)	62 (5)	29 (10)	61 (6)
VEAN intra viol[d] > 10^0	71	123 (12)	56 (4)	18 (8)	54 (3)

[a]For description of different docking runs, see footnote e of Table 2. The statistics are for the 10 lowest energy structures of the lowest energy cluster. See also Figure 2 for a histogram of deviations from experimental RDC values.
[b]Q-factor=rms(D^{calc}-D^{obs})/rms(D^{obs})[42], where D^{calc} and D^{obs} are calculated and observed RDCs, respectively.
[c]RDC viol.: number of violations of direct RDC restraints (using cutoffs of 1 and 3 Hz, respectively); total number of direct RDC restraints is 92.
[d]VEAN inter viol and intra viol: number of violations of intermolecular and intramolecular VEAN-restraints, respectively (using cutoff of 10^0). The VEAN restraints consist of 981 inter- and 972 intra-molecular restraints.

Docking from CSP and direct RDC restraints (SANI)
Figure 1c shows that the SANI-run does not converge as well as the VEAN run; there are only 65 structures in the lowest energy cluster. The Q-factor is low, 0.14 +/- 0.01, which indicates excellent agreement with the experimental RDC-data. Table 3 shows that the SANI structures have less direct RDC violations than the VEAN structures, but more intervector projection angle restraint violations (although still less than the CSP-only docked structures or the crystal structure). This indicates that the two kinds of RDC restraints do not define or restrict the orientation of the N-H vectors in a fully similar way. The average backbone RMSD of the SANI structures from the crystal structure is 1.6 +/- 0.3 Å and from the average CSP structure 1.4 +/- 0.3 Å.

Table 4. Average pairwise RMSD [Å] between the crystal structure and the various docked structures.[a]

	Crystal	CSP	VEAN	SANI	VEAN-SANI
Crystal	--	1.0 (0.2)	2.4 (0.5)	1.6 (0.3)	1.7 (0.3)
CSP	1.4 (0.2)	0.8 (0.2) / 0.9 (0.2)	1.5 (0.3)	1.4 (0.3)	1.5 (0.3)
VEAN	2.7 (0.5)	1.8 (0.4)	0.8 (0.3)/0.9 (0.3)	1.0 (0.2)	1.0 (0.3)
SANI	2.0 (0.3)	1.6 (0.4)	1.3 (0.2)	0.7 (0.2)/0.9 (0.2)	0.9 (0.2)
VEAN-SANI	2.0 (0.4)	1.8 (0.3)	1.2 (0.2)	1.2 (0.2)	0.7 (0.2) / 0.8 (0.3)

[a]Upper right: backbone RMSD; lower left: all-heavy atom RMSD. The RMSD values are calculated for the respective ensembles of structures as the averages from the respective average structure. Standard deviations are indicated between brackets. Residues 1 to 71 were used to calculate RMSD for both distal and proximal domain (excluding residues 72-76, which are a flexible tail). For a description of the different docking runs, see footnote e of Table 2.

Comparison of direct RDC (SANI) and intervector projection angle (VEAN) run
The VEAN-structures and the SANI-structures both are more consistent with the experimental RDC data than the crystal structure or the CSP-structures. Still, there are differences between those structures, reflected in an average backbone RMSD from the mean of VEAN from SANI of 1.0 +/- 0.2 Å. To compare the consistency of both sets of structures with the RDC data in more detail, a histogram of the differences between calculated and observed RDCs is

presented in Figure 2. Both runs show a Gaussian-like distribution around a mean value of 0, but the distribution for SANI is tighter than the one for VEAN. The maximum violation in SANI is 4.2 Hz, which, considering the 50 Hz range of the measured RDCs, is below 10% of the experimental values. Note also that we are using a flat bottom harmonic potential for the RDCs that allows larger deviations from experimental values than the often used purely harmonic potential.

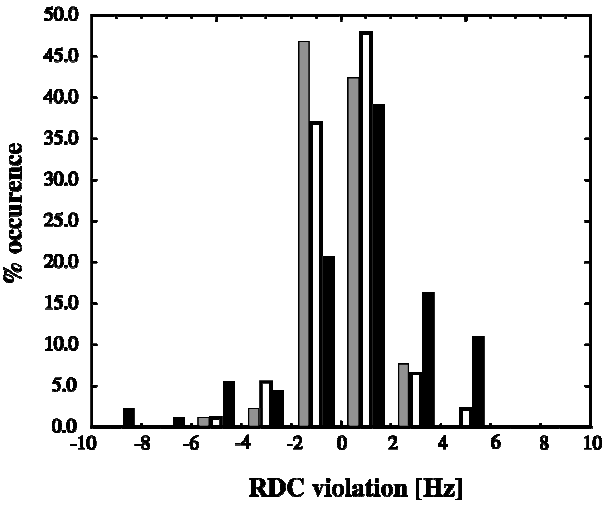


Figure 2. Histogram of differences between experimental and calculated (best-fitted using the program Pales[34]) RDC values for the structures from the different docking runs using RDCs (open black bar: SANI; filled black bar: VEAN; filled grey bar: VEAN-SANI; see footnote e of Table 2). The bin size used is 2 Hz.

Final protocol: docking with CSP and combination of VEAN and SANI

From the results described above it is clear that VEAN has the advantage of better convergence (107 versus 65 structures for SANI in the lowest energy cluster), but that SANI shows better consistency with the experimental RDC-data. To combine those advantages, we defined a protocol in which we use intervector projection angle restraints (VEAN) in the first two stages of our docking protocol (rigid body EM and semi-flexible SA) and direct RDC restraints (SANI) in the final water refinement (for details, see Material and Methods); we refer to this as VEAN-SANI. In Figure 1d, it can be seen that the convergence of this protocol is comparable to that of VEAN; the lowest energy cluster is the most highly populated with 129 members. Moreover, the structures satisfy the RDC restraints as well as the SANI structures: the Q-factor for the best 10 structures after water refinement is 0.15 +/- 0.01; see also Table 3 and Figure 2. The structures have an average backbone RMSD from the average structure of 1.0 +/- 0.3 Å to VEAN and 0.9 +/- 0.2 Å to SANI, respectively. These differences are within the precision of the respective ensembles, indicating that the various structures calculated using RDCs (SANI, VEAN and VEAN-SANI) are to a major extent similar.

Figure 3 shows a plot of the experimental versus calculated RDC values for the crystal and the VEAN-SANI structures, together with both structures color-coded according to the differences between experimental and calculated RDCs. From these data, it is clear that our solution structure has a much better consistency with the RDC data. It can also be seen that the deviations in the crystal structure are not only present in loop regions but also in secondary structure elements, indicating that it is unlikely that they could originate from intramolecular flexibility only.

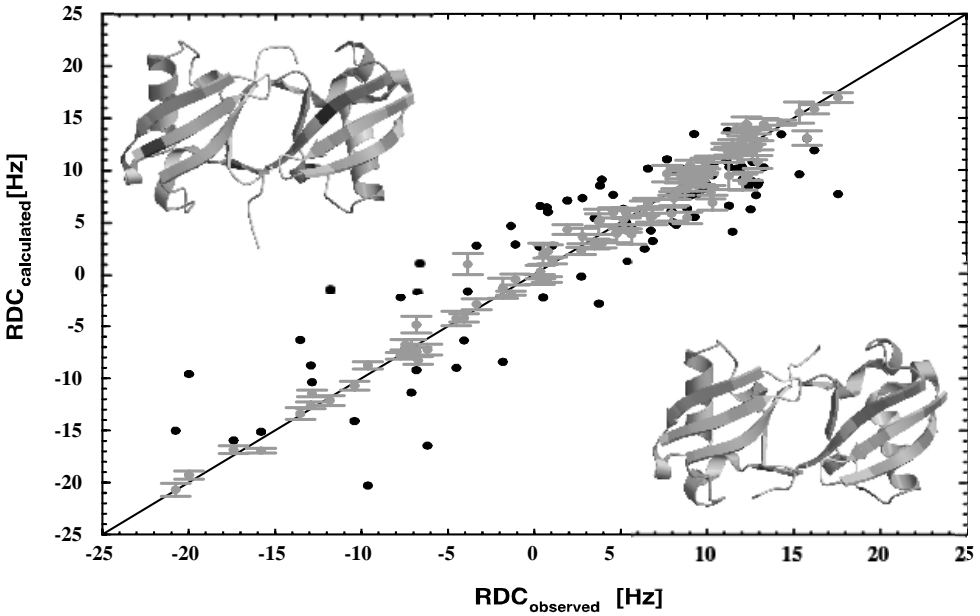


Figure 3. Comparison of the experimentally measured dipolar couplings $RDC_{observed}$ with the back-calculated values using the program Pales ($RDC_{calculated}$) for the ensemble of the best 10 solution structures (grey, error bars are indicated) and for the crystal structure (black). Upper left corner: Mapping on the crystal structure of the relative difference between experimental and calculated value of RDC ($abs(D^{obs}-D^{calc})/max[1.0,abs(D^{obs})]$. Residues are color-coded on a scale from light grey (relative difference 0.0) to dark grey (relative difference 6.1). White indicates that no RDCs were measured. The isopeptide bond is indicated in ball-and-stick. Lower right corner: idem, but for the solution structure (closest to the mean).

Validation using ^{15}N relaxation data

As an independent validation of the Ub_2 structures we determined how well they agree with the ^{15}N relaxation data. The crystal structure of Ub_2 and the ten final Ub_2 structures of each run were used as input for the computer program ROTDIF [25,38,39] that fits relaxation data to a fully anisotropic rotational diffusion model. The experimental data were also fitted for each domain separately, as control. The results are summarized in Table 5; details are presented in the Supplementary Material (see Appendix).

The orientation of the principal axes frame of the diffusion tensor with respect to the molecule is very similar for SANI, VEAN, and VEAN-SANI ensembles (see Figure 4). The CSP structures and the crystal structure resulted in a somewhat different orientation of the diffusion tensor, consistent with the different interdomain orientation for these two structures. It is worth mentioning that the Ub_2 diffusion tensor orientation derived from ^{15}N relaxation data is in reasonable agreement with the tensor orientation predicted solely from hydrodynamics properties of our solution structures using HYDRONMR [44] (Supplementary Material).

We have also compared the orientation of the diffusion tensor axes of Ub_2 obtained from fitting both domains simultaneously with those derived from fitting each individual domain alone. For a proper alignment of the domains in a molecule, these three sets of axes should agree with each other. There is an excellent agreement between the "both-domains"-axes and those for the proximal domain in all NMR-derived structures (Table 5). For the distal domain, the agreement is very good for the VEAN-SANI ensemble, while there is a somewhat greater variation in the axes orientation for the other structures, with CSP showing the least agreement. Note that the diffusion tensor for the distal domain appears more axially

symmetric [25], and therefore the orientation of its x- and y-axes (given by the angle γ) is less well defined, resulting in a greater variation in the γ angle among the structures. To better assess the agreement between the orientations of the diffusion tensors, we also determined the angles between the corresponding axes for every structure in each ensemble (Supplementary Material). The average values of the intervening angle between the z-axes of the diffusion tensors ("both-domains" *versus* "distal-only") were 6±2° (VEAN-SANI), 8±1° (SANI), 8±3° (VEAN), and 14±5° (CSP), while the corresponding angles between the "both-domains" and the "proximal-only" diffusion tensors were 5±1° (VEAN-SANI), 4±1° (SANI), 7±5° (VEAN), and 13±6° (CSP). A comparison between the axes of the "distal-only" and the "proximal-only" diffusion tensors gives for the same intervening angle 9±3° (VEAN-SANI), 10±1° (SANI), 13±6° (VEAN), and 24±9° (CSP). For the crystal structure, the z-z angles were 14°, 16°, and 28°, respectively.

Table 5. Summary of the results of ROTDIF analysis of ^{15}N relaxation data for Ub_2, applied to both domains simultaneously and to each Ub domain separately.

	Both domains					Distal					Proximal					$\chi^2_{D\&P}/[\chi^2_D+\chi^2_P]$ [d]	F [e]
Str	α [a]	β [a]	γ [a]	χ^2 [b]	χ^2/df [c]	Str	α [a]	β [a]	γ [a]	χ^2 [b]	χ^2/df [c]	Str	α [a]	β [a]	γ [a]		χ^2/df [c]
CSP	-16 (3)	90 (3)	3 (4)	330.1 (60.7)	4.8 (0.9)	-23 (3)	103 (6)	-49 (27)	145.9 (19.6)	4.6 (0.6)	-4 (8)	90 (2)	8 (3)	90.2 (19.3)	2.9 (0.6)	1.39 (0.16)	1.27 (0.14)
VEAN	15 (4)	85 (2)	182 (10)	330.0 (46.9)	4.8 (0.7)	14 (4)	77 (3)	160 (29)	134.2 (26.6)	4.2 (0.8)	13 (8)	87 (4)	181 (10)	127.1 (40.3)	4.1 (1.3)	1.27 (0.10)	1.16 (0.09)
SANI	18 (2)	87 (3)	178 (5)	232.2 (27.6)	3.4 (0.4)	18 (2)	79 (3)	158 (28)	115.5 (8.5)	3.6 (0.3)	14 (3)	88 (3)	178 (3)	61.2 (6.0)	2.0 (0.2)	1.31 (0.08)	1.20 (0.08)
VEAN -SANI [f]	97 (2)	90 (2)	120 (3)	230.2 (13.4)	3.3 (0.2)	103 (2)	87 (3)	109 (9)	109.1 (8.8)	3.4 (0.3)	94 (2)	87 (2)	120 (3)	81.3 (10.2)	2.6 (0.3)	1.21 (0.07)	1.11 (0.06)
Cryst	101	87	163	329.2	4.8	113	79	130	105.4	3.3	85	84	173	88.7	2.9	1.70	1.55

Shown are average values of the parameters over the corresponding ensemble of 10 selected structures (except for the crystal structure), the numbers in the parentheses represent the standard deviations within each ensemble.

[a] Euler angles {α, β, γ}, in degrees, determine the orientation of the diffusion tensor frame with respect to the molecular coordinate frame of the specified Ub_2 structure.

[b] The values of the target function representing the residuals of fit using ROTDIF, see Equation 4.

[c] The value of the target function per degree of freedom.

[d] The ratio of χ^2 value from the simultaneous fit of both Ub domains, $\chi^2_{D\&P}$, to the sum of the residuals of fit when treating the two Ub domains separately (χ^2_D and χ^2_P). A value of this ratio close to 1 indicates that the relative orientation of the two domains in the analyzed structure agrees well with the relaxation data.

[e] The F-test value determined here as a ratio of the variance from the simultaneous fit of both domains, $var(D\&P) = \chi^2_{D\&P}/df_{D\&P}$, to the combined variance from fitting the individual domains, $var(D+P) = [\chi^2_D+\chi^2_P]/[df_D+df_P]$, where df represents the number of degrees of freedom. At a 95% confidence level, the critical F-test value is 1.50755, for $df_{D\&P} = 69$, $df_D+df_P=63$.

[f] Note that the difference in angles of the VEAN-SANI structures with respect to the SANI and VEAN structures is only the result of a different orientation of the input coordinates and that the diffusion tensors for the three sets of structures are nicely collinear (see Figure 4).

The structures were then compared in terms of the residuals of fit, represented by the value of the target function, χ^2 (Table 5). The average χ^2 value per degree of freedom, χ^2/df, for the core residues was 4.78 ± 0.88 (SD) for CSP, 4.78 ± 0.68 for VEAN, 3.37 ± 0.40 for SANI, and 3.34 ± 0.19 for VEAN-SANI. These values should be compared with 4.77 for the crystal structure. In order to assess how well the relative orientation of the two domains in the derived structures agrees with the ^{15}N relaxation data, we also compared the residuals of fit for Ub_2 with the sum of χ^2 values obtained when fitting the same experimental data for the

two Ub domains separately. Such a comparison is necessary because the total χ^2 could be affected by greater residuals of fit for the individual domains (as e.g. in the case of the distal domain).

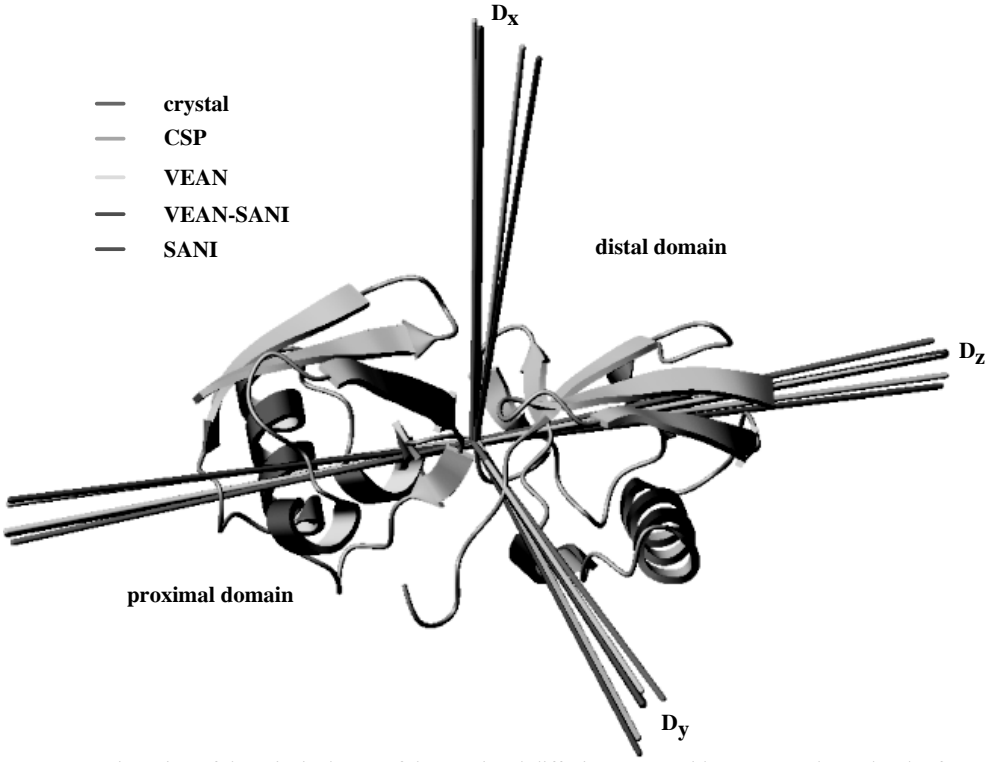


Figure 4. Orientation of the principal axes of the rotational diffusion tensor with respect to the molecular frame of Ub_2 for the various structural models discussed here. The diffusion tensors were derived from ^{15}N relaxation data using the computer program ROTDIF [25,38,39]. The rods represent the average orientation of the tensor axes (X, Y, and Z indicated here as D_x, D_y, and D_z) for each ensemble of the docked structures and for the crystal structure. The Ub_2 structures from these ensembles were superimposed onto the representative VEAN-SANI structure (shown here as a ribbon cartoon) using the backbone atoms of the residues belonging to the elements of secondary structure. The figure was generated using MolMol [55].

The ratios of the corresponding χ^2 values (Table 5) indicate that all docked structures are in a better agreement with the experimental relaxation data than the crystal structure. A statistical F test [45] comparing the corresponding variances for CSP, VEAN, SANI, and VEAN-SANI, results in F-test values (Table 5, last column) below the critical value of 1.50755 at the 95% confidence level. This suggests that an increase in χ^2 when fitting both domains according to their relative orientation in a given structure *versus* fitting them separately is not statistically significant for the docked structures. In comparison, the F-test value for the crystal structure suggests that the observed increase in χ^2/df could be significant: the probability for such an increase to occur by chance is 0.04. The VEAN-SANI ensemble, which has the lowest χ^2/df values, agrees best with the experimental relaxation data. The SANI ensemble is of comparable quality (although it has a higher standard deviation), while the VEAN ensemble is in slightly lesser agreement with the ^{15}N relaxation data, although it is still somewhat better than the crystal structure and the CSP-only structures.

Structural analysis of the Ub_2 solution structure: quality of the structure

Based on the good convergence, the consistency with the RDC data and the independent validation with the relaxation data, we conclude that the VEAN-SANI structures are our best representation of the solution structure of Ub_2. The final ensemble of the ten lowest-energy structures is shown in Figure 5. In Table 6 we analyze the solution structure and compare it with the crystal structure (1AAR [29]). The solution structure has only a small number of violations of AIR restraints and RDC data. This is not true for the crystal structure, which has 10 violations of ambiguous interaction restraints and has a higher Q-factor (see Table 3).

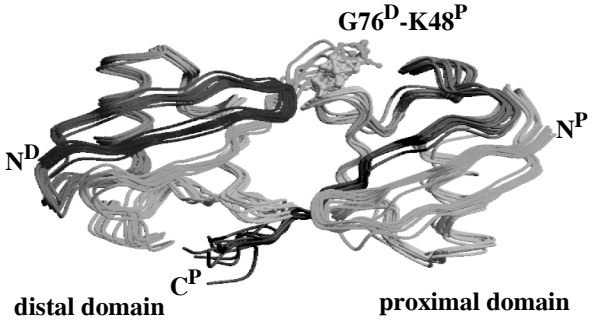


Figure 5. Ensemble of 10 lowest energy structures from the lowest energy cluster for the solution structure (calculated using the final VEAN-SANI protocol) of Ub_2. The structures are fitted on the interface residues (see Table I) their orientation is the same as in Figure 3. The isopeptide bond and the N- and C-terminal residues of the distal (D) and proximal (P) domains are indicated. This figure was generated using MolScript [56] and Raster3D [57].

Table 6. Structural statistics of the Ub_2 solution structure and comparison with the crystal structure (PDB entry 1AAR [29]). [a]

	1AAR	Solution structure
Docking statistics		
E_{vdW} [b] (kcal/mol)	--	-61 (9)
E_{elec} [b] (kcal/mol)		-425 (35)
Cluster-population[c]	--	129
AIR-energy (kcal/mol)	--	0.4 (0.9)
AIR-violations > 0.5 Å	10	0.3 (0.6)
AIR RMS (Å)	--	0.08 (0.1)
Structural statistics		
RMSD backbone[d] (Å)	--	0.7 (0.2)
BSA[e] (Å^2)	1528	1749 (54)
Helix-angle[f] (^{0})	128	154 (6)
RMS Deviations from ideal geometry		
angles (^{0})	3.2	0.43 (0.01)
bonds (10^{-3} Å)	18	3 (0.1)
impropers (^{0})	1.64	0.47 (0.02)
Ramachandran analysis (%)		
- Most favored	93.9	87.0
- Additional allowed	6.1	13.0
- Generously allowed	0.0	0.0
- Disallowed	0.0	0.0

[a] The reported values for the solution structure are averages over the 10 final structures with standard deviation indicated between brackets.
[b] The non-bonded energies Evdw and Eelec were calculated with an 8.5 Å distance cutoff using the OPLS nonbonded parameters [53] from the parallhdg5.3.pro parameter file [54].
[c] Number of structures in the lowest energy cluster (out of a total of 200 structures).
[d] Average RMSD from the average structure.
[e] Buried surface area (calculated with NACCESS [48]).
[f] The helix-angle is defined as the angle between the helices of the distal and proximal domain, calculated from the helix axis of residues 24-33 of each domain.

To make sure that the solutions resulting from the docking with RDC-restraints, while in better agreement with the RDC and relaxation data, are not artificial in the sense that their geometry is worse, we analyzed the RMSD from idealized covalent geometry and the

Ramachandran plot statistics with Procheck [46]. As shown in Table 6, the results are satisfactory. Although the crystal structure has a somewhat higher percentage of backbone dihedrals in the most favored region of the Ramachandran plot, the solution structure is still very good, with 100% of the residues having their backbone dihedrals in the most favored or additionally allowed regions of the Ramachandran plot. The RMSDs from idealized geometry are also very good, indicating that the additional orientational restraints do not result in local distortions of the covalent geometry. We also checked if the improvement in Q-factor for the structures docked using RDC-restraints was not due to local distortions of N-H bonds; this was done by removing the hydrogens and adding them again in CNS without using experimental restraints. The Q-factor of the structures generated in this way was on average 0.19, compared to 0.15 for the refined structure, indicating again that the use of RDCs does not artificially distort the geometry of the structure (otherwise it would be expected that adding the N-H protons according to standard geometry would give much higher Q-factors).

Structural analysis of the Ub_2 solution structure: orientation of the two domains

A rotation of the two domains (distal and proximal) relative to each other is observed when comparing solution structure and crystal structure. This is confirmed by analysis with Dyndom [47], which allows to identify and characterize domain motions. The rotation axis is 40% (+/- 10%) a twist axis (normal to the interface) and 60% +/- 10% a closure axis (parallel to the interface); the rotation around this axis is 20^0 (+/- 3^0). Figure 6 gives a graphical representation of this rotation; the rotation axis as determined by Dyndom for the representative structure (closest to the mean) of the ensemble is shown, together with this structure and the crystal structure. We also calculated the angle between the helix axes in the distal and proximal domain: this angle is 128^0 in the crystal structure and 154^0 ($+/-6^0$) in the solution structure. This difference in domain orientation between solution and crystal structure is also reflected in the RMSD value of 1.7 Å +/- 0.3 Å between the two.

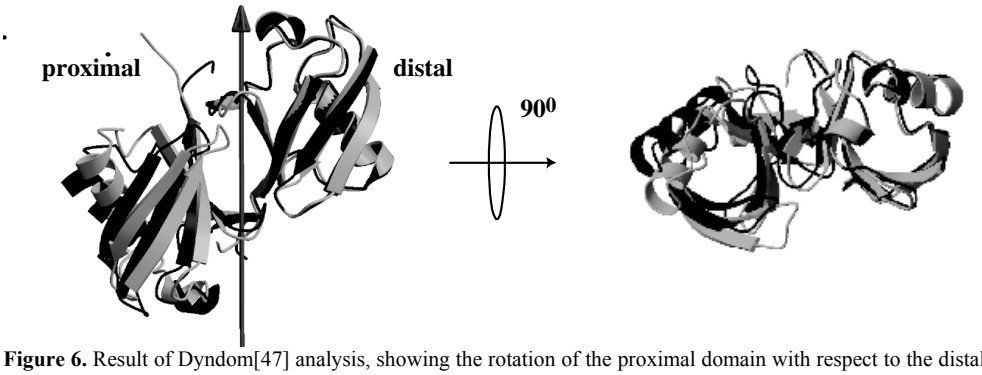


Figure 6. Result of Dyndom[47] analysis, showing the rotation of the proximal domain with respect to the distal domain when comparing the representative solution structure (black) with the crystal structure (grey). The structures are fitted on the distal domain, and secondary structure elements are indicated. Two orthogonal views are shown, corresponding to a 90^0 rotation around a horizontal axis in the plane of the paper. The rotation axis as determined by Dyndom is indicated. This figure was generated using MolScript [56] and Raster3D [57].

Structural analysis of the Ub_2 solution structure: the interface

It was previously shown based on NMR data [25,38], that the functionally important residues Leu8, Ile44 and Val70 at the ubiquitin Ub_2 interface are not rigidly buried but that some interdomain dynamics is retained that could be sufficient to expose these groups and allow them to interact with other molecules. It was also suggested that the interface in the closed conformation (at pH 6.8, for which we determined the structure here) is somewhat more open than that in the crystal structure.

Our solution structure supports this latter conclusion. Specifically, we looked at the percentage of decrease in solvent accessibility (calculated using NACCESS [48]) observed upon going from the ubiquitin monomer to the Ub_2 structure. For Val70 (both distal and proximal domain) there is not much difference between the solution and the crystal structure; for both, the decrease in accessibility is between 95 and 100%. For Ile44 (both distal and proximal domain), in the crystal structure the accessibility decreases by more than 99%, whereas in the solution structure, Ile44 remains a little bit more accessible (decrease in accessibility of 94% +/- 7%). The accessibility of Leu8 in the crystal structure is decreased by 65% and 70% for the distal and proximal domains, respectively, compared with only 52% +/- 11% and 58% +/- 7% for these domains in the solution structure. This shows that in the solution structures these hydrophobic residues are slightly more accessible than in the crystal structure. Note that these data also suggest that the interactions between the functionally important hydrophobic residues (Leu8, Ile44 and Val70) in Ub_2 and various recognition factors will involve significant interdomain motions, e.g. a transition between the closed conformation (which we study here) and open conformations of Ub_2.

We also analyzed the intermolecular contacts using the program DIMPLOT [43]. A number of non-bonded contacts were detected which are similar in the solution and the crystal structure. There are however additional contacts that are specific for the solution structure (see Supplementary Material Table 3). Figure 7 shows a detailed view of the interface of the solution structure.

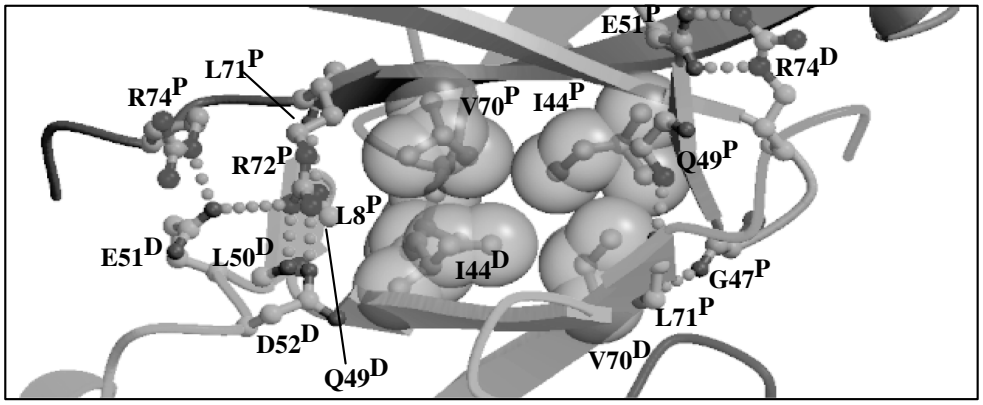


***Figure 7.** Detailed view of the interface of the Ub_2 solution structure. Residues involved in hydrophobic non-bonded contacts (ball-and-stick and transparent CPK representation) or in inter-domain hydrogen bonds or salt-bridges (ball-and-stick representation) are shown (see also Supporting Table 3; note that for a better visualization not all contacts are shown). Dotted lines represent hydrogen bonds. The residues are labeled with one-letter residue code and residue number, followed by D or P to indicate the distal or proximal domain, respectively. This figure was generated using MolScript [56] and Raster3D [57].

Discussion

The structures obtained by docking using only CSP data did not satisfy RDC nor relaxation data. To some extent, this indicates that using solely chemical shift perturbation data to drive protein-protein docking should be done with care. However, it is still reassuring that the resulting CSP structures are quite close to the crystal structure (RMSD 1.0 Å). In our opinion, this shows that the (easily available) CSP-data are very valuable to get a fair representation of the solution structure of a protein-protein complex.

When introducing RDCs in the docking protocol, both SANI and VEAN gave good and comparable results for Lys48-linked Ub_2. In order to get an unbiased account of the way

the structures fulfill the experimental RDC data, cross-validation was performed: 10% of the RDC data was excluded from the structure calculations; these data were subsequently used for independent validation. After cross-validation, the values of the Q-factor were 0.48 for SANI and 0.58 for VEAN. This shows that the direct RDC structure and the intermolecular intervector projection angle structure are of comparable quality. However, proper assessment would require full cross-validation. This was not done because it is computationally very expensive. We also checked whether the higher Q-factors of the structures calculated with intervector projection angle restraints compared to those obtained with direct RDC restraints were due to the force constants for the VEAN energy term being too small: we repeated the docking with a two times larger value of the VEAN force constants. This did not result in a better agreement with the experimental data (results not shown), indicating that looser restraining might well be an intrinsic property of this type of restraints: the accumulation of errors when taking pairs of RDCs could lead to less restrictive restraints.

The two approaches (SANI and VEAN) were combined, using intervector projection angle restraints in the first stages of the calculations to benefit from better convergence, and direct RDC restraints in the final explicit solvent refinement to fine-tune the structures. The structures calculated with this final protocol better fit the experimental RDC-restraints compared to the structures calculated using only chemical shift perturbations or the crystal structure. This is to be expected since those structures have been optimized against these RDC data. The anisotropic relaxation data provide, however, an independent validation; the structures calculated using the RDCs also better fit those data. It was estimated previously [25], on the basis of RDC and relaxation data, that the angle between the α-helices in the two domains should be between 161^0 (relaxation data) and 145^0 (RDC data) for the structure in solution, compared to 128^0 in the crystal. In agreement with this, the structures obtained here have an average angle between the α-helices of 154^0.

As previously discussed[25,38], the interface in Lys48-linked Ub_2 is not rigidly locked, which allows ligands to compete with the Ub-Ub interactions [49]. A model based on chemical shifts titration suggests that at pH6.8 Lys48-linked Ub_2 is in dynamic equilibrium between the predominantly populated (>85%) "closed" state and a minor, "open" state. This equilibrium to some extent represents the on/off events in protein-protein interactions. It is important to note that the so-called "open" state of Lys48-linked Ub_2 at pH6.8 is a compact conformation, probably characterized by somewhat weaker interdomain contacts compared to the "closed" state but not very different structurally. This is supported by recent small angle X-ray scattering data [50] that indicate a mono-dispersed compact globular conformation of Lys48-linked Ub_2 under similar conditions in solution.

From the RDC data alone it is difficult to determine whether this solution structure at pH 6.8 represents a single static conformation or results from motion averaging of RDCs; a similar problem has been noticed in the case of other proteins [14,51]. It is not clear to what extent the measured RDCs are affected by the dynamic equilibrium in Ub_2. Several interface residues in Ub_2 exhibit line-broadening indicative of conformational exchange in a μs-ms time range [25,38], although it is not possible to determine the amplitudes of these motions from the relaxation data. An estimate of the effect of motions on RDC measurements can be obtained by calculating the 'internal generalized degree of order' as proposed by Tolman *et al.*[52]; its value is correlated with possible structural distortions due to dynamics. In our case, for the distal and proximal domains, the values range from 0.96 to 1.0, indicating only very small structural distortions due to motion, if any. In addition, independent validation using ^{15}N relaxation data, which are probing dynamics in a much narrower time range compared to RDCs, clearly indicates that both RDC and relaxation data are affected by motions in Ub_2 in approximately the same way. This suggests that whatever conformational averaging is present, it happens primarily on the time scale of the overall rotational diffusion or faster. If dynamic events on a slower time scale would influence the structure calculated here, one

would expect a statistically significant increase in the value of χ^2 when comparing the fits of relaxation data for the distal or proximal domain with di-ubiquitin. This is however not the case. All this indicates (1) that the conformational equilibrium between open and closed conformations on a slow time scale does not significantly affect the measured RDCs and thus our proposed model and (2) that the value of ~15% for the population of the open state is most likely an overestimate. It is not yet possible to rule out the possibility that our structure reflects conformational averaging on a ~10 ns or faster time scale.

Conclusion

Our results show that a docking method based on chemical shift perturbation data, HADDOCK [4], can successfully be combined with residual dipolar couplings following a combined strategy in which RDCs are first introduced as intervector projection angle restraints and at the later refinement stage as direct restraints. The use of intervector projection angles has some definite advantages. First, it leads to a better convergence of the calculations, which, in turn, gives more confidence in the results, and also means that structural characteristics (like interdomain orientation in our case) can be calculated more precisely, because relevant parts of the configuration space are sampled more extensively. The better convergence of the docking run using intervector projection angle restraints could be due to the fact that it is possible to distinguish between intermolecular and intramolecular restraints; by using only the intermolecular restraints in the rigid-body docking phase, it is possible to define directly the relative orientation of the two components. Finally, by avoiding the use of an external tensor in structure calculations, intervector projection angles offer a convenient means to combine RDC data obtained from different alignment media or simultaneously refine structures against both RDCs and relaxation data.

Acknowledgements

This work was supported by NIH grant GM65334 to D.F., and a “Jonge Chemici” grant from the Netherlands Organization for Scientific Research (N.W.O.) to A.B. We thank Charles Geraghty (University of Maryland) for performing hydrodynamic calculations of the rotational diffusion tensor, and Ranjani Varadan (University of Maryland) and Cyril Dominguez (Utrecht University) for data and helpful discussions.

References

1. Alm E, Arkin AP. Biological networks. Curr Opin Struct Biol 2003;13(2):193-202.
2. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 2002;47(4):409-443.
3. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ, . CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 2003;52(1):2-9.
4. Dominguez C, Boelens R, Bonvin AM. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. Journal of the American Chemical Society 2003;125(7):1731-1737.
5. Prestegard JH, al_Hashimi HM, Tolman JR. NMR structures of biomolecules using field oriented media and residual dipolar couplings. Quarterly Reviews of Biophysics 2000;33(4):371-424.
6. de Alba E, Tjandra N. NMR dipolar couplings for the structure determination of biopolymers in solution. Progress in Nuclear Magnetic Resonance Spectroscopy 2002;40(2):175-197.
7. Bax A. Weak alignment offers new NMR opportunities to study protein structure and dynamics. Protein Science : a Publication of the Protein Society 2003;12(1):1-16.
8. Tolman JR. Dipolar couplings as a probe of molecular dynamics and structure in solution. Current Opinion in Structural Biology 2001;11(5):532-539.
9. Bax A, Kontaxis G, Tjandra N. Dipolar couplings in macromolecular structure determination. Methods in Enzymology 2001;339:127-174.
10. Fushman D, Cowburn D. Protein NMR for the Millenium. Krishna, editor: Kluwer; 2002. 53-78 p.

11. Zuiderweg ER. Mapping protein-protein interactions in solution by NMR spectroscopy. Biochemistry 2002;41(1):1-7.
12. Fischer MWF, Losonczi JA, Weaver JL, Prestegard JH. Domain orientation and dynamics in multidomain proteins from residual dipolar couplings. Biochemistry 1999;38(28):9013-9022.
13. Fushman D, Xu R, Cowburn D. Direct determination of changes of interdomain orientation on ligation: Use of the orientational dependence of N-15 NMR relaxation in Abl SH(32). Biochemistry 1999;38(32):10225-10230.
14. Skrynnikov NR, Goto NK, Yang DW, Choy WY, Tolman JR, Mueller GA, Kay LE. Orienting domains in proteins using dipolar couplings measured by liquid-state NMR: Differences in solution and crystal forms of maltodextrin binding protein loaded with beta-cyclodextrin. Journal of Molecular Biology 2000;295(5):1265-1273.
15. Mollova ET, Hansen MR, Pardi A. Global structure of RNA determined with residual dipolar couplings. Journal of the American Chemical Society 2000;122(46):11561-11562.
16. Bayer P, Varani L, Varani G. Refinement of the structure of protein-RNA complexes by residual dipolar coupling analysis. Journal of Biomolecular Nmr 1999;14(2):149-155.
17. Clore GM. Accurate and rapid docking of protein-protein complexes on the basis of intermolecular nuclear overhauser enhancement data and dipolar couplings by rigid body minimization. Proceedings of the National Academy of Sciences of the United States of America 2000;97(16):9021-9025.
18. McCoy MA, Wyss DF. Structures of protein-protein complexes are docked using only NMR restraints from residual dipolar coupling and chemical shift perturbations. Journal of the American Chemical Society 2002;124(10):2104-2105.
19. Xu XP, Case DA. Automated prediction of N-15, C-13(alpha), C-13(beta) and C-13 ' chemical shifts in proteins using a density functional database. J Biomol NMR 2001;21(4):321-333.
20. Clore GM, Schwieters CD. Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from 1H/15N chemical shift mapping and backbone 15N-1H residual dipolar couplings using conjoined rigid body/torsion angle dynamics. Journal of the American Chemical Society 2003;125(10):2902-2912.
21. Dobrodumov A, Gronenborn AM. Filtering and selection of structural models: combining docking and NMR. Proteins 2003;53(1):18-32.
22. Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. Journal of Molecular Biology 1997;272(1):106-120.
23. Meiler J, Blomberg N, Nilges M, Griesinger C. A new approach for applying residual dipolar couplings as restraints in structure elucidation. Journal of Biomolecular NMR 2000;16(3):245-252.
24. Tjandra N, Garrett DS, Gronenborn AM, Bax A, Clore GM. Defining long range order in NMR structure determination from the dependence of heteronuclear relaxation times on rotational diffusion anisotropy. Nature Structural Biology 1997;4(6):443-449.
25. Varadan R, Walker O, Pickart C, Fushman D. Structural properties of polyubiquitin chains in solution. Journal of Molecular Biology 2002;324(4):637-647.
26. Weissman AM. Themes and variations on ubiquitylation. Nature Reviews Molecular Cell Biology 2001;2(3):169-178.
27. Glickman MH, Ciechanover A. The ubiquitin-proteasome proteolytic pathway: Destruction for the sake of construction. Physiological Reviews 2002;82(2):373-428.
28. Pickart CM. Mechanisms underlying ubiquitination. Annual Review of Biochemistry 2001;70:503-533.
29. Cook WJ, Jeffrey LC, Carson M, Chen Z, Pickart CM. Structure of a diubiquitin conjugate and a model for interaction with ubiquitin conjugating enzyme (E2). The Journal of Biological Chemistry 1992;267(23):16467-16471.
30. Phillips CL, Thrower J, Pickart CM, Hill CP. Structure of a new crystal form of tetraubiquitin. Acta Crystallographica Section D, Biological Crystallography 2001;57(Pt 2):341-344.
31. Cook WJ, Jeffrey LC, Kasperek E, Pickart CM. Structure of tetraubiquitin shows how multiubiquitin chains can be formed. Journal of Molecular Biology 1994;236(2):601-609.
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Research 2000;28(1):235-242.
33. Varadan R, Assfalg M, Haririnia A, Raasi S, Pickart C, Fushman D. Solution Conformation of Lys63-linked Di-ubiquitin Chain Provides Clues to Functional Diversity of Polyubiquitin Signaling. The Journal of Biological Chemistry 2004;279(8):7055-7063.
34. Zweckstetter M, Bax A. Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR. Journal of the American Chemical Society 2000;122(15):3791-3792.
35. Tjandra N, Omichinski JG, Gronenborn AM, Clore GM, Bax A. Use of dipolar 1H-15N and 1H-13C couplings in the structure determination of magnetically oriented macromolecules in solution. Nature Structural Biology 1997;4(9):732-738.

36. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse_Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallographica Section D, Biological Crystallography 1998;54 (Pt 5):905-921.
37. Meiler J, Blomberg N, Nilges M, Griesinger C. A new approach for applying residual dipolar couplings as restraints in structure elucidation (vol 16, pg 245, 2000). Journal of Biomolecular NMR 2000;17(2):185-185.
38. Fushman D, Varadan R, Assfalg M, Walker O. Determining domain orientation in macromolecules by using spin-relaxation and residual dipolar coupling measurements. Progress in Nuclear Magnetic Resonance Spectroscopy 2004;44:189-214.
39. Walker O, Varadan R, Fushman D. Efficient and accurate determination of the overall rotational diffusion tensor of a molecule from 15N relaxation data using computer program ROTDIF. Journal of Magnetic Resonance 2004;168:336-345.
40. Fushman D, Tjandra N, Cowburn D. Direct measurement of N-15 chemical shift anisotropy in solution. Journal of the American Chemical Society 1998;120(42):10947-10952.
41. Fushman D, Tjandra N, Cowburn D. An approach to direct determination of protein dynamics from N-15 NMR relaxation at multiple fields, independent of variable N-15 chemical shift anisotropy and chemical exchange contributions. Journal of the American Chemical Society 1999;121(37):8577-8582.
42. Cornilescu G, Marquardt JL, Ottiger M, Bax A. Journal of the American Chemical Society 1998;120(27):6836-6837.
43. Wallace AC, Laskowski RA, Thornton JM. Ligplot - a Program to Generate Schematic Diagrams of Protein Ligand Interactions. Protein Engineering 1995;8(2):127-134.
44. de la Torre JG, Huertas ML, Carrasco B. HYDRONMR: Prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. Journal of Magnetic Resonance 2000;147(1):138-146.
45. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C. New York: Cambridge University Press; 1992.
46. Laskowski R, MacArthur M, Moss D, Thornton J. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 1993;26:283-291.
47. Hayward S, Berendsen HJC. Systematic analysis of domain motions in proteins from conformational change: New results on citrate synthase and T4 lysozyme. Proteins-Structure Function and Genetics 1998;30(2):144-154.
48. Hubbard SJ, Thornton JM. NACCESS. London: Department of Biochemistry and Molecular Biology, University College; 1993.
49. Verma R, Peters NR, D'Onofrio M, Tochtrop GP, Sakamoto KM, Varadan R, Zhang MS, Coffino P, Fushman D, Deshaies RJ, King RW. Ubistatins inhibit proteasome-dependent degradation by binding the ubiquitin chain. Science 2004;306(5693):117-120.
50. Tenno T, Fujiwara K, Tochio H, Iwai K, Morita EH, Hayashi H, Murata S, Hiroaki H, Sato M, Tanaka K, Shirakawa M. Structural basis for distinct roles of Lys63- and Lys48-linked polyubiquitin chains. Genes to Cells 2004;9(10):865-875.
51. Lukin JA, Kontaxis G, Simplaceanu V, Yuan Y, Bax A, Ho C. Quaternary structure of hemoglobin in solution. Proceedings of the National Academy of Sciences of the United States of America 2003;100(2):517-520.
52. Tolman JR, Al-Hashimi HM, Kay LE, Prestegard JH. Structural and dynamic analysis of residual dipolar coupling data for proteins. Journal of the American Chemical Society 2001;123(7):1416-1424.
53. Jorgensen WL, Tirado-Rives J. Journal of the American Chemical Society 1998;110:1657-1666.
54. Linge JP, Williams MA, Spronk CAEM, Bonvin AMJJ, Nilges M. Refinement of protein structures in explicit solvent. Proteins-Structure Function and Genetics 2003;50(3):496-506.
55. Koradi R, Billeter M, Wuthrich K. MOLMOL: A program for display and analysis of macromolecular structures. Journal of Molecular Graphics 1996;14(1):51-&.
56. Kraulis PJ. MOLSCRIPT: A program to produce both detailled and schematic plots of protein structures. J Appl Cryst 1991;24:946-950.
57. Merrit EA, Murphy MEP. Raster3D version 2.0: A program for photorealistic molecular graphics. Acta Cryst 1994;D50:869-873.

Chapter 4.
Combining NMR relaxation with chemical shift perturbation data to drive protein-protein docking.

Aalt D.J. van Dijk, Robert Kaptein, Rolf Boelens and Alexandre M.J.J. Bonvin,
Journal of Biomolecular NMR **34** (2006), 237 - 244.

Abstract

The modeling of biomolecular complexes by computational docking using the known structures of their constituents is developing rapidly to become a powerful tool in structural biology. It is especially useful in combination with even limited experimental information describing the interface. Here we demonstrate for the first time the use of diffusion anisotropy in combination with chemical shift perturbation data to drive protein-protein docking. For validation purposes we make use of simulated diffusion anisotropy data. Inclusion of this information, which can be derived from NMR relaxation rates and reports on the orientation of the components of a complex with respect to the rotational diffusion tensor, substantially improves the docking results.

Introduction

A major innovation in modeling biomolecular complexes/interactions has been the development of docking algorithms that aim to elucidate the structure of a complex based on the known structures of its constituents [1,2]. The docking process can be facilitated by inclusion of experimental information such as the NMR chemical shift perturbations (CSP) that are observed when titrating the molecules together [2-4]. The structure determination of biomolecules by NMR, which is traditionally based mainly on NOEs, has progressed recently by including new experimental information, most notably residual dipolar couplings (RDCs) [5,6] and diffusion anisotropy (relaxation) data [6] which contain valuable long-range orientational information. It has previously been shown that the inclusion of RDCs into docking improves the results [4,7,8].

Diffusion anisotropy data have been used previously in various ways to characterize biomolecular complexes. One approach is to fit the data to the structures of the individual components and align the resulting tensors [6,9,10]. It is also possible to compare back-calculated and experimental relaxation data for various structural models [11-13]. Relaxation data have also been used as restraints in NMR structure calculations to refine a multi domain protein structure in combination with classical NOE information [14,15]. Here we demonstrate that NMR relaxation data can be used to drive protein-protein docking in combination with chemical shift perturbation data. These data have been implemented as additional restraints in our data-driven docking approach HADDOCK [3] that encodes experimental information about interaction surfaces into ambiguous interaction restraints (AIRs). Compared to the use of RDCs, relaxation data do not require dissolving the protein complexes in liquid crystalline media and can be measured from regular solution samples, which can offer a serious advantage.

Methods

Structures and data

Docking was performed on the E2A-HPr complex (PDB 1GGR) [16], using the structures of the unbound components, 1F3G [17] and 1HDN [18], respectively. Histidine 90 on E2A was used in its phosphorylated form. Experimental chemical shift perturbation data [19] were introduced in the form of Ambiguous Interaction Restraints (AIRs) as described previously [3]. In all docking runs, 50% of the AIRs were randomly removed, a procedure which helps to deal with inaccuracies and false positives in the chemical shift perturbation data.

Theoretical relaxation data were generated with HydroNMR [20], using a shell thickness parameter of 2.4Å (this represents the sum of the average atomic van der Waals radius plus the thickness of the hydration shell). HydroNMR uses the “shell-modelling”

strategy where a shell-model composed of 'mini-beads' of radius σ is derived from the primary hydrodynamic model, and extrapolation to the limit σ=0 is carried out. We used the option NSIG=-1, which means that the program estimates the extrapolation limits. The hydrodynamic calculations were performed using a temperature T = 300K, a viscosity $\eta = 7\ 10^{-4}$ kg/(m s) and a field strength of 600MHz. The output of HydroNMR consists of T_1 and T_2 relaxation times and heteronuclear NOE as well as the rotational diffusion tensor. The latter was used to generate three additional artificial sets of relaxation data in CNS [21]: the tensor orientation was kept but larger anisotropy and/or rhombicity were defined, in order to probe the influence of these parameters on the docking results. Note that three parameters are needed to describe the magnitude of the rotational diffusion tensor components. These can be either τ_c, Da and R, or D_x, D_y and D_z, which are related to another *via*: $\tau_c = 0.5/(Dx+Dy+Dz)$, Da = 2Dz/(Dx+Dy) and R = 1.5(Dy-Dx)/(Dz-0.5(Dx+Dy)). In our case, τ_c was 9.8 ns for all sets; the original parameters calculated by HydroNMR (set1) are Da=1.35 and R =0.33; for the other three sets these are 1.35 and 0.7 (set2), 1.8 and 0.33 (set3) and 1.8 and 0.7 (set4), respectively. Dx, Dy and Dz for set1 are given by 1.5, 1.6 and 2.1 10^7 s^{-1}, for set2 by 1.4, 1.7 and 2.1 10^7 s^{-1}, for set3 by 1.2, 1.5 and 2.4 10^7 s^{-1}, and for set4 by 1.1, 1.6 and 2.4 10^7 s^{-1}, respectively.

In order to investigate the influence of experimental noise we generated additional data sets by adding 2% or 5% noise on both T_1 and T_2 for each set defined above. Note that 2% corresponds to a typical experimental noise level while 5% is already pretty high. We used only data from secondary structure elements (70 out of 150 residues for E2a and 57 out of 85 for HPr) for the docking. This somewhat mimics the filtering that is often performed to exclude residues affected by flexibility and/or chemical exchange. (Note that our theoretical data do not suffer from such effects).

Relaxation data as restraints in docking

The use of relaxation data as restraints in NMR structure calculations is described in [14]. These have been implemented in various software, among which XPLOR-NIH [22], SCULPTOR [23] , and CNS [21]. In the latter, the relaxation data are introduced into an energy function defined as:

$$E_{dani} = k_{dani}\left(\left(\frac{T_1}{T_2}\right)_{\exp} - \left(\frac{T_1}{T_2}\right)_{back}\right)^2 \quad (1)$$

Here $(T_1/T_2)_{exp}$ is the ratio of experimental relaxation times and $(T_1/T_2)_{back}$ the back-calculated ratio. For the latter, a floating diffusion tensor is used during the structure calculations. We used a 'square potential' with an error range of 0.2; if the difference between experimental and back-calculated values is lower than this value, E_{dani} is set to 0. In order to back-calculate T_1 and T_2, the rotational diffusion tensor parameters need first to be determined. In the case of docking, the 3D structures of the isolated components are usually known and can be used to fit the T_1/T_2 ratios. This was done using the software Tensor2 [24]. The synthetic relaxation data were fit to the unbound E2A structure and the 10 models of the unbound HPr structure, respectively; the resulting tensor parameters from the best-fitting structure were used subsequently. To probe the influence of the goodness of fit on the tensor parameters, we also did for each of the four sets a docking run using the parameters resulting from the worst fit.

The HADDOCK docking protocol consists of three consecutive stages (for details, see ref. [3]):

(i) randomization of orientations followed by rigid body energy minimization (EM);

(ii) semi-flexible simulated annealing in torsion angle space (TAD-SA), which consists of (ii-a) a rigid body Molecular Dynamics search and first simulated annealing, (ii-b) a second

semi-flexible simulated annealing during which side chains at the interface are free to move, and (ii-c) a third semi-flexible simulated annealing during which both side chains and backbone at the interface are free to move; and

(iii) final refinement in Cartesian space with explicit solvent.

The tensor is introduced at stage (i) with a random orientation; a rotational minimization is used to find its optimal orientation (this is repeated a few times to find the global minimum). During stages (ii) and (iii) the tensor is free to rotate. The values of the force constants during the various stages of the protocol are listed in Table 1.

Table 1. Force constants used during the different stages of the docking protocol.

Stage[a]	k_{dani} (kcal mol^{-1})	k_{air} (kcal mol^{-1} A^{-2})
i) Rigid body EM	1-10	1 - 10
ii-a) SA	1-5	10
ii-b) SA	5-10	10 - 50
ii-c) SA	10	50
iii)Water Refinement	10	50

a) i,ii and iii refer to the rigid body, simulated annealing (SA) and water refinement stages of the protocol, respectively; ii-a, ii-b and ii-c refer to the different parts of the semi-flexible simulated annealing (see Methods for details); k_{dani}, force constant for DANI restraints. k_{air}, force constant for Ambiguous Interaction Restraints.

After each of the different stages, a score is calculated by using weights for the different energy terms (see Table 2). The rigid body docking stage is performed a number of times (in our case: 5 times), and the best resulting structure of those is saved. Note that after rigid body docking out of 1000 structures, the best 200 based on this score are selected for further refinement.

Table 2. Scoring scheme[a] in HADDOCK2.0_devel.

Docking stage Term	Rigid body EM	SA	Water refinement
Elec	1.0	0.2	0.2
vdW	0.01	1.0	1.0
BSA	-0.01	-0.01	0.0
Desolv[b]	1.0	1.0	1.0
AIR	0.01	0.1	0.1
DANI	0.01	0.1	0.1

a) The overall score is calculated as a weighted sum of different terms, using the weights as listed. Elec, electrostatic energy; vdW, van der Waals energy; BSA, buried surface area; Desolv, desolvation energy; AIR, ambiguous interaction restraints; DANI, diffusion anisotropy restraint energy.
b) The desolvation energy is calculated using the atomic desolvation parameters of Fernandez-Recio *et al.* [25].

Results and Discussion

Orientational information to distinguish docking solutions

As CSP data do not define the specific contacts that are made across the interface and thus the relative orientation of the components of a complex, distinguishing between different, possibly symmetry-related, binding modes can be difficult. This is indeed a common problem in our data-driven docking approach: symmetry related solutions are obtained where one molecule is rotated by approximately 180^0 around an axis orthogonal to the binding surface. In favourable cases this ambiguity can be removed by mutagenesis and biochemical methods for screening interactions. An alternative approach that requires no further biochemical manipulation is the inclusion of anisotropic information. Next to RDCs, relaxation data can also be useful for this purpose [6]. To illustrate this, we used HydroNMR [20] to generate a theoretical set of ^{15}N relaxation data (set1) for the E2A-HPr complex [16], which were subsequently converted into diffusion anisotropy restraints (Tjandra et al. 1997) in CNS [14,21]. To monitor the ability of this type of information to distinguish between various

relative orientations in a complex, one of the two components of the complex was rotated around either the z-axis of the rotational diffusion tensor or the z-axis of the inertia tensor, both being approximately orthogonal to the interface. This rotation is meant to represent different docking solutions that could be obtained when only information about the interface would be used: since experimental data such as CSP data define in principle the binding interface, there is basically only one rotational degree of freedom left to describe the relative orientation of the two components of the complex. The restraint energy as a function of rotation angle is shown in Figure 1. When the rotation is performed exactly around the diffusion tensor z-axis, one finds multiple degenerate minima (one at 0^0 and one at 180^0): this is a well-known characteristic of relaxation data. However, even a small difference between rotation and tensor axes is enough to lift this degeneracy: this is illustrated by the continuous lines in Figure 1 that were obtained by rotating the structures around the z-axis of the inertia tensor instead of the diffusion tensor. This second axis would represent the axis of rotation between possible symmetry related docking solutions and its exact orientation will depend on the properties of the binding surface. It has no physical meaning and is not per se related to the diffusion tensor axes.

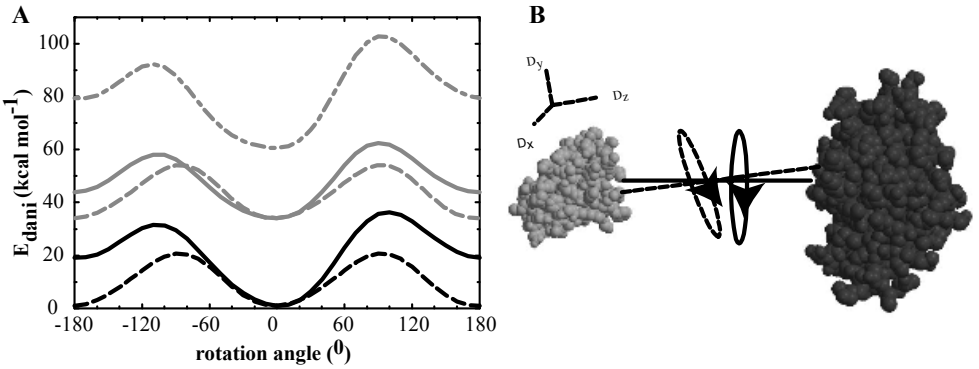


Figure 1. (A) Diffusion anisotropy restraint energy as a function of the rotation angle around the rotational diffusion tensor z-axis (dashed lines) or inertia tensor z-axis (continuous lines) for the bound (black) and unbound (gray) structures of the E2A-HPr complex. In addition, in the case of the unbound structures, a wrong tensor parameter (Da = 1.5 instead of 1.35) was used on purpose for the rotation around the inertia tensor axis (dash-dotted gray line): this did not affect the position of the minimum. **(B)** Representation of the rotation around the z-axis of the diffusion anisotropy (dashed line) and inertia (continuous line) tensor of the E2A-HPr complex (E2A: black, HPr: gray).

The orientation of the two axes differs by only 5^0 in this particular case. Explicit modeling of the interaction by computational docking should thus be able to lift the degeneracy present in the relaxation data provided the interface shows some degree of asymmetry, for example, in the chemical shift perturbation data, the electrostatic potential or the surface shape.

When using the unbound structures (which differ slightly from the bound forms), there is an overall upward shift of the curves (gray lines); however, the minimum is still at 0^0. For the unbound structures we also tested the effect of wrong Da and R values on the shape of the curve. Roughly, errors of up to 20% and 30% in the estimation of Da and R, respectively, still result in a minimum at the correct rotation angle, while distorting the curve (this is illustrated in Figure 1 for the case of using an anisotropy of 1.5 instead of the correct value of 1.35).

Relaxation data in docking

The use of relaxation data was implemented in our data-driven docking approach HADDOCK [3]. HADDOCK encodes experimental information about interaction surfaces into ambiguous

interaction restraints (AIRs). These are defined between any residue which, based on experimental data (e.g. CSP), is believed to be at the interface, and all such residues plus their surface neighbors on the partner molecule. The AIRs are incorporated as an additional energy term into the energy function that one tries to minimize during sampling.

Table 3. Theoretical and fitted tensor parameters for the E2A-HPr complex.[a]

	Set1[a]	Set2[a]	Set3[a]	Set4[a]
Theoretical				
- τ_c(ns)	9.80	9.80	9.80	9.80
- Da	1.35	1.35	1.80	1.80
- R	0.33	0.70	0.33	0.70
Best fit for 0% noise set				
- τ_c(ns)	9.81	9.77	9.76	9.76
- Da	1.26	1.25	1.55	1.52
- R	0.44	0.96	0.46	1.00
- χ^2/df	0.17	0.18	0.62	0.63
Worst fit for 0% noise set				
- τ_c(ns)	9.95	9.91	10.0	10.0
- Da	1.24	1.24	1.49	1.52
- R	0.51	0.87	0.54	0.94
- χ^2/df	0.88	0.92	3.26	3.74
2% noise set				
- τ_c(ns)	9.80	9.75	9.77	9.79
- Da	1.28	1.23	1.56	1.58
- R	0.21	1.11	0.46	0.90
- χ^2/df	0.45	0.66	0.88	1.19
5% noise set				
- τ_c(ns)	9.83	9.70	9.59	9.58
- Da	1.40	1.35	1.40	1.43
- R	0.33	1.09	0.70	1.25
- χ^2/df	2.79	1.88	3.31	3.26

[a] Set1 is the original set calculated by HydroNMR; set2, set3 and set4 are sets with artificially increased anisotropy and/or rhombicity (see Methods).

We introduced the relaxation data as additional restraints; the protocol is comparable to the one presented before for RDCs [7]. Briefly, during the rigid body energy minimization step, the diffusion tensor is introduced with a random orientation. Then, a rotational minimization is used to find its optimal orientation. During the remaining of the protocol, the tensor is free to rotate; for details see Methods. The E2A-HPr complex was docked without and with inclusion of theoretical relaxation data. The latter were obtained with HydroNMR (set1), which predicted an anisotropy Da of 1.35 and rhombicity R of 0.33, and also with CNS using the tensor orientation obtained with HydroNMR but, for testing purposes, a higher anisotropy and/or rhombicity (see Methods; set2: Da = 1.35, R=0.7; set3: Da=1.8 and R=0.33; set4: Da=1.8 and R=0.7). These sets were used to probe the influence of the amount of anisotropy on the docking results. AIRs were defined based on the available experimental CSP-data for the complex [19]. Starting from the unbound conformations, 1000 structures were generated in the rigid body docking phase, out of which 200 were further refined (using semi-flexible simulated annealing and water refinement). The tensor parameters needed as input for the protocol were obtained from the known structures of the unbound constituents using Tensor2

[24] as commonly done for RDCs (see Methods). The best fit was obtained in all cases for E2A; for HPr, similar tensor parameters were obtained, but with somewhat higher χ^2 values. The tensor parameters used in docking are listed in Table 3. Note that the difference in goodness of fit is due to the fact that the unbound E2A structure is closer to its bound form than the HPr structure (which is reflected in the backbone RMSD values between free and bound structures: 0.3Å for E2A vs. 1.3 (+/- 0.1) Å for HPr).

Inclusion of relaxation data in general improves the docking results, even in the case of substantial amounts of noise (Table 4). The HADDOCK score of the resulting structures is plotted against the interface RMSDs from the target in Figure 2 for each of the four 2% noise sets corresponding to various amounts of anisotropy. The interface RMSD is defined as the backbone RMSD from the structure of the complex for those residues making contacts across the interface within a 10.0Å cutoff. The inclusion of relaxation data results clearly in a larger number of structures with low interface RMSD (see also Table 4) and a larger energy difference between correct and incorrect solutions, which improves the scoring of the solutions. In all cases the scoring of the solutions is improved when diffusion anisotropy data are included, as can be seen from the number of low-RMSD structures among the 10 best-scoring structures (Table 4). For the set with 5% noise, this is still the case, although the total number (out of 200) of correct solutions varies between the various sets (due to the influence of the tensor parameters). It is also clear from the results in Table 4 that larger anisotropies result in a larger number of correct solutions.

Table 4. Docking results for the E2A-HPr complex.[a]

Docking run	Number of structures with iRMSD[b] < 1 / 4 Å			
	Rigid body		Refined	
	All 1000	Top 200	All 200	Top 10
Reference	3 / 376	1 / 159	3 / 159	0 / 7
DANI set1 - no noise	4 / 414	4 / 168	3 / 168	3 / 10
- worst fit[c]	0 / 309	0 / 154	5 / 155	2 / 10
- 2% noise	5 / 327	5 / 154	2 / 154	1 / 10
- 5% noise	1 / 401	1 / 161	2 / 161	0 / 10
DANI set2 - no noise	2 / 528	1 / 191	9 / 191	3 / 10
- worst fit[c]	3 / 474	3 / 191	12 / 191	4 / 10
- 2% noise	2 / 532	1 / 190	7 / 190	4 / 10
- 5% noise	4 / 474	2 / 196	3 / 196	2 / 10
DANI set3 - no noise	4 / 487	3 / 192	5 / 192	0 / 10
- worst fit[c]	13 / 441	13 / 192	26 / 192	5 / 10
- 2% noise	1 / 481	1 / 179	11 / 178	3 / 10
- 5% noise	1 / 511	1 / 193	5 /192	1 / 10
DANI set4 - no noise	7 / 544	4 / 194	20 / 194	3 /10
- worst fit[c]	8 / 545	7 / 197	27 / 197	5 / 10
- 2% noise	7 / 551	5 / 198	22 / 198	6 / 10
- 5% noise	2 / 494	2 / 186	10 / 185	5 / 10

a) One reference docking run was performed using only CSP and no diffusion anisotropy data; docking runs with T_1/T_2 restraints corresponding to various amounts of anisotropy were performed; see Table 3 for the tensor parameters corresponding to the various sets.

b) Interface backbone RMSD from the NMR structure (PDB id 1GGR) .

c) Worst fit: docking run using tensor parameters with highest as opposed to lowest χ^2 values (see Table 3).

Interestingly, the results from the docking where the worst-fit tensor parameters were used are in general still better than the reference docking run without relaxation data. This indicates that the method is not too sensitive to incorrect tensor parameters which is in line with our analysis of the energy function discussed above (see Fig. 1).

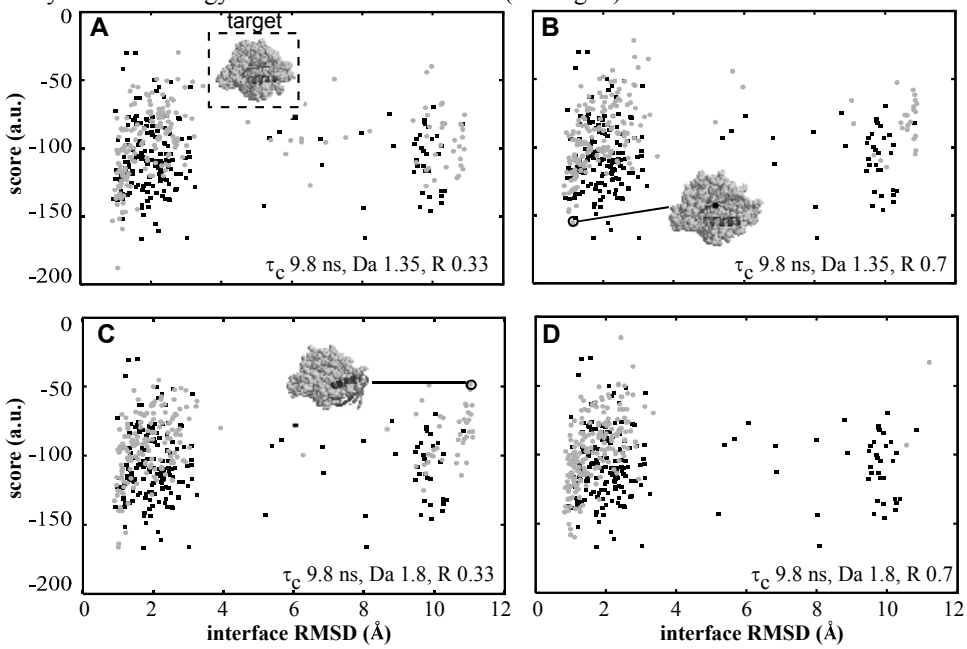


Figure 2. HADDOCK score versus interface RMSDs from the target for docking runs without (black) and with (gray) relaxation data for various amounts of anisotropy **(A)** set1; **(B)** set2; **(C)** set3; and **(D)** set4 (see Table 3 and 4). The reference structure is shown in panel A and examples of docked structures in panels B and C. The HADDOCK score corresponds to the weighted sum of the van der Waals, electrostatic, and restraint energy (DANI + AIR) (see Table 2). The interface RMSDs are calculated on the backbone atoms of the residues making contacts across the interface within a 10.0Å distance cutoff.

In the case of docking with relaxation data, structures with large RMSDs have in most cases a low relaxation restraint energy, as they correspond to 180º symmetrical solutions around the tensor axis (see inset in Fig. 2C); their interface is however different, resulting in higher AIR and especially electrostatic energies. This is related to the fact that asymmetry in the interface and possibly in the available information describing it (e.g. CSP) is able to lift the degeneracy present in the relaxation data (see above).

One practical limitation of using relaxation data is that chemical exchange and/or flexibility can influence the experimental data. Such effects are absent from our theoretical data sets; however, results obtained in the presence of noise indicate that diffusion anisotropy data will be useful in defining the intermolecular orientation of the components of a complex provided the experimental errors are not too large.

Conformational differences between unbound and bound forms could possibly affect the tensor parameters determined by fitting the data to the unbound components, and thus the docking results. It has however been shown that the tensor parameters can be quite accurately determined even in the presence of substantial experimental errors [6]. In addition, our results obtained with wrong parameters show that the diffusion anisotropy energy (DANI) function is not too sensitive to such errors and still allows to identify the correct minimum as can be seen for the unbound case in Fig. 1 (gray dash-dotted curve). This is also illustrated by our docking results using the "worst fit" tensor parameters.

In conclusion, we demonstrated the usefulness of introducing NMR relaxation data in protein – protein docking. These improve the convergence of our docking protocol and both the accuracy and discrimination of the correct solutions. Compared to the use of RDCs, relaxation data have the advantage that their measurement does not require dissolving the protein complexes in liquid crystalline media. This methodology should be useful for the modeling of large protein-protein complexes.

Acknowledgments

This work was supported by a "Jonge Chemici" grant from the Netherlands Organization for Scientific Research (N.W.O.) to A.B.

References

1. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 2002;47(4):409-443.
2. van Dijk ADJ, Boelens R, Bonvin AMJJ. Data-driven docking for the study of biomolecular complexes. Febs J 2005;272(2):293-312.
3. Dominguez C, Boelens R, Bonvin AM. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 2003;125(7):1731-1737.
4. Clore GM, Schwieters CD. Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from 1H/15N chemical shift mapping and backbone 15N-1H residual dipolar couplings using conjoined rigid body/torsion angle dynamics. J Am Chem Soc 2003;125(10):2902-2912.
5. Bax A, Grishaev A. Weak alignment NMR: a hawk-eyed view of biomolecular structure. Curr Opin Struct Biol 2005;15(5):563-570.
6. Fushman D, Varadan R, Assfalg M, Walker O. Determining domain orientation in macromolecules by using spin-relaxation and residual dipolar coupling measurements. Prog Nucl Magn Reson Spectrosc 2004;44:189-214.
7. van Dijk AD, Fushman D, Bonvin AM. Various strategies of using residual dipolar couplings in NMR-driven protein docking: Application to Lys48-linked di-ubiquitin and validation against (15)N-relaxation data. Proteins 2005;60(3):367-381.
8. Dobrodumov A, Gronenborn AM. Filtering and selection of structural models: combining docking and NMR. Proteins 2003;53(1):18-32.
9. Bruschweiler R, Liao XB, Wright PE. Long-Range Motional Restrictions in a Multidomain Zinc-Finger Protein from Anisotropic Tumbling. Science 1995;268(5212):886-889.
10. Hwang PM, Skrynnikov NR, Kay LE. Domain orientation in beta-cyclodextrin-loaded maltose binding protein: Diffusion anisotropy measurements confirm the results of a dipolar coupling study. J Biomol NMR 2001;20(1):83-88.
11. Barbato G, Ikura M, Kay LE, Pastor RW, Bax A. Backbone Dynamics of Calmodulin Studied by N-15 Relaxation Using Inverse Detected 2-Dimensional Nmr-Spectroscopy - the Central Helix Is Flexible. Biochemistry 1992;31(23):5269-5278.
12. Fushman D, Xu R, Cowburn D. Direct determination of changes of interdomain orientation on ligation: Use of the orientational dependence of N-15 NMR relaxation in Abl SH(32). Biochemistry 1999;38(32):10225-10230.
13. Bernado P, Akerud T, de la Torre JG, Akke M, Pons M. Combined use of NMR relaxation measurements and hydrodynamic calculations to study protein association. Evidence for tetramers of low molecular weight protein tyrosine phosphatase in solution. J Am Chem Soc 2003;125(4):916-923.
14. Tjandra N, Garrett DS, Gronenborn AM, Bax A, Clore GM. Defining long range order in NMR structure determination from the dependence of heteronuclear relaxation times on rotational diffusion anisotropy. Nat Struct Biol 1997;4(6):443-449.
15. Hashimoto Y, Smith SP, Pickford AR, Bocquier AA, Campbell ID, Werner JM. The relative orientation of the fibronectin (6)F1(1)F2 module pair: A N-15 NMR relaxation study. J Biomol NMR 2000;17(3):203-214.
16. Wang G, Louis JM, Sondej M, Seok YJ, Peterkofsky A, Clore GM. Solution structure of the phosphoryl transfer complex between the signal transducing proteins HPr and IIA(glucose) of the Escherichia coli phosphoenolpyruvate:sugar phosphotransferase system. Embo J 2000;19(21):5635-5649.

17. Worthylake D, Meadow ND, Roseman S, Liao DI, Herzberg O, Remington SJ. Three-dimensional structure of the Escherichia coli phosphocarrier protein IIIglc. Proc Natl Acad Sci U S A 1991;88(23):10382-10386.
18. van Nuland NA, Hangyi IW, van Schaik RC, Berendsen HJ, van Gunsteren WF, Scheek RM, Robillard GT. The high-resolution structure of the histidine-containing phosphocarrier protein HPr from Escherichia coli determined by restrained molecular dynamics from nuclear magnetic resonance nuclear Overhauser effect data. J Mol Biol 1994;237(5):544-559.
19. Chen Y, Reizer J, Saier MH, Jr., Fairbrother WJ, Wright PE. Mapping of the binding interfaces of the proteins of the bacterial phosphotransferase system, HPr and IIAglc. Biochemistry 1993;32(1):32-37.
20. de la Torre JG, Huertas ML, Carrasco B. HYDRONMR: Prediction of NMR relaxation of globular proteins from atomic-level structures and hydrodynamic calculations. J Magn Reson 2000;147(1):138-146.
21. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse_Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54 (Pt 5):905-921.
22. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. J Magn Reson 2003;160(1):65-73.
23. Hus JC, Marion D, Blackledge M. Model-dependent artifacts in long-range structure determination using orientational restraints derived from NMR relaxation. J Am Chem Soc 1999;121(10):2311-2312.
24. Dosset P, Hus JC, Blackledge M, Marion D. Efficient analysis of macromolecular rotational diffusion from heteronuclear relaxation data. J Biomol NMR 2000;16(1):23-28.
25. Fernandez-Recio J, Totrov M, Abagyan R. Identification of protein-protein interaction sites from docking energy landscapes. J Mol Biol 2004;335(3):843-865.

Chapter 5.
Solvated docking: introducing water into the modelling of biomolecular complexes.

Aalt D.J. van Dijk and Alexandre M.J.J. Bonvin, *Bioinformatics*, in press.

Abstract

Interfacial water, which plays an important role in mediating biomolecular interactions, has been neglected in the modelling of biomolecular complexes. Here, we present a solvated docking approach that explicitly accounts for the presence of water in protein-protein complexes. Our solvated docking protocol is based on the concept of the first encounter complex in which a water layer is present in-between the molecules. It mimics the pathway from this initial complex towards the final assembly in which most waters have been expelled from the interface. Docking is performed from solvated biomolecules and waters are removed in a biased Monte Carlo procedure based on water-mediated contact propensities obtained from an analysis of high-resolution crystal structures. We demonstrate the feasibility of this approach for protein-protein complexes representing both "wet" and "dry" interfaces. Solvated docking leads to improvements both in quality and scoring. Water molecules are recovered that closely match the ones in the crystal structures.

Introduction

The modelling of protein-protein complexes by means of docking (a computational approach which models the unknown structure of a complex from its constituents) has become increasingly popular, as witnessed by the CAPRI (Critical Assessment of PRedicted Interactions) experiment [1]. Docking approaches have benefited from knowledge obtained by detailed analyses of binding interfaces [2, 3]. As discussed in a recent review, water molecules are expected to influence the assembly of biomolecular complexes [4], and, as such, to be important for protein-protein docking. An analysis based on Voronoi volume showed that only upon inclusion of interfacial solvent molecules are protein-protein interfaces as densely packed as protein interiors [5]. So far, however, water has generally been neglected in biomolecular docking. Its role and importance in single proteins have been discussed [6-11] and several case studies have analyzed its conservation in 3D structures of homologs [6, 12-20]. There has also been quite some interest in identifying and predicting the positions of water molecules in known structures: this can be quite successfully performed for example by GRID [9, 10, 21] or Fold-X [22]. These kind of approaches, however, are not very well suited for docking purposes, since the structure of the complex is not known *a priori*. Ideally, water should be accounted for directly during the docking process since its presence might affect the resulting models. So far this has only be done for protein-ligand [23-27] and nucleic acid-ligand docking [28] .

Only very recently has the role of water molecules at protein-protein interfaces been investigated. A hydrogen bonding potential for water-mediated contacts, in combination with a solvated rotamer library for describing side chain conformations, has been shown to predict rather successfully the positions of water molecules in complexes with known structures [29]. In another study [30], various properties of interfacial water molecules such as residue preference and their number per unit of interface area were investigated.

We have previously experimented with the inclusion of water in the NMR structure calculation of a protein - non-specific DNA complex [31]: in that case, an extensive set of NOEs could be used, which forced the solvated biomolecules to come together and the unnecessary waters to leave the interface in a simulated annealing molecular dynamic approach. In general, in docking, this kind of experimental information is not available and, in the absence of a driving force, the water molecules will remain trapped at the interface. Alternative approaches are thus needed to remove the unnecessary water molecules from the interface. We have developed for this purpose a solvated docking protocol implemented in our data-driven

docking approach HADDOCK [32] and demonstrate here for the first time that water can be explicitly included in protein-protein docking.

Methods

Database Analysis

In order to obtain information on water in high-resolution crystal structures of complexes, the non-redundant dataset of Keskin *et al.* [33] was analyzed using CNS [34] and a set of home written Python scripts. Interface residues were defined as residues having at least one heavy-atom contact with a residue from the partner chain, within a 10 Å cut-off distance. Water-mediated contacts were defined between pairs of interface residues, provided a water molecule is making at least one heavy atom contact within 5 Å with both residues. Water-mediated contacts were designated main chain when at least one contact was made via a backbone atom; otherwise they were designated side chain.

To investigate whether the various types of water-mediated contacts adopt specific, well-defined conformations, we clustered them on the basis of positional RMSD values: the RMSD values were calculated after least-square positional fitting on the coordinates of the water oxygen, its contacting heavy atoms within 5Å on both chains and their respective first bonded partner (total of five atoms). Since several atoms of a given side-chain can make contacts with the water oxygen atom within 5Å, various combinations of atoms were tested for the calculation of the RMSD matrix and the one resulting in the best clustering (most populated first cluster) was selected for each amino acid – amino acid pair. Clustering was performed separately for main chain – water – main chain, side chain – water – side chain and main chain – water – side chain contacts. In the case of main chain contacts, N and O were defined as contacting atoms, with CA and C, respectively, as bonded neighbors.

RMSDs were calculated using g_rms [35] and Profit (www.bioinf.org.uk/software/profit). Clustering was performed using the greedy algorithm described by Daura *et al.* [36], with a cut-off of 1.5 Å. This cut-off was based on an analysis of the distribution of all RMSD values (data not shown). Contacts involving two close waters that would fall into the same cluster were counted only once.

Protein – Protein Docking using Explicit Water

HADDOCK incorporates information about the interface in ambiguous interaction restraints (AIRs) that drive the docking. An AIR is defined as an ambiguous intermolecular distance (d_{iAB}) with a maximum value of typically 2Å between any atom m of an active residue i of protein A (m_{iA}) and any atom n of both active and passive residues k (N_{res} in total) of protein B (n_{kB}) (and inversely for protein B). The effective distance d_{iAB}^{eff} for each restraint is calculated using the equation:

$$d_{iAB}^{eff} = \left(\sum_{m_{iA}=1}^{N_{atoms}} \sum_{k=1}^{N_{resB}} \sum_{n_{kB}=1}^{N_{atoms}} \frac{1}{d_{m_{iA}n_{kB}}^{6}} \right)^{-\frac{1}{6}}$$

where N_{atoms} indicates all atoms of a given residue and N_{res} the sum of active and passive residues for a given molecule. Note that the effective distance calculated in this way will always be shorter than the shortest distance entering the sum, which is the reason why we can use a rather short upper bound of 2Å. The definition of passive residues ensures that residues which are at the interface but are not detected, are still able to satisfy the AIR restraints, i.e. contact active residues of the partner molecule. For details, see ref. [3, 32]. HADDOCK consists of a collection of scripts derived from ARIA1.2 [37] and CNS [34]. The respective position and orientation of the two molecules are first randomized. Then docking is performed consisting

of a rigid body energy minimization, followed by semi-flexible simulated annealing in torsion angle space and final refinement in explicit solvent. Rigid body docking is performed a number of times (1000); each time, out of a number of trials (typically 5) only the best model is selected and written to disk.

We modified the rigid body docking stage to explicitly include water. We start by solvating the two chains using a box of TIP3P [38] water. All waters outside a cut-off range (<4.0Å - >8.0Å) from the protein are removed. A short Molecular Dynamics run is performed to optimize the water positions while keeping the proteins fixed (4000 MD steps consisting of four times 1000 steps at a temperature of 600, 500, 400 and 300K, respectively). After that, all waters further away than 5.5Å are removed. An ensemble of different solvation shells (typically 5) is generated by randomly rotating the protein before adding the solvation shell. We also experimented with the use of GRID [21] to place the initial waters around the separate protein chains. The results of the subsequent docking did not depend much on the choice of the solvating method (data not shown). The solvated docking protocol itself is presented in the Results section.

The standard semi-flexible refinement of HADDOCK consists of two rigid body simulated annealing stages followed by two simulated annealing stages with flexibility introduced first on side chains and then on backbone. For solvated docking we only used the latter two semi-flexible simulated annealing stages.

Non-bonded energies (sum of van der Waals and electrostatic terms) are calculated with an 8.5Å distance cut-off using the OPLS non-bonded parameters [39] from the parallhdg5.3.pro parameter file [40]; the dielectric constant ε is set to 10.0 to damp the electrostatic contribution in vacuum. The overall score is calculated as a weighted sum of different terms, using the default HADDOCK2.0 values for the weights (rigid body stage: E_{vdW} 0.01, E_{elec} 1.0, E_{AIR} 0.01, Buried Surface Area -0.01, E_{desolv} 1.0; semi-flexible refinement: E_{vdW} 1.0, E_{elec} 1.0, E_{AIR} 0.1, Buried Surface Area -0.01, E_{desolv} 1.0). Here vdW is van der Waals energy; Elec, electrostatic energy; AIR, ambiguous interaction restraints; BSA, buried surface area; and Desolv, desolvation energy. The desolvation energy is calculated using the atomic desolvation parameters of Fernandez-Recio *et al.* ([41]). The various weights were obtained by a grid search to optimize scoring over the complexes tested so far including CAPRI targets. These were optimized separately for the various stages of HADDOCK to reflect the various levels of complexity and refinement (from rigid body docking in vacuum to flexible refinement in explicit solvent).

Test systems

We tested our protocol on ten protein-protein complexes (see Table 2). Note that there is only a limited number of complexes that are suitable as test cases: the resolution should be high enough (>2Å) in order to have reliable positions for interfacial water molecules, and the free structures of the components of the complex should be available. We used all structures from the docking benchmark [42] satisfying those criteria and a few other complexes which we have been testing before. For two of these, E2A-HPr [43] and cohesin – dockerin [44], we used experimental data available from the literature (NMR chemical shift perturbation data for E2A-HPr [32] and mutagenesis and conservation data as used previously for docking cohesin – dockerin, which was one of the targets in round 4 of the Critical Assessment of Predicted Interactions [45]. For the others, AIRs were defined based on the interface residues identified in the crystal structure; for those complexes, to simulate a more realistic case, 50% of the restraints were randomly removed for each docking trial. When free structures of the complex components were available (7 cases, see Table 2), we performed unbound docking followed by semi-flexible refinement as well as bound docking. For cohesin-dockerin, bound-unbound

docking was performed in addition to bound docking, and for the other 2 cases only bound docking was performed.

Results

Our "solvated docking" protocol is based on the physical concept that, in the first encounter complex, a water layer will be present in-between the two protein chains. To proceed from the encounter complex to the final structure, most of the interfacial waters have to be removed. Our protocol mimics this process by starting the docking from solvated molecules. Water is subsequently removed in a biased Monte Carlo procedure based on water-mediated contact propensities. The latter are obtained from an analysis of a database of high resolution crystal structures of protein-protein complexes. In the following we will first describe the results of this analysis and then present our solvated docking protocol, demonstrating its feasibility for a number of protein-protein complexes.

Analysis of water mediated contacts

In order to extract statistics of water-mediated contacts, we analyzed the high-resolution structures (≤ 2.0 Å) in the non-redundant dataset of protein-protein interfaces of Keskin *et al.* [33]. The corresponding PDB id's are provided in Suppl. Table 5. Some general statistics of our dataset are listed in Table 1.

In Figure 1, the fraction of water-mediated side chain and main chain contacts for all 20x20 amino acid combinations is shown. It is clear from this figure that preferences do exist for specific water-mediated contacts, an information which should be useful in the modelling of protein-protein complexes by docking (see below). In order to assess the statistical significance of the fractions of water-mediated contacts we compared the values obtained from the non-redundant filtered set with those obtained using the complete redundant set of structural homologues. Since these have a lower resolution, the derived fractions are lower than those from the filtered set (data not shown); there is, however, a clear correlation between the two datasets (R=0.6). It is however clear that the propensities reported here should be refined in the future by making use of the (rather slowly) increasing number of protein complexes deposited into the PDB.

Table 1. Analysis of water in Keskin dataset.

Resolution (Å)	1.1-2.0
$N_{structures}$	19
<Chain length>	158 (111)
<# of waters>	346 (264)
<# of water per residue>	0.80
<Number of water at interface>[a]	24 (18)
N_c/N_{wmc}[b]	7155/1544
f_{wmc}: sc/mc[c]	0.16/0.05

[a] Number of waters within heavy atom distance cut-off of 5.0 Å from both chains.
[b] N_c: total number of interface contacts (defined using a 10 Å heavy atom distance cut-off) in dataset; N_{wmc}: total number of water-mediated contacts in dataset.
[c] f_{wmc}: fraction of water-mediated side chain (sc) and main chain (mc) contacts

To find out whether interfacial water molecules adopt specific, well-defined conformations, we clustered the water-mediated contacts based on pairwise RMSDs (see Methods and Suppl. Materials for details). The rationale behind this analysis is that, if water molecules do adopt well-defined specific positions in an interface, one might be able to derive for each type of water-mediated contact a few preferred conformations (an analogy in protein structures would

be the rotameric states of side chains). Such information might be useful in the modeling of water-mediated contacts. The clustering statistics are reported in Suppl. Table 7. Using a 1.5 Å clustering cut-off almost 90% (118 out of 133) of the side chain contacts that could be clustered (133 out of 210) fall into one or two clusters (note that contacts for which less than two water-mediated instances were found could not be clustered at all). Figure 2 shows examples of clusters found for the most populated water-mediated contacts in the resolution-filtered Keskin dataset; in addition, the main backbone-backbone contact (O - H2O - O) and the best-clustering backbone – side chain contact (Ser side chain – N) are shown.

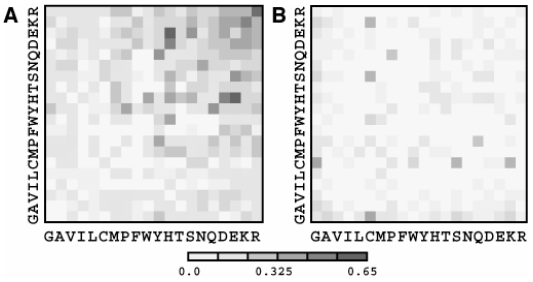


Figure 1. Fraction of water-mediated contacts for each amino acid pairwise combination. Amino acid - amino acid contacts are colour-coded according to the fraction of water-mediated contacts for **(A)** side chain and **(B)** main chain contacts. The corresponding numbers for the matrix elements are provided as supplementary material (Suppl. Table 6).

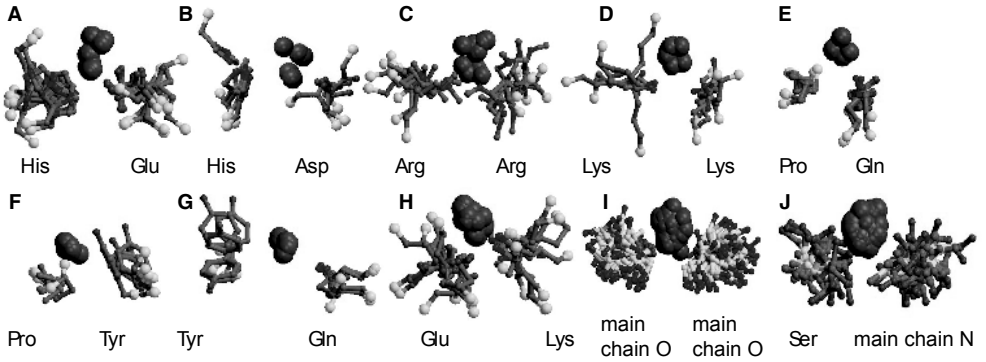


***Figure 2.** View of the most populated water-mediated clusters as found for the resolution-filtered Keskin dataset. **(A-G)** most often occurring side chain contacts; **(H)** best clustering side chain contact; **(I)** best clustering O-O main chain contacts; and **(J)** Ser side chain – main chain N contact. The clustering was performed based on positional RMSD (see Methods). Water oxygens are shown in red spheres. The amino acid side chain atoms are shown in ball-and-stick (red: oxygen; blue: nitrogen; grey: carbon), together with the Cα in white CPK. In the cases involving main chain contacts (I,J), only main chain atoms are shown.

Solvated Docking

Our solvated docking approach is based on the concept of the first encounter complex in which the proteins are separated by a hydration layer. Prior to docking, we solvate the protein chains with one hydration layer as described in Methods. Then, the conventional HADDOCK rigid body docking protocol is followed; for this, each protein and its associated solvation shell is considered as one rigid body. This results in an encounter complex with a water-layer in between the two protein chains. All non-interfacial water molecules are removed from this complex and the remaining waters, together with the protein chains, are treated as separate rigid bodies in a subsequent energy minimization stage (1000 EM steps were found to be sufficient for convergence). Water molecules are then removed in a biased Monte Carlo procedure: randomly chosen water molecules are probed for their closest amino acid residues on both chains; their probability to be kept is set equal to the observed fraction of water-mediated contacts for this specific amino acid combination as derived from the resolution-filtered Ke-

skin set (see above). This procedure is repeated until only 25% of the initial interfacial water molecules remain. Subsequently, water molecules with an unfavorable interaction energy (sum of van der Waals and electrostatic water-protein energies > 0.0 kcal/mol) are removed.

Finally, the remaining waters and the protein chains are again subjected to a rigid body energy minimization (see Suppl. Fig. 6 for an overview). Note that we checked that the use of water-mediated propensities to bias water removal does lead to improvement compared to a simple random removal of waters. The number of retained waters at the end of our protocol is usually lower than 25% because of the energy criterion, typically between 10 and 20%. This fraction is roughly in accordance with a recent study [30] where it was found that, on average, 90% of the interface waters are removed upon assembly. In fact, we observe a substantial variation in the final number of water molecules in the docked structures for the complexes that we used to test our protocol (see below, Table 4).

The solvated docking protocol as described above corresponds to the rigid body docking stage in HADDOCK. The resulting structures are then further refined using semi-flexible simulated annealing. Since water is introduced during rigid body docking we focus the discussion of our results on this stage, but we will also show some initial results for the semi-flexible refinement.

We tested our solvated docking approach on ten complexes representing both "wet" and "dry" interfaces (Table 2). An accurate docking protocol accounting for the presence of water should not only be able to correctly position water molecules at the interface, thereby improving the docking results in the case of "wet" interfaces, but it should also avoid retaining waters in "dry" interfaces in order not to deteriorate the docking results. Assessed by the number of fully buried water molecules, the α-amylase - αAI and barnase - barstar complexes are representative of "wet" interfaces, the PKC interacting protein complex represents a completely "dry" interface and most of the other complexes are in-between. Only the E2A-HPr complex is an NMR structure for which no information on water positions is available.

Table 2. Protein-protein complexes used in solvated docking.

	PDB-id[a]	Res (Å)[b]	$N_{w,bur}$[c]	BSA (Å^2)[d]	Docking[e]
α-amylase - αAI [46]	1dhk	1.9	25	3020	B/B; I
Barnase - barstar [47]	1brs; 1a2p; 1a19	2.0	18	1556	B/B + U/U; I
Subtilisin – subtilisin inhibitor [48]	2sic; 1sup; 3ssi	1.8	8	1617	B/B + U/U; I
Colicin E7 – Im7 [49]	7cei; 1ayi, 1cei, 1unk; 1m08	2.3	8	1384	B/B + U/U; I
bovine trypsin - CMTI-1 squash inhibitor [50]	1ppe; 1btp; 1lu0	2.0	6	1688	B/B + U/U; I
Cohesin – dockerin [44]	1ohz; 1anu	2.2	5	1504	B/B + U/B; E
GRB2 C-ter SH3 domain – N-ter SH3 domain [51]	1gcq; 1gcp; 1gri	1.7	4	1208	B/B + U/U; I
porcine trypsin – soybean trypsin inhibitor [52]	1avx; 1ba7; 1qqu	1.9	1	1585	B/B + U/U; I
PKC interact. protein [53]	1kpf	1.5	0	3700	B/B; I
E2A - HPr [43]	1ggr; 1f3g; 1hdn	-[b]	-	1374	B/B+U/U; E

[a] PDB-id of the complex followed by the PDB-id of the unbound structures if available.
[b] Resolution; note that E2A-HPr (1ggr) is an NMR structure.
[c] Number of fully buried interfacial water molecules.
[d] Buried surface area as calculated using NACCESS [54].
[e] Bound (B) or unbound (U) docking; experimental data (E) or interface residues (I) were used to define AIRs.

The docking was performed using either the bound (B) structures from the complex or the unbound (U) structures; in the latter case rigid body docking was followed by flexible refinement. Experimental data (E) or interface residues (I) in the complex were used to define the ambiguous interaction restraints, 50% of which were randomly discarded for each docking trial in the latter case (see Methods). Further details on these complexes and the information used to drive the docking can be found in the Methods section.

For each complex, two runs were performed: one reference run without water and one following our new solvated docking approach (see Methods). This was done for bound docking (using the bound structures of the components of the complex) and, if unbound structures were available, repeated for unbound docking. The bound docking results are presented in Suppl. Table 8. Table 3 gives an overview of the unbound docking results, assessed by interface-RMSD (i-RMSD) to the target structure. The i-RMSD is defined as the backbone RMSD from the reference structure of the complex for those residues making contacts across the interface within a 10Å cut-off (i-RMSDs below 2Å and 4Å are considered as medium quality and acceptable predictions, respectively, according to the CAPRI criteria [1]).

Table 3. Unbound solvated and unsolvated docking results.[a]

		Rigid body				**Refined**			
		Top 200[b] <4Å	Top RMSD[c]	Best Rank[d]	Best RMSD[e]	All 200[b] <4 Å	Top RMSD[c]	Best Rank[d]	Best RMSD[e]
1brs	- R	**5**	**8.8**	**78***	**2.8**	**5**	**9.1**	**119***	**2.7**
	- S	**26**	**8.8**	**4**	**1.5**	**25**	**9.0**	**12**	**1.4**
2sic	- R	168	**1.9**	**1**	**1.6**	168	1.7	1	**1.3**
	- S	72	**1.9**	**1**	**1.5**	72	7.7	2	**1.3**
7cei	- R	**196**	**11.0**	**2**	**1.1**	**196**	1.3	**1**	0.8
	- S	**199**	**1.6**	**1**	**1.0**	**199**	1.4	**1**	0.9
1ppe	- R	198	**5.1**	**3**	**1.4**	198	1.1	**1**	**1.0**
	- S	186	**1.5**	**1**	**1.4**	186	1.2	**1**	**0.8**
1ohz (U/B)	- R	**19**	**5.8**	25	**0.7**	**17**	**6.2**	**38**	**1.1**
	- S	**33**	**3.1**	38	**0.7**	**33**	**6.0**	**2**	**0.7**
1gcq	- R	70	7.0	**19**	**1.4**	71	**4.0**	**2**	**1.1**
	- S	64	8.3	**3**	**1.4**	63	**1.7**	**1**	**1.1**
1avx	- R	194	1.4	**1**	1.4	194	1.6	**1**	1.0
	- S	171	1.7	**1**	1.5	171	1.9	**1**	1.1
1ggr	- R	106	**2.6**	**80**	**1.5**	106	**10.0**	**2**	**1.0**
	- S	96	**1.5**	**2**	**1.3**	95	**1.4**	**1**	**1.0**

[a] Results from reference (R) unsolvated and solvated (S) protein – protein docking for the various test-cases (see Table 2). Bold indicates cases where solvated docking performs equal to or better than unsolvated docking.
[b] The number of structures below the indicated interface-RMSD (i-RMSD) values is reported (<4Å: acceptable quality). The i-RMSD is calculated over the backbone atoms of all residues making contacts across the interface within a 10Å cut-off.
[c] i-RMSD of top ranking solution.
[d] Rank of best-ranked structure below 2Å i-RMSD; when there are no structures below 2Å i-RMSD this is the rank of the best-ranked structure below 4Å i-RMSD (indicated with *).
[e] i-RMSD value of best structure (closest to target).

As can be seen from Table 3, the inclusion of water in docking generally improves the scoring of the solutions. This is clear from the i-RMSD of the top ranking solution: for the solvated docking, this is in 5 cases a medium quality solution and in 1 case an acceptable solution, whereas for the unsolvated docking this is in only 2 cases a medium quality solution and in 1 case an acceptable solution. In addition, the rank of the best-ranked medium quality solution is in most cases lower for the solvated docking. Finally, the lowest RMSD found in all top 200 ranked structures is on average lower for the solvated docking. Note that scoring in our solvated docking protocol includes the water-water and water-protein non-bonded energy contributions, which clearly improves the performance (data not shown).

After flexible refinement (see Table 3) the same conclusions are valid, although the differences between solvated and unsolvated docking are smaller. For example, the unsolvated docking has 4 medium and 1 acceptable solutions and the solvated docking has 5 medium quality solutions. For the "wet" interfaces, a large fraction of the waters in our docking solutions have positions very close to those in the crystal (Figure 3 and Suppl. Fig. 7-9). These correspond to both fully buried waters and waters present at the rim of the interface. Especially the results from the bound barnase - barstar docking are impressive, with around 80% of the water molecules within 2Å of crystal water positions. The distributions of distances between predicted and native waters in Figure 3 compare favourably with the results from Jiang *et al.* [29]; in that study, no docking was performed, but water positions at the interface were predicted from the crystal structures of a set of complexes. We also found that the quality of the water predictions does not change much after the semi-flexible refinement (see Suppl. Fig. 9). Note however that those are only preliminary results and the flexible refinement protocol needs further optimization.

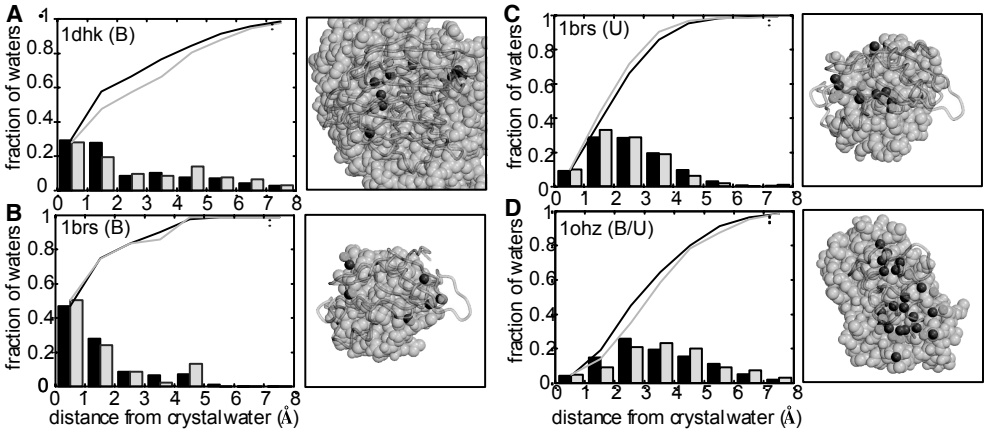


***Figure 3.** Accuracy of predicted water molecules in solvated docking. **(A)** α-amylase – αAI (bound docking); **(B)** barnase – barstar (bound docking); **(C)** barnase – barstar (unbound docking); and **(D)** cohesin – dockerin (bound/unbound docking). Left panel: Histograms (bars) and cumulative fractions (lines) of closest distances between modelled and crystal waters are shown for all acceptable structures (black) and for the top 10 acceptable structures (light grey) out of the top 200 ranked structures. Right panel: view of the best-scoring acceptable solvated docking solution, together with its predicted waters (red) and the corresponding ones in the crystal (green).

We analyzed the recovery of totally buried crystal water molecules over all acceptable (i-RMSD < 4Å) solutions out of the top 200 ranked models (Table 4 and Suppl. Table 9). On average, each docking solution contains between 6 and 12 water molecules (both buried and rim). Buried water molecules are generally more consistently recovered (i.e. found in a larger fraction of the solutions) than those at the rim of the interface (Figure 4 and Suppl. Fig. 9-11). On average, 94% of the buried crystal waters are recovered and each one is observed in 17%

of the acceptable solutions. We find that those crystal waters that are not recovered are making most of their contacts with only one of the two components of the complex.

We also analyzed the fraction of native water-mediated contacts recovered after flexible refinement: this is on average 30% for all acceptable structures, 46% for the highest-ranked acceptable structure and even 66% in the most favorable case. These are quite high fractions considering that on average, per structure, only 32% of the crystal waters are recovered within 4Å. Those numbers are on average 25% smaller for rigid body docking solutions. As was already observed previously [45], flexible refinement significantly improves the fraction of native contacts across the interface. In CAPRI, high/medium/acceptable – quality solutions require at least 50/30/10% fraction native contacts.

Table 4. Recovery of water molecules in solvated docking.[a]

		Rigid body		**Refined**	
	<#waters>[b]	recovery[c] #	$f_{recover}$ (%)[d]	recovery[c] #	$f_{recover}$ (%)[d]
1brs	10.4 (2.6)	14 /18	12 (11)	14/18	14 (12)
2sic	11.7 (4.0)	7/8	29 (19)	7/8	29 (17)
7cei	8.2 (3.2)	8/8	5 (1)	8/8	3 (2)
1ppe	11.0 (3.7)	6/6	12 (10)	6/6	14 (9)
1ohz (U/B)	5.6 (4.6)	5 /5	32 (11)	5/5	13 (8)
1gcq	10.9 (3.0)	4/4	34 (13)	4/4	28 (13)
1avx	8.9 (3.3)	1/1	10	1/1	11

[a] Solvated docking results for the acceptable solutions out of the top 200 models for the various test-cases (see Table 3). (B) and (U) indicate bound and unbound docking, respectively.
[b] Average number (standard deviation) of water molecules per structure.
[c] Number of fully buried crystal waters recovered (i.e. within 2.0Å of a modeled water) / total number of buried crystal waters (see Table 2).
[d] Average fraction (standard deviation) of acceptable structures in which a fully buried water is recovered.

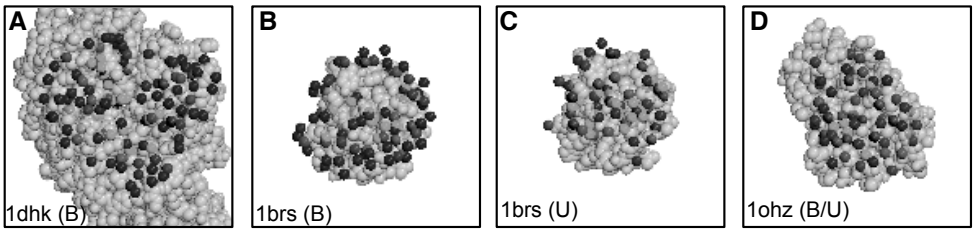


***Figure 4.** Recovery of interfacial water molecules in solvated docking. **(A)** α-amylase – αAI (bound docking); **(B)** barnase – barstar (bound docking); **(C)** barnase – barstar (unbound docking); and **(D)** cohesin – dockerin (bound/unbound docking). For each complex, the largest component is shown with its associated crystal waters (transparent green) together with cluster representatives from all predicted water in the acceptable solutions. The latter are colour-coded according to the fraction of acceptable structures in which they are observed, from blue 0% to red 40% (maximal observed fraction). Waters from all acceptable solutions were clustered based on pairwise distances using a 2.5Å cut-off.

Crystal waters are recovered not only in "wet" interfaces (e.g. α-amylase - αAI and barnase - barstar) but also for example in the case of 1gcq, where all 4 fully buried interface waters are found in several of the docking solutions (this complex shows the highest average fraction of structures in which crystal waters are observed (34%)). For the "dry" PKC interacting protein, the water molecules in the resulting docked structures are placed mostly at the rim

of the interface. The same applies to E2A-HPr. For the latter, however, we cannot compare their positions to experimental ones since the reference complex was solved by NMR. While decreasing somewhat the number of acceptable solutions for that particular complex, explicit inclusion of water led to an improvement in the ranking and in the number of medium quality solutions, both before and after flexible refinement. Taken all together, these results demonstrate the general applicability of our method.

Explicit inclusion of water molecules in our solvated docking protocol results in a factor 3 to 4 increase in computational time requirements for the rigid body docking stage. The most time-consuming part of HADDOCK is however the semi-flexible refinement stage, in which the presence of some additional water molecules does not make much difference. Explicit inclusion of water in docking thus only results in about a factor 2 increase in the overall run time, which is reasonable considering the improvements in both success rate and accuracy, and the fact that as a result water positions are predicted.

Conclusions and Perspective

For the first time, water has been explicitly introduced in protein-protein docking. We followed for this purpose a strategy mimicking the concept of the solvated initial encounter complex. By performing the docking from solvated protein chains in combination with a Monte Carlo water removal procedure based on water contact propensities, we successfully recovered interfacial crystal water molecules and improved our docking results both in bound and unbound docking cases. Further improvements could be achieved by making use of the geometrical information obtained from the cluster analysis of water-mediated contacts.

The very promising results obtained here and the rather reasonable additional computational burden make us confident that solvated docking is a viable approach to model biomolecular complexes. We actually started applying solvated docking in the last two rounds of CAPRI (targets 25 and 26; see http://capri.ebi.ac.uk) but will have to wait for the release of the targets in order to assess its performance. Solvated docking should also benefit the field of protein-DNA modelling since it is well-known that protein-DNA complexes have rather wet interfaces. We therefore intend to extend our approach to the modelling of such complexes, which, as we recently demonstrated, can be successfully modelled using HADDOCK [55].

Acknowledgements

This work was supported by a “Jonge Chemici” grant (grant number 700.50.512) from the Netherlands Organization for Scientific Research (N.W.O.) to AB and by the European Community, FP6 STREP project “ExtendNMR” (contract no. LSHG-CT-2005-018988).

References

1. Mendez, R., Leplae, R. *et al.* Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. Proteins (2005), **60,** 150-169.
2. Halperin, I., Ma, B. *et al.* Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins (2002), **47,** 409-43.
3. van Dijk, A. D. J., Boelens, R. *et al.* Data-driven docking for the study of biomolecular complexes. FEBS Journal (2005), **272,** 293-312.
4. Chandler, D. Interfaces and the driving force of hydrophobic assembly. Nature (2005), **437,** 640-647.
5. Lo Conte, L., Chothia, C. *et al.* The atomic structure of protein-protein recognition sites. J. Mol. Biol. (1999), **285,** 2177-2198.
6. Robert, C. H. & Ho, P. S. Significance of Bound Water to Local Chain Conformations in Protein Crystals. Proc. Natl. Acad. Sci. USA (1995), **92,** 7600-7604.
7. Rashin, A. A., Iofin, M. *et al.* Internal cavities and buried waters in globular proteins. Biochemistry (1986), **25,** 3619-25.

8. Hubbard, S. J., Gross, K. H. *et al.* Intramolecular Cavities in Globular-Proteins. Protein Eng. (1994), **7,** 613-626.
9. Wade, R. C., Clark, K. J. *et al.* Further Development of Hydrogen-Bond Functions for Use in Determining Energetically Favorable Binding-Sites on Molecules of Known Structure .1. Ligand Probe Groups with the Ability to Form 2 Hydrogen-Bonds. J. Med. Chem. (1993), **36,** 140-147.
10. Wade, R. C. & Goodford, P. J. Further Development of Hydrogen-Bond Functions for Use in Determining Energetically Favorable Binding-Sites on Molecules of Known Structure .2. Ligand Probe Groups with the Ability to Form More Than 2 Hydrogen-Bonds. J. Med. Chem. (1993), **36,** 148-156.
11. Raschke, T. M. Water structure and interactions with protein surfaces. Curr. Opin. Struct. Biol. (2006), **16,** 152-159.
12. Loris, R., Langhorst, U. *et al.* Conserved water molecules in a large family of microbial ribonucleases. Proteins (1999), **36,** 117-134.
13. Carugo, O. Correlation between occupancy and B factor of water molecules in protein crystal structures. Protein Eng. (1999), **12,** 1021-1024.
14. Carugo, O. & Bordo, D. How many water molecules can be detected by protein crystallography? Acta Crystallogr. D (1999), **55,** 479-483.
15. Houborg, K., Harris, P. *et al.* Impact of the physical and chemical environment on the molecular structure of Coprinus cinereus peroxidase. Acta Crystallogr. D (2003), **59,** 989-996.
16. Zhang, X. J. & Matthews, B. W. Conservation of Solvent-Binding Sites in 10 Crystal Forms of T4-Lysozyme. Protein Sci. (1994), **3,** 1031-1039.
17. Sreenivasan, U. & Axelsen, P. H. Buried Water in Homologous Serine Proteases. Biochemistry (1992), **31,** 12785-12791.
18. Tame, J. R. H., Sleigh, S. H. *et al.* The role of water in sequence-independent ligand binding by an oligopeptide transporter protein. Nat. Struct. Biol. (1996), **3,** 998-1001.
19. Mustata, G. & Briggs, J. M. Cluster analysis of water molecules in alanine racemase and their putative structural role. Protein Eng. (2004), **17,** 223-234.
20. Babor, M., Sobolev, V. *et al.* Conserved positions for ribose recognition: Importance of water bridging interactions among ATP, ADP and FAD-protein complexes. J. Mol. Biol. (2002), **323,** 523-532.
21. Boobbyer, D. N. A., Goodford, P. J. *et al.* New Hydrogen-Bond Potentials for Use in Determining Energetically Favorable Binding-Sites on Molecules of Known Structure. J. Med. Chem. (1989), **32,** 1083-1094.
22. Schymkowitz, J. W. H., Rousseau, F. *et al.* Prediction of water and metal binding sites and their affinities by using the Fold-X force field. Proc. Natl. Acad. Sci. USA (2005), **102,** 10147-10152.
23. Yang, J. M. & Chen, C. C. GEMDOCK: A generic evolutionary method for molecular docking. Proteins (2004), **55,** 288-304.
24. Osterberg, F., Morris, G. M. *et al.* Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. Proteins (2002), **46,** 34-40.
25. Rejto, P. A. & Verkhivker, G. M. Mean field analysis of FKBP12 complexes with FK506 and rapamycin: Implications for a role of crystallographic water molecules in molecular recognition and specificity. Proteins (1997), **28,** 313-324.
26. Rarey, M., Kramer, B. *et al.* The particle concept: Placing discrete water molecules during protein-ligand docking predictions. Proteins (1999), **34,** 17-28.
27. Verdonk, M. L., Chessari, G. *et al.* Modeling water molecules in protein-ligand docking using GOLD. J. Med. Chem. (2005), **48,** 6504-6515.
28. Moitessier, N., Westhof, E. *et al.* Docking of Aminoglycosides to hydrated and flexible RNA. J. Med. Chem. (2006), **49,** 1023-1033.
29. Jiang, L., Kuhlman, B. *et al.* A "solvated rotainer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. Proteins (2005), **58,** 893-904.
30. Rodier, F., Bahadur, R. P. *et al.* Hydration of protein-protein interfaces. Proteins (2005), **60,** 36-45.
31. Kalodimos, C. G., Biris, N. *et al.* Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. Science (2004), **305,** 386-389.
32. Dominguez, C., Boelens, R. *et al.* HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. J. Am. Chem. Soc. (2003), **125,** 1731-1737.
33. Keskin, O., Tsai, C. J. *et al.* A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. Protein Sci. (2004), **13,** 1043-1055.
34. Brunger, A. T., Adams, P. D. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D (1998), **54 (Pt 5),** 905-21.
35. Lindahl, E., Hess, B. *et al.* GROMACS 3.0: a package for molecular simulation and trajectory analysis. J. Mol. Model. (2001), **7,** 306-317.

36. Daura, X., Gademann, K. *et al.* Peptide folding: When simulation meets experiment. Angew. Chem. Int. Edit. (1999), **38,** 236-240.
37. Linge, J. P., Habeck, M. *et al.* ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics (2003), **19,** 315-316.
38. Jorgensen, W. L., Chandrasekhar, J. *et al.* Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. (1983), **79,** 926-935.
39. Jorgensen, W. L. & Tirado-rives, J. The OPLS Potential functions for proteins. Energy minimizations for crystals of cyclin peptides and crambin. J. Am. Chem. Soc. (1988), **110,** 1657-1666.
40. Linge, J. P., Williams, M. A. *et al.* Refinement of protein structures in explicit solvent. Proteins (2003), **50,** 496-506.
41. Fernandez-Recio, J., Totrov, M. *et al.* Identification of protein-protein interaction sites from docking energy landscapes. J. Mol. Biol. (2004), **335,** 843-865.
42. Mintseris, J., Wiehe, K. *et al.* Protein-protein docking benchmark 2.0: An update. Proteins (2005), **60,** 214-216.
43. Wang, G., Louis, J. M. *et al.* Solution structure of the phosphoryl transfer complex between the signal transducing proteins HPr and IIA(glucose) of the Escherichia coli phosphoenolpyruvate:sugar phosphotransferase system. EMBO J. (2000), **19,** 5635-49.
44. Carvalho, A. L., Dias, F. M. V. *et al.* Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. Proc. Natl. Acad. Sci. USA (2003), **100,** 13809-13814.
45. van Dijk, A. D. J., de Vries, S. J. *et al.* Data-driven docking: HADDOCK's adventures in CAPRI. Proteins-Structure Function and Bioinformatics (2005), **60,** 232-238.
46. BompardGilles, C., Rousseau, P. *et al.* Substrate mimicry in the active center of a mammalian alpha-amylase: Structural analysis of an enzyme-inhibitor complex. Structure (1996), **4,** 1441-1452.
47. Buckle, A. M., Schreiber, G. *et al.* Protein-Protein Recognition - Crystal Structural-Analysis of a Barnase Barstar Complex at 2.0-Angstrom Resolution. Biochemistry (1994), **33,** 8878-8889.
48. Takeuchi, Y., Satow, Y. *et al.* Refined Crystal-Structure of the Complex of Subtilisin Bpn' and Streptomyces Subtilisin Inhibitor at 1.8 a-Resolution. J. Mol. Biol. (1991), **221,** 309-325.
49. Ko, T. P., Liao, C. C. *et al.* The crystal structure of the DNase domain of colicin E7 in complex with its inhibitor Im7 protein. Structure (1999), **7,** 91-102.
50. Bode, W., Greyling, H. J. *et al.* The Refined 2.0 a X-Ray Crystal-Structure of the Complex Formed between Bovine Beta-Trypsin and Cmti-I, a Trypsin-Inhibitor from Squash Seeds (Cucurbita-Maxima) - Topological Similarity of the Squash Seed Inhibitors with the Carboxypeptidase a Inhibitor from Potatoes. Febs Lett. (1989), **242,** 285-292.
51. Nishida, M., Nagata, K. *et al.* Novel recognition mode between Vav and Grb2 SH3 domains. EMBO J. (2001), **20,** 2995-3007.
52. Song, H. K. & Suh, S. W. Kunitz-type soybean trypsin inhibitor revisited: Refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from Erythrina caffra and tissue-type plasminogen activator. J. Mol. Biol. (1998), **275,** 347-363.
53. Lima, C. D., Klein, M. G. *et al.* Structure-based analysis of catalysis and substrate definition in the HIT protein family. Science (1997), **278,** 286-290.
54. Hubbard, S. J. & Thornton, J. M. (1993) *NACCESS* (Department of Biochemistry and Molecular Biology, University College, London).
55. van Dijk, M., van Dijk, A. D. J. *et al.* Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. Nucl. Acids Res. (2006), **34,** 3317-3325.

Chapter 6.
Data-driven docking: HADDOCK's adventures in CAPRI.

Aalt D.J. van Dijk, Sjoerd J. de Vries, C. Dominguez, Huiling Chen, Huang-Xian Zhou and Alexandre M.J.J. Bonvin, *Proteins: Structure, Function and Bioinformatics* **60** (2005), 232 – 238.

Abstract

We have shown previously that, given high-resolution structures of the unbound molecules, structure determination of protein complexes is possible by including biochemical and/or biophysical data as highly ambiguous distance restraints in a docking approach. We applied this method, implemented in the HADDOCK (High Ambiguity Driven DOCKing) package [Dominguez *et al.*, *J. Am. Chem. Soc.* **2003**, *125*, 1731], to the targets in the fourth and fifth rounds of CAPRI (Critical Assessment of PRedicted Interactions). Here we describe our results and analyze them in detail. Special attention is given to the role of flexibility in our docking method, and the way in which this improves the docking results. We also describe extensions to our approach that were developed as a direct result of our participation to CAPRI. In addition to experimental information, we also included interface residue predictions from PPISP [Zhou and Shan, *Proteins* **2001**, *44*, 336], a neural network method. Using HADDOCK we were able to generate acceptable structures for 6 of the 8 targets, and to submit at least one acceptable structure for 5 of them. Of these 5 submissions, 3 were of medium quality (target 10, 11 and 15) and 2 of high quality (target 13 and 14). In all cases, predictions were obtained containing at least 40% of the correct epitope at the interface for both ligand and receptor simultaneously.

Introduction

Biochemical and biophysical experiments such as mutagenesis, NMR and mass-spectrometry are widely used to gain insight into biomolecular interactions. The information generated in this way can in principle be used to model the structure of the corresponding complex when conventional NMR and crystallographic approaches fail. Taking the step from experimental data to modeling is however not common practice. This can be done using docking approaches that model the structure of a complex based on the structure of the constituents. Although clear progress has been achieved in the field of 'ab-initio docking', as illustrated by the previous rounds of CAPRI (Critical Assessment of PRedicted Interactions) [1], most current approaches have difficulties in generating consistently reliable predictions. However, as highlighted in a recent review [2], in many cases of biological interest, some kind of experimental information is available which can be used to filter docking solutions or even to drive the docking. We developed for this purpose the data-driven docking method HADDOCK that can incorporate any kind of information about interface residues [3]. Using HADDOCK, we participated to rounds 4 and 5 of CAPRI. Our method and its performance within CAPRI are presented here.

Materials and Methods

HADDOCK: data

Experimental and/or prediction information is incorporated in HADDOCK by defining active residues (which based on the data are supposed to be part of the interface) and passive residues (surface neighbors of active residues). The docking is driven by ambiguous interaction restraints (AIRs) defined between any atom of the active residues and all atoms of all active and passive residues on the partner protein [3]. In cases where the data were very fuzzy or judged unreliable, we randomly removed 25% of the data for each docking trial. In the absence of experimental data, we experimented with the use of all accessible residues as active and/or passive residues, in the combinations: A-active B-passive, A-passive B-active and both A and B active.

HADDOCK: flexible docking protocol

Flexibility is introduced at several levels in the algorithm:

i) by docking from ensembles of structures and taking all possible pairwise combinations,
ii) by introduction of flexibility in the side-chain at the interface, and
iii) by allowing both side-chain and backbone flexibility in the final refinement stage.

This is then followed by a final refinement in explicit solvent. In the rigid body docking stage we typically save to disk 1000 to 2000 solutions. The best 200 are then subjected to a refinement step consisting of 3 consecutive simulated annealing, first treating the molecules as rigid bodies and then with introduction of flexibility (ii and iii above); for details see [3].

For each target we typically performed a number of docking runs with various definitions of the AIRs. The initial scoring was based on the sum of intermolecular vanderWaals, electrostatic and AIR energies. The non-bonded energies are calculated with a 8.5Å cutoff using the OPLS parameters [4]. In round 5, we set ε to 10 during the vacuum part of the docking (rigid body docking and semi-flexible refinement). The electrostatic energy contribution was scaled down by a factor 0.1 in the final scoring after water refinement. The scoring was performed on a cluster basis considering only the best 5 to 10 structures of each cluster to remove cluster size effects. The clustering is based on pairwise RMSDs. The lowest energy structure of the lowest energy cluster is considered the highest ranking solution. Due to the inclusion of experimental information only a limited number of solutions is typically obtained (<25). For CAPRI, the selection from the clusters from different docking runs was performed manually based on energy considerations and visual inspection.

Interface predictions

In PPISP, sequence profiles produced by PSI-BLAST and solvent accessibility of spatially neighboring surface residues calculated by DSSP [5] were used as input to a neural network [6]. The network was trained on interface residues collected from the protein-protein complexes in the Protein Data Bank. A residue is considered as a surface residue if at least 10% of its surface area is solvent accessible. A surface residue is considered as an interface residue if at least one of its heavy atoms is within 5 Å of a heavy atom of the partner protein. Since the original publication of PPISP, extensive improvements have been made (Chen and Zhou, personal communication). The training set now consists of 1,156 protein chains with less than 30% identity. These chains each have at least 20 interface residues. The training set contains a total of 225,238 surface residues, of which 52,624 (or 23%) are interface residues. Test on a different set of 100 non-homologous protein chains shows that the overall prediction accuracy is ~80%. One problem was that interface residues were over-predicted for some proteins but under-predicted for others. To tackle this problem, we developed a consensus method that combines predictions from a series of neural networks with different levels of accuracy and coverage of native interface residues (F_{IR}). Usually a neural network with higher F_{IR} predicts more interface residues but with lower accuracy, and vice versa. Each predicted interface residue was ranked by consensus score (the number of neural networks predicting it as interface) and the top ranked ones were mapped onto the protein surface. The residues clustered together were collected. This process was stopped early if there were enough interface residues collected (to prevent over-prediction), or extended to less confidently predicted residues if there were only a few predictions made (to avoid under-prediction).

Results and Discussion

HADDOCK data

Information to drive the docking process was derived from literature searches, in combination with interface residue predictions by PPISP [6] (see below). This information is highly

ambiguous, in the sense that it gives information about the interface but not about contacts made across it. In order to use such information to drive the docking, we distinguish between "active" and "passive" residues: active residues correspond to solvent accessible, experimentally identified or predicted interface residues while passive residues correspond to their surface neighbors. These are used to define ambiguous interaction restraints (AIRs) between each active residue of one chain and all active and passive residues of the other chain (for further discussion, see Dominguez *et al.* [3]).

Performance of the PPISP interface residues predictions

PPISP is a neural network based method that is trained on known structures of protein complexes. The input of the neural network consists of sequence profiles and solvent accessibility of spatially neighboring residues. Sequence profiles, obtained from multiple sequence alignment by PSI-BLAST, capture characteristics of interface residues such as conservation and hydrophobicity. The PPISP interface predictions for CAPRI 4 and 5 targets are summarized in Table 1. For a total of 11 proteins in the targets (no prediction was necessary for the two antibodies), four (TBE monomer of T10, cohesin of T11/T12, MYPT1 of T14, and colicin D of T15) had higher than 50% prediction accuracy, and two others (dockerin of T11/T12 and the xylanase of T18) had very good values for F_{IR} (1 and 0.75, respectively) but moderate accuracy. Interface predictions were poor for the two antigens (in T13 and T19) and two large proteins (PP-1 of T14 and TAXI of T18). These predictions were used to define AIRs (see above) to drive the docking.

Table 1. Results of interface residue prediction on CAPRI targets

Target	Chain	Seq. length	#Native int. res.	PPISP predictions			HADDOCK	
				#Pred.	F_{IR} [a]	Accuracy[b]	F_{IR} [a]	Accuracy[b]
10	TBE monomer	381	95	19	0.11	0.53	0.44	0.90
11/12	Cohesin	140	25	13	0.48	0.92	0.89	0.91
	Dockerin	56	13	32	1	0.41	0.74	0.67
13	antigen	245	25	5	0	0.0	0.88	0.79
14	PP-1	309	49	10	0	0.0	0.81	0.85
	MYPT1	291	60	36	0.32	0.53	0.67	0.85
15	Colicin D	107	22	20	0.59	0.65	0.85	0.85
	Immunity protein	87	19	14	0.11	0.14	0.85	0.79
18	TAXI	370	23	25	0.04	0.04	0.42	0.38
	A. niger xylanase	182	24	51	0.75	0.35	0.62	0.71
19	Ovine prion	102	21	15	0.14	0.20	0.44	0.42

[a]) F_{IR}: fraction of correctly predicted (by PPISP) or observed (HADDOCK) interface residues among all native interface residues. Native interface residues were defined as those forming native contacts (< 5Å).

[b]) Accuracy: fraction of correct interface residues among all predicted (by PPISP) or observed (HADDOCK best model) interface residues.

HADDOCK results for the CAPRI 4 & 5 targets

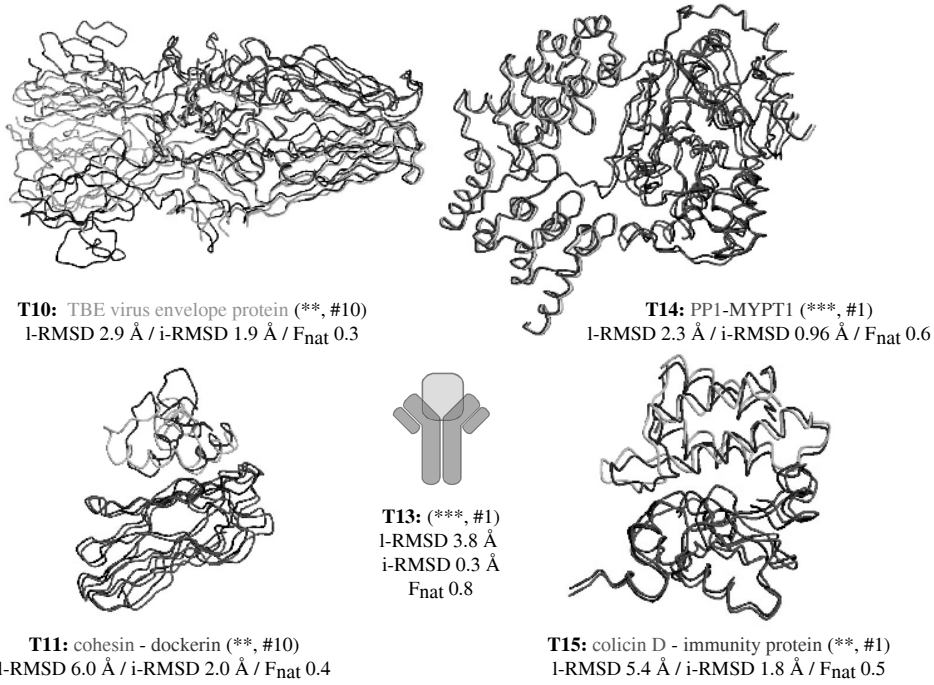
For 5 of the 8 targets (including the cancelled target 15), we had at least one acceptable solution in our submissions; of these three were of medium quality and two of high quality. The corresponding models superimposed on their respective targets are shown in Figure 1 (the still unpublished target 13 is shown as cartoon). The ranking we report for the various targets is based on the fraction of correct native contacts. In all 5 cases we ended within the top 10 of all submissions. In three cases, our best prediction was correctly ranked number one within our ten submissions. In the following we discuss our results for the individual targets.

Target 10. For target 10 [7], we modified the docking protocol in order to be able to deal with three molecules and impose C3 symmetry. This was done by using symmetry restraints as introduced for NMR structure calculation of symmetrical dimers [8]. To impose C3 symmetry, we defined triplets of distance pairs (AB/BC, BC/AC and CA/AB), requiring that the distances must be equal within a pair. We used in addition non-crystallographic symmetry restraints in CNS [9], which enforce that the molecules be similar without defining any symmetry operation between them. The ambiguous interaction restraints (AIRs) used to drive the docking were derived from various sources [10-13] and included epitope mapping, mutagenesis and sequence conservation data, in combination with interface predictions.

The results of the docking could be partitioned into two classes: one with rather 'flat' triangular arrangements of the three monomers, and one consisting of 'spikes'. We based our selection on electron microscopy data [14] that suggested a triangular flat shape of the timer. Accordingly, we mainly selected such structures, and only put one 'spike' conformation in our submission, which we ranked tenth. The latter turned out to be a very close prediction, with a ligand RMSD of 2.9Å, an interface RMSD of 1.9Å and a fraction of native contacts of 0.3. It is worth noting that this "spike" solution had the best intermolecular energy but that misjudgment of literature data led us to rank it last.

Target 11/12. For target 11 [15], a homology model of dockerin was built using SWISSMODEL [16] and WHATIF [17]. The model was subjected to a 200 ps molecular dynamics (MD) simulation in explicit solvent using Gromacs3.1.4 [18]. The starting structure plus ten structures taken every 20 ps were selected as input for the docking. The backbone RMSD of our starting homology model from the bound form was 4.3 Å, whereas it ranged from 4.1 Å to 4.8 Å for the structures taken from the MD-trajectory. These values however decrease to 2.3 Å for the model and 2.2 Å to 2.9 Å for the MD structures if only the two Ser/Thr pairs that are known to be important for the binding are considered. This means that some of the MD structures had moved closer to the bound form than the homology-model itself. The unbound form of cohesin was also subjected to 200 ps MD in explicit solvent. Again, 11 structures were selected for the docking. These were also used for target 12.

To drive the docking, we used previously published mutagenesis data [19-22], in particular two amino acid pairs Ser11/Thr12 and Ser45/Thr46, and other predicted interface residues. As discussed in the paper describing the experimental crystal structure of the complex [15], the dockerin sequence contains a tandem repeat, with residues 1-23 showing high homology to residues 35-57. These two stretches of sequences also adopt very similar 3D structures (main chain atom RMSD 0.36Å). Moreover, dockerin contains near perfect internal twofold symmetry, such that residues 1-22 overlay onto residue 35-56, and vice versa (see figure 5 in ref [15]). This symmetry is also reflected in the mutagenesis data that we used for docking. As stated in ref [15], it would seem likely that both symmetry-related halves of dockerin could interact with cohesin in almost identical manners. The accuracy of our docking results, especially for target 12, were mainly hampered by the fact that both sites were included in our restraints while only the second one is actually involved in binding in the crystal structure. Since both sites were defined, we also obtained several docking solutions corresponding to a 180° rotated binding mode. For target 11 we nevertheless had 1 medium and 3 acceptable solutions in our submission (with the best one having a ligand RMSD of 6.0Å, an interface RMSD of 2.0 Å, and a fraction of native contacts of 0.4 ranking at the third position overall). For target 12 no acceptable solutions were submitted, although acceptable ones within 2.3 Å interface RMSD were generated. However, if the symmetry related binding mode would be considered, all our submissions for target 11 and 12 would be within 10Å ligand RMSD.



T10: TBE virus envelope protein (**, #10)
l-RMSD 2.9 Å / i-RMSD 1.9 Å / F_{nat} 0.3

T14: PP1-MYPT1 (***, #1)
l-RMSD 2.3 Å / i-RMSD 0.96 Å / F_{nat} 0.6

T13: (***, #1)
l-RMSD 3.8 Å
i-RMSD 0.3 Å
F_{nat} 0.8

T11: cohesin - dockerin (**, #10)
l-RMSD 6.0 Å / i-RMSD 2.0 Å / F_{nat} 0.4

T15: colicin D - immunity protein (**, #1)
l-RMSD 5.4 Å / i-RMSD 1.8 Å / F_{nat} 0.5

***Figure 1. Best HADDOCK models and corresponding experimental structures.** Closest models overlaid with corresponding crystal structures for target 10 (pdb: 1urz[7]), 11 (pdb: 1ohz[15]), 14 (pdb: 1s70[23]) and 15 (pdb: 1v74[24]). Color coding: red and blue for crystal structures, purple and light-blue for model; for target 10 in addition: green for crystal structure, black for model. Note that target 10 is a trimer; for clarity, for two of its (C3-symmetric) chains we show only part of the chain. Segments 148-159 and 204-209 are missing in the crystal structure. The still unpublished target 13 is shown as a cartoon. The quality of our best model (stars) and its rank within our ten submissions is indicated between brackets. The CAPRI quality criteria are: "high" (***): F_{nat} > 0.5, l-RMSD or i-RMSD < 1.0 Å and "medium" (**): F_{nat} > 0.3, l-RMSD < 5.0 Å or i-RMSD < 2.0 Å, where F_{nat} is the fraction of native contacts, l-RMSD the ligand backbone RMSD from the target, after superimposition on the receptor and i-RMSD the interface residues backbone RMSD from the target.

Target 13. For target 13, the bound form of the antibody and the unbound form of the antigen were provided. The latter was subjected to a 200 ps MD in explicit solvent and 11 structures were selected as input for the docking as described for targets 11/12. In this case, the unbound form was 0.5 Å away from the bound form, and it drifted away to 0.9 Å during the MD simulation.

For the antibody we defined the CDR region as active residues. For the antigen, several epitopes were identified from literature data [25-27]. This information was introduced differently in various runs: either using all epitopes simultaneously or only a subset of them to define AIRs. We also performed a docking run in which all accessible residues of the antigen were defined as passive.

The clusters from all runs were pooled and ranked mainly based on favorable values of van der Waals energy and on packing arguments after visual inspection. Our number one submission was of high-quality (l-RMSD 3.8 Å; i-RMSD 0.3 Å; F_{nat} 0.8).

Target 14. The docking was performed from the bound form of MYPT1 and a homology model of PP1 created using SWISSMODEL [16] and WHATIF [17]. For this target [23], excellent information was available from literature [28-30] including a crystal structure of the RRVSFA peptide bound to PP1. In addition, interface predictions from PPISP involving the N-terminal helix of MYPT1 were used. Because of the rather closed, "embracing" conformation of MYPT1, which might prevent a proper rigid body docking, we turned off the

intermolecular interactions of PP1 with the C-terminal residues of MYPT1 starting from residue 52 and scaled down the intermolecular interactions by a factor 0.01. Interaction with the C-terminal residues was reintroduced in the second flexible refinement stage of the docking. As the quality of the data in this case was really high, it was not surprising that one of our predictions was of high quality (l-RMSD 2.3Å; i-RMSD 0.96Å; F_{nat} 0.6) with two additional acceptable ones out of our 10 submitted structures.

Target 15. Target 15 [24] was a bound docking case, but with shaved surface side-chains. The latter were generated with CNS [9] and the structures were submitted to a short refinement in explicit solvent [31] with position restraints on the backbone. For each protein, 10 structures were generated and used as starting point for the docking. Only little data were found in the literature [32]. We used as always predicted interface residues but also experimented (as for target 13, but now for both proteins) taking all accessible residues as passive and/or active, in the combinations active-passive, passive-active and active-active. The results of these three docking runs were analyzed in terms of the frequency that a given residue contacts the partner protein. The top 10% most frequently found residues at the interface were then selected for a new docking run.

Our number one submission, resulting from a run with the latter restraints, was a close hit (l-RMSD 5.4 Å; i-RMSD 1.8 Å; F_{nat} 0.5) corresponding to a medium quality prediction. Out of the three runs with a total of 600 structures from which we selected our submissions, only 5 structures had l-RMSD below 10 Å (resulting from the runs with all accessible residues on the immunity protein as passive). Our selection criterion (based on Evdw + 0.1 Eelec) thus correctly led to the best conformation in this case.

Target 18. For target 18 [33], the predicted interface residues were quite good for the xylanase with 75% of the epitope correctly predicted (see F_{IR}, fraction of correctly predicted native interface residues, in Table I). In addition, some mutagenesis information was available[34]. However, for TAXI the interface prediction was poor (F_{IR} = 0.04), and we did not find any experimental data for defining the interface. Although we tried the same kind of active/passive protocols as for target 15, in this case our docking runs did not generate any acceptable solution. We did however obtain solutions with more than 42% of the native epitope of each partner simultaneously at the interface but these corresponded to rotated solutions.

Target 19. For target 19 [35], the ovine prion was modeled as described above from the NMR entry 1DWY [36]. We found various epitopes in the literature [37, 38], which we used again in various combinations of active and passive residues, together with the CDR residues of the antibody. A docking run was also performed using all accessible residues of the antigen. However, no acceptable solution was obtained. As for target 18, our submissions however contained rotated solutions containing as much as 44 and 79% of the correct epitopes at the interface for the antibody and the antigen, respectively.

The effect of flexibility on our docking results

In our docking protocol, flexibility is introduced stepwise, first for interface side-chains and then for both backbone and side-chains at the interface (see Materials and Methods). To analyze in detail if it is worth paying the additional computational price, we compared various quality parameters for structures after rigid body docking and after final refinement in explicit solvent (Table 2). Although there is some variability between targets, it is clear that the number of good solutions as monitored by ligand RMSDs increases when comparing the rigid body docking results with the results after flexibility has been introduced. The quality of the prediction also improves significantly as can clearly be seen from the fraction of native contacts. Finally, flexibility also considerably improved the ranking of structures.

Table 2. Impact of flexibility on docking results.

Target	10	11	12	13	14	15
Number of solutions within the given l-RMSD range: rigid body/refined[a]						
0-5Å	9/8	36/63	33/36	26/26	11/12	2/3
Average fraction of native contacts [a] Rigid body:						
0-5Å	0.14 (0.05)	0.27 (0.11)	0.11 (0.03)	0.57 (0.07)	0.54 (0.15)	0.45 (0.07)
Average fraction of native contacts [a] Refined:						
0-5Å	0.18 (0.09)	0.36 (0.11)	0.12 (0.03)	0.74 (0.07)	0.50 (0.10)	0.35 (0.13)
Ranking of best submitted structure: rigid body/refined [b]						
	6/3	181/51	-	36/1	23/2	8/2

[a] For target 11, the RMSD range is 5 - 7.5Å ; for target 12 and 15 it is 5 – 10Å.
[b] Rank of best submitted structure according to our scoring scheme. Lower ranking indicates better result.

In addition to the explicit inclusion of flexibility in the refinement stage, we also implicitly included it in the rigid body docking stage by starting from ensembles of structures obtained from short MD simulations in explicit solvent. Depending on the quality of the starting model, the RMSD to the bound form increased or decreased during the MD. Still, we often observed that some conformations preferentially lead to better docking solutions. This effect must originate from better side-chain conformations since large backbone conformational changes cannot be expected within such short simulation times (200 ps). For example, for target 11, our best model originated from MD structures taken at 40 ps for cohesin (backbone RMSD to bound form 0.5 Å) and 200 ps for dockerin (backbone RMSD to bound form 4.6Å). For target 13, good solutions were obtained mainly from the unbound form of the antigen (backbone RMSD to bound form 0.5Å) and from a MD snapshot taken at 140 ps (backbone RMSD to bound form 0.75Å).

What did we learn from CAPRI?

CAPRI was a very good stimulus to develop new tools, as well as a possibility to validate new features in an unbiased way. As explained above, the fact that we did often not have very reliable experimental data (or no data at all) inspired us to experiment with all accessible residues as active-active, active-passive or passive-active. For target 13 and 15, this was successful, while for target 18 and 19 this did not generate any acceptable solution. Our solutions for the latter two cover however the correct epitopes but correspond to rotated solutions. Although there are several possible reasons for this difference, it might be related to the fact that target 18 and 19 consisted of bigger proteins, which means that the configuration space that needs to be covered is larger; it is therefore possible that our sampling was insufficient. In cases where data were fuzzy or scarce we used all accessible residues and also experimented with the random removal of a fraction (typically 25%) of our data for each solution that was generated. In this way, the fuzziness of the interface definition is increased, and wrong or inconsistent data need not have a disastrous influence. We started using this feature after the results for round 4 were known, when we realized the effect that the wrong definition of active residues for dockerin had on our docking of target 12. Note that random removal of as much as 50% of the restraints leads to high quality predictions for target 12 (data not shown). Considering the interface predictions, the blind test results on the CAPRI targets were consistent with results on other test sets (Chen and Zhou, personal communication). In general, predictions for small proteins are much better than for large proteins. Complexes of small proteins on which our training set is based are indeed well represented in the PDB. In terms of classes of protein complexes, the PPISP method generally

makes good prediction for enzyme-inhibitor interfaces but is not suited for antigen-antibody interactions. The former complexes have presumably evolved over time to optimize the interface. In contrast, antigen-antibody interactions are not subjected to evolutionary optimization. A strength of PPISP is that it is relatively insensitive to conformational changes accompanying complex formation. Results on CAPRI targets also showed that PPISP performed equally well with homology models (dockerin in target 11).

We also analyzed in detail the influence of flexibility on our docking results. As described here and in related studies [39, 40], flexibility is a very important factor for docking. Indeed, we found that, in general, the docking results improve after the flexible refinement stages. The largest improvement is found in the fraction of native contacts, while the effect of flexibility on the RMSDs from the target is the most pronounced when the unbound form is further away from the bound form. Finally, we also observed that flexibility improves the ranking of correct solutions. Taken all together, this indicates that the inclusion of flexibility in docking is clearly beneficial, even if it increases the required computational time.

Our participation to CAPRI revealed both the strengths and weaknesses of our data-driven docking approach HADDOCK. Good results can be expected when the data are of high quality. Using very fuzzy or ambiguous data, HADDOCK sometimes still generated acceptable results (as for target 13 and 15) but in other cases failed to do so (target 18 and 19). The scoring of flexible docking solutions, which have a significantly larger number of degrees of freedom, remains a difficult process, which can clearly be improved in the future.

Acknowledgments

This work was supported in part by NIH grant GM58187 (to HXZ) and by a “Jonge Chemici” grant from the Netherlands Organization for Scientific Research (N.W.O.) to AB.

References

1. Janin, J., Henrick, K. *et al.* CAPRI: A Critical Assessment of PRedicted Interactions. Proteins (2003), **52,** 2-9.
2. van Dijk, A. D. J., Boelens, R. & Bonvin, A. M. J. J. Data-driven docking for the study of biomolecular complexes. FEBS Journal (2005), **272,** 293-312.
3. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc (2003), **125,** 1731-1737.
4. Jorgensen, W. L. & Tirado-rives, J. The OPLS Potential functions for proteins. Energy minimizations for crystals of cyclin peptides and crambin. J Am Chem Soc (1988), **110,** 1657-1666.
5. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers (1983), **22,** 2577-2637.
6. Zhou, H. X. & Shan, Y. B. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins (2001), **44,** 336-343.
7. Bressanelli, S., Stiasny, K. *et al.* Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. Embo J (2004), **23,** 728-738.
8. Nilges, M. A Calculation Strategy for the Structure Determination of Symmetrical Dimers by H-1-Nmr. Proteins (1993), **17,** 297-309.
9. Brunger, A. T., Adams, P. D. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr (1998), **54 (Pt 5),** 905-21.
10. Stiasny, K., Allison, S. L. *et al.* Structural requirements for low-pH-induced rearrangements in the envelope glycoprotein of tick-borne encephalitis virus. J Virol (1996), **70,** 8142-8147.
11. Allison, S. L., Schalich, J. *et al.* Mutational evidence for an internal fusion peptide in flavivirus envelope protein E. J Virol (2001), **75,** 4268-4275.
12. Stiasny, K., Allison, S. L. *et al.* Membrane interactions of the tick-borne encephalitis virus fusion protein E at low pH. J Virol (2002), **76,** 3784-3790.
13. Allison, S. L., Stiasny, K. *et al.* Mapping of functional elements in the stem-anchor region of tick-borne encephalitis virus envelope protein E. J Virol (1999), **73,** 5605-5612.
14. Ferlenghi, I., Clarke, M. *et al.* Molecular organization of a recombinant subviral particle from tick-borne encephalitis. Mol Cell (2001), **7,** 593-602.

15. Carvalho, A. L., Dias, F. M. V. *et al.* Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. P Natl Acad Sci USA (2003), **100,** 13809-13814.
16. Schwede, T., Kopp, J. *et al.* SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res (2003), **31,** 3381-3385.
17. Rodriguez, R., Chinea, G. *et al.* Homology modeling, model and software evaluation: three related resources. Bioinformatics (1998), **14,** 523-528.
18. Lindahl, E., Hess, B. & van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model (2001), **7,** 306-317.
19. Miras, I., Schaeffer, F. *et al.* Mapping by site-directed mutagenesis of the region responsible for cohesin-dockerin interaction on the surface of the seventh cohesin domain of Clostridium thermocellum CipA. Biochemistry-Us (2002), **41,** 2115-2119.
20. Schaeffer, F., Matuschek, M. *et al.* Duplicated dockerin subdomains of Clostridium thermocellum endoglucanase CelD bind to a cohesin domain of the scaffolding protein CipA with distinct thermodynamic parameters and a negative cooperativity. Biochemistry-Us (2002), **41,** 2106-2114.
21. Mechaly, A., Fierobe, H. P. *et al.* Cohesin-dockerin interaction in cellulosome assembly - A single hydroxyl group of a dockerin domain distinguishes between nonrecognition and high affinity recognition. J Biol Chem (2001), **276,** 9883-9888.
22. Mechaly, A., Yaron, S. *et al.* Cohesin-dockerin recognition in cellulosome assembly: Experiment versus hypothesis. Proteins (2000), **39,** 170-177.
23. Terrak, M., Kerff, F. *et al.* Structural basis of protein phosphatase 1 regulation. Nature (2004), **429,** 780-784.
24. Graille, M., Mora, L. *et al.* Structural inhibition of the colicin D tRNase by the tRNA-mimicking immunity protein. Embo J (2004), **23,** 1474-1482.
25. Godard, I., Estaquier, J. *et al.* Antigenicity and Immunogenicity of P-30-Derived Peptides in Experimental-Models of Toxoplasmosis. Mol Immunol (1994), **31,** 1353-1363.
26. Velge-Roussel, F., Dimier-Poisson, I. *et al.* Anti-SAG1 peptide antibodies inhibit the penetration of Toxoplasma gondii tachyzoites into enterocyte cell lines. Parasitology (2001), **123,** 225-233.
27. Velge-Roussel, F., Chardes, T. *et al.* Epitopic Analysis of the Toxoplasma-Gondii Major Surface-Antigen Sag1. Mol Biochem Parasit (1994), **66,** 31-38.
28. Egloff, M. P., Johnson, D. F. *et al.* Structural basis for the recognition of regulatory subunits by the catalytic subunit of protein phosphatase 1. Embo J (1997), **16,** 1876-1887.
29. Toth, A., Kiss, E. *et al.* Study of the subunit interactions in myosin phosphatase by surface plasmon resonance. Eur J Biochem (2000), **267,** 1687-1697.
30. Hirano, K., Phan, B. C. & Hartshorne, D. J. Interactions of the subunits of smooth muscle myosin phosphatase. J Biol Chem (1997), **272,** 3683-3688.
31. Linge, J. P., Williams, M. A. *et al.* Refinement of protein structures in explicit solvent. Proteins (2003), **50,** 496-506.
32. de Zamaroczy, M., Mora, L. *et al.* Cleavage of colicin D is necessary for cell killing and requires the inner membrane peptidase LepB. Mol Cell (2001), **8,** 159-168.
33. Sansen, S., De Ranter, C. J. *et al.* Structural basis for inhibition of Aspergillus niger xylanase by Triticum aestivum xylanase inhibitor-I. J Biol Chem (2004), **279,** 36022-36028.
34. Gebruers, K., Brijs, K. *et al.* Properties of TAXI-type endoxylanase inhibitors. Bba-Proteins Proteom (2004), **1696,** 213-221.
35. Eghiaian, F., Grosclaude, J. *et al.* Insight into the PrPC -> PrPSc conversion from the structures of antibody-bound ovine prion scrapie-susceptibility variants. P Natl Acad Sci USA (2004), **101,** 10254-10259.
36. Garcia, F. L., Zahn, R. *et al.* NMR structure of the bovine prion protein. P Natl Acad Sci USA (2000), **97,** 8334-8339.
37. Peretz, D., Williamson, R. A. *et al.* A conformational transition at the N terminus of the prion protein features in formation of the scrapie isoform. J Mol Biol (1997), **273,** 614-622.
38. Korth, C., Stierli, B. *et al.* Prion (PrPSc)-specific epitope defined by a monoclonal antibody. Nature (1997), **390,** 74-77.
39. Rajamani, D., Thiel, S. *et al.* Anchor residues in protein-protein interactions. P Natl Acad Sci USA (2004), **101,** 11287-11292.
40. Ehrlich, L. P., Nilges, M. & Wade, R. C. The Impact of Protein Flexibility on Protein-Protein Docking. Proteins-Structure Function and Bioinformatics (2005), **58,** 126.

Chapter 7.
Modelling protein-protein complexes along the cytochrome c oxidase Cu-delivery pathway.

Aalt D.J. van Dijk, Simone Ciofi-Baffoni, Lucia Banci, Ivano Bertini, Rolf Boelens and Alexandre M.J.J. Bonvin, *manuscript in preparation.*

Abstract

Cytochrome c oxidase (the terminal enzyme in the electron transport system) is of vital importance since it generates the proton gradient that drives synthesis of ATP. Its correct assembly depends on proper delivery of copper, a process which has been implicated in several human diseases. Structural details about the interactions in this pathway are however unknown. Here we model several complexes of the human proteins involved in Cu-delivery to cytochrome c oxidase: cox17-sco1, sco1-cox2, cox17-cox11 and cox11-cox11. We make use for this purpose of information-driven docking, with restraints based on bioinformatic interface predictions. The resulting models provide structural information about copper delivery towards cytochrome c oxidase. In particular, we analyze the details of the copper transfer from donating to accepting protein in the case of cox17-sco1, sco1-cox2 and cox17-cox11. For the cox11-cox11 homodimer our model suggests a possible membrane or protein interaction site, which might be involved in copper delivery. Finally, our models provide testable hypotheses for mutagenesis experiments, which we are currently performing.

Introduction

The reduction of molecular oxygen, catalyzed by cytochrome c oxidase (the terminal enzyme in the electron transport system) is of vital importance since it generates the proton gradient that drives synthesis of ATP. Multiple subunits and several cofactors are necessary for catalytic activity including two hemes a, a magnesium ion, a zinc ion, and three copper ions. In particular, the copper ions are located in subunits cox1 and cox2, which contain the CuB and CuA centers, respectively. Cytochrome c oxidase is located in the inner mitochondrial membrane and the copper binding subunits are encoded by the mitochondrial genome, so copper insertion must occur within this organelle. The insertion of these cofactors and assembly of the cytochrome c oxidase complex requires accessory proteins [1], but not much details are known about the assembly process.

In particular, the mechanism of copper delivery to cytochrome c oxidase is a long-standing puzzle. In eukaryotes, copper insertion involves the key copper metallochaperone cox17 which acts as a donor of Cu(I) to both sco1 and cox11 [2]. Sco1 is known to interact with cox2 [3], and is proposed to donate Cu(I) to the CuA center, whereas cox11 is a mitochondrial co-chaperone that provides copper to the CuB center [4]. Cox2 is an integral membrane protein anchored to the inner membrane of mitochondria through two transmembrane helices. Its soluble C-terminal part contains the copper binding site and projects out in the inner membrane space. Both sco1 and cox11 are anchored to the mitochondrial inner membrane by a single transmembrane helix with the soluble domain located in the inner membrane space, whereas cox17 is a non-membrane protein. See also Figure 1 for a schematic overview of the pathway.

Consistent with the importance of this copper insertion pathway, these proteins are associated with a number of human diseases [5-7]. Although 3D structures of the individual proteins are known, nothing is known about their interaction mode and full experimental characterization of the complexes is hampered among others by the instability of a number of components.

Docking is a computational methodology that uses the known structures of components of a complex to generate a model for the complex. Docking approaches are increasingly popular for investigating intermolecular complexes whose structures cannot be studied otherwise [8-10]. They have been applied for studying a variety of complexes, among which a complex related to the pathway we investigate here, cytochrome c – cytochrome c

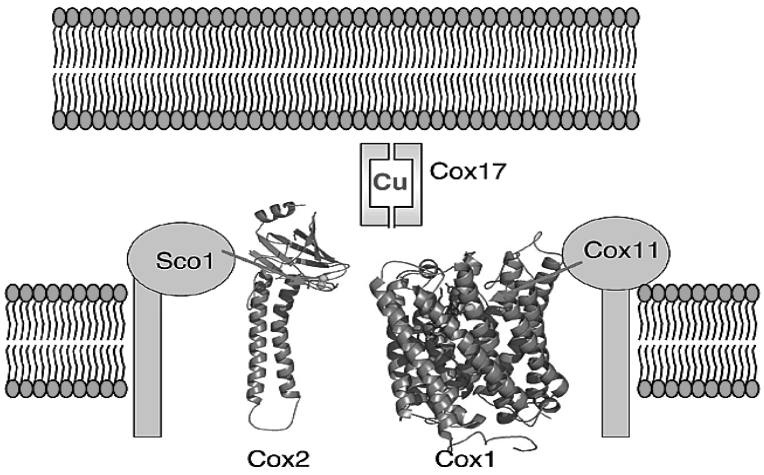


Figure 1. Scheme of Cu metalation of cox1 and cox2 mediated by cox11 and sco1, respectively. We model the interactions depicted in this figure: cox17-sco1, sco1-cox2 and cox17-cox11, and in addition the cox11 homodimer (see text for details); we do not model the cox11-cox1 interaction since Cu is thought to be inserted in cox1 co-translational, before folding of cox1.

oxidase [11], and some copper transfer complexes, $PacS_N$-ScAtx1 [12] and Atx1-ccc2 [13]. Here we apply the data-driven docking approach HADDOCK [14] to study the complexes along the cytochrome c oxidase Cu-delivery pathway for the human proteins. Since for the various proteins involved no experimental information is available for the complexes that are formed (at variance with the previously studied complexes mentioned above), we used bioinformatics methods to predict interface residues in order to generate restraints to drive the docking. Specifically, we model the interactions cox17 – sco1, sco1 – cox2, cox17 – cox11 and cox11 – cox11 (since cox11 is known to form a homodimer [15]). Note that we do not model the cox11 – cox1 complex, since the CuB site is buried deeply inside cox1, and it is thought that copper insertion occurs co-translational, i.e. before cox1 folding [16].

Results and discussion

To obtain insight into the cytochrome c oxidase Cu-delivery pathway we modeled the structures of the human cox17, cox11 and cox2 (for details see Methods); the resulting structures, together with the recently solved structure of human sco1 [17] are shown in Figure 2 and some characteristics are listed in Table 1.

Table 1. Models for the human cytochrome c oxidase Cu-delivery proteins

Protein	Model based on	% sequence identity	Ensemble –RMSD[a]
Cox17	1Z2G [18]	32 / 40[b]	5.8 (1.5) / 0.5 (0.1) [b]
Sco1	2GT5, 2GQM [17] [c]	- [c]	2.7 (0.5)
Cox11	1SP0 [15]	45	1.3 (0.4)
Cox2	1OCR [19]	73	0.2 (0.1)

[a] Average RMSD (standard deviation) of all 10 structures with respect to each other.
[b] Not taking into account unstructured N-terminal tail (residues 1-22).
[c] The NMR structure of the human sco1 which is available was used in this case; 2gt5 corresponds to the apo- and 2gqm to the holo-form.

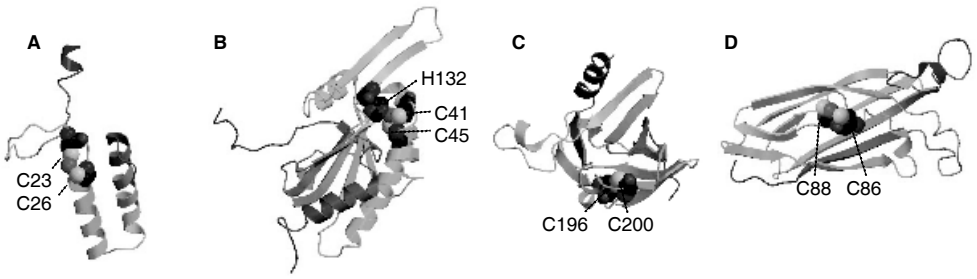


***Figure 2.** Ribbon representation of the homology model of cox17 **(A)**, the NMR structure of sco1 **(B)**, and the homology models of cox2 **(C)** and cox11 **(D)**. Proteins are color coded using a color ramp from blue (N-term) to red (C-term). The copper coordinating residues are shown in cpk and the copper atom is colored magenta (in cox17, sco1 and cox11; cox2 is shown without copper since only apo-cox2 is used in the docking). Homology modelling was performed using Modeller as described in Methods using templates listed in Table 1. Figure was generated using MolScript [20] and Raster3D [21].

These structures were used as input structures for modelling the following complexes: cox17(Cu)-sco1, sco1(Cu)-cox2, cox17(Cu)-cox11 and cox11(Cu)-cox11(Cu). Since there is no experimental information on the interface in those complexes, we used WHISCY [22] and PROMATE [23] interface predictions (see Methods) to generate ambiguous interaction restraints (AIRs) to drive the docking. The predictions are listed in Table 2 and are also shown mapped onto the respective structures in Figure 3.

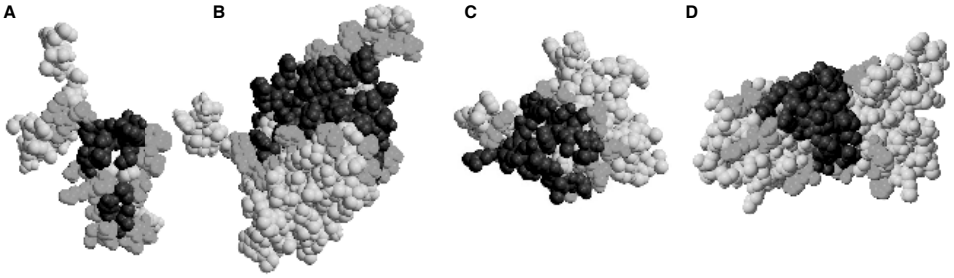


Figure 3. Spacefilling representation of the homology model of cox17 **(A)**, the NMR structure of sco1 **(B)**, and the homology models of cox2 **(C)** and cox11 **(D)**. Grayscale code indicates active (black) and passive (medium gray) residues used for docking (see Table 2 for corresponding residue numbers). Active residues are defined based on interface predictions using WHISCY [22], PROMATE [23] and their combination WHISCYMATE [22]; passive residues are surface neigbours of active residues (see Methods).

In all cases the predictions map to patches around the Cu-binding site, which is reasonable since these proteins interact in order to transfer copper (note that we did not restrict the prediction towards this site). In addition to the interface prediction based restraints we also used restraints between the Cu atoms and each of the coordinating cysteines (for details see Methods). We performed for each complex three different docking runs, using either the AIRs based on interface predictions, or the copper based restraints or both simultaneously. Since the latter gave the best convergence, we present results from the docking runs using the combined set of restraints.

To select the best model after docking, the structures were clustered based on their pairwise RMSDs. In Table 3 the characteristics of the resulting clusters are shown. In general, the structures with the lowest HADDOCK score (see Methods) of the cluster with the lowest HADDOCK score would constitute the predicted model. For the sco1-cox2 and cox17-cox11 this leads to the unambiguous selection of one cluster; for cox17-sco1 and cox11-cox11, however, there are two clusters which score almost equal (see further discussion below).

Table 2. List of active and passive residues used for docking.

Cox17	
Active	C23, C24, A25, C26, P27, R33, D34, C36, I37, C55, M56, A58, L59, G60, F61
Passive	P19, L20, K21, E28, K30, K31, A32, A35, E39, K40, G41, E42, E43, H44, G46, H47, E50, K53, E54, R57, K62
Cu	Cys23, Cys26
Sco1	
Active	G10, F38, T39, C41, P42, D43, C45, P46, L49, F72, I75, D76, Y88, F92, A110, R111, Y113, R114, Y116, Y117, S118, Y128, I129, V130, D131, T133, I134, Y137, G143, F145, Y148
Passive	L7, L8, S13, Y27, L28, G29, E47, E50, E78, T81, A84, N87, K90, E91, P94, R103, E104, D107, Q108, A112, P119, G120, P121, K122, D125, E126, D127, D142, E144, L146, N152
Cu	Cys41,Cys45
Cox11	
Active	Q38, Q39, L51, F53, K83, I84, Q85, C86, F87, C88, E90, V100, D101, M102, P103, V104, F105
Passive	G2, K20, N24, A25, D26, V27, N34, R36, T40, E41, Y43, E48, T49, A50, T69, N71, F75, Q79, Q92, R93, E98, E99, D122, L123, S127, T129, F131
Cu	Cys86, Cys88; Cu-Cu restraint
Cox2	
Active	Q103, T104, Y121, R134, D139, D158, V159, Q195, S197, I199, G201, A202, N203, S205
Passive	N119, P124, P125, L128, L133, N140, R141, Q157, L160, P166, K171, D173, P176, V191
Cu	Cys196, Cys200

For all complexes, ambiguous interaction restraints were defined using active and passive residues as listed in the Table, based on WHISCYMATE predictions. In addition, copper-based restraints were defined as described in Methods using the residues listed in the Table.

The docking provided thus structural models for these complexes along the cytochrome c oxidase Cu-delivery pathway, which had been uncharacterized so far. The models are shown in Figure 4. In the following we will discuss them and their biological implications. Table 4 data gives statistics of copper - cysteine contacts and Table 5 summarizes the intermolecular contacts.

Analysis of copper-transfer complexes

Cox17 – sco1

As noted above we find two clusters of solutions for the cox17-sco1 complex, with comparable HADDOCK score (although cluster 1 scores slightly better when assessed by FastContact, Dfire and Probe, see Methods). As shown in Figure 4 the orientation of the two molecules in these models is rather different. This is reflected in the difference between intermolecular contacts in the two clusters (see Table 5). One key difference is the role of cox17 Arg33 (which based on mutations is expected to be involved in complex formation

Table 3. Characteristics of resulting clusters[a]

Cl #[c]	N_{str}	Rmsd-Emin	Score	Dfire	BSA (Å^2)	Eair (kcal/mol)	dCu-Cys (Å) [b]
Cox17 (Cu) Sco1 (apo)							
1	**33**	**3.8 (2.0)**	**-140 (15)**	**-19.2 (1.9)**	**2289 (247)**	**408 (126)**	**2.6**
2	**56**	**13.6 (0.6)**	**-135 (10)**	**-16.1 (1.8)**	**2412 (285)**	**447 (71)**	**2.2**
3	27	9.3 (1.0)	-112 (12)	-16.5 (2.0)	2192 (254)	438 (112)	2.2
4	9	13.6 (0.4)	-106 (44)	-19.6 (2.3)	2354 (388)	520 (66)	2.2
5	24	10.7 (2.0)	-95 (17)	-14.7 (1.8)	1962 (238)	475 (88)	2.2
6	15	5.9 (1.4)	-93 (22)	-16.6 (3.1)	2031 (284)	543 (81)	2.3
7	8	9.2 (1.7)	-74 (7)	-15.6 (2.2)	1887 (220)	615 (65)	2.2
Sco1 (Cu) Cox2 (apo)							
1	**132**	**6.5 (3.3)**	**-106 (5)**	**-18.2 (1.3)**	**1943 (181)**	**446 (72)**	**2.5**
2	12	13.9 (0.6)	-86 (6)	-18.3 (1.0)	1938 (76)	601 (62)	16.5
3	14	14.0 (2.2)	-65 (5)	-14.7 (0.5)	1798 (194)	611 (72)	5.7
4	9	11.2 (0.7)	-45 (17)	-14.0 (1.3)	1691 (219)	644 (54)	4.0
5	8	16.5 (1.1)	-44 (8)	-14.4 (0.7)	1685 (237)	712 (61)	8.1
6	7	9.4 (1.2)	-34 (9)	-15.2 (1.4)	1858 (214)	782 (55)	11.4
Cox17 (Cu) Cox11 (apo)							
1	**125**	**4.5 (2.9)**	**-159 (8)**	-17.8 (0.6)	1835 (134)	333 (73)	**2.1**
2	15	6.1 (1.7)	-148 (6)	-17.6 (0.9)	1998 (158)	280 (76)	2.1
3	26	8.9 (1.4)	-137 (7)	-18.0 (0.6)	1885 (162)	285 (81)	2.1
4	12	14.4 (1.2)	-113 (9)	-15.0 (1.6)	1779 (129)	372 (64)	2.1
5	10	8.0 (1.9)	-112 (11)	-14.4 (1.0)	1828 (119)	260 (86)	2.1
Cox11 (Cu) Cox11 (Cu)							
1	**40**	**1.8 (1.2)**	**-117 (8)**	**-18.8 (1.0)**	**2291 (135)**	**137 (38)**	**-**
2	**22**	**8.2 (0.3)**	**-117 (6)**	**-19.5 (2.0)**	**2366 (114)**	**196 (43)**	**-**
3	114	16.9 (0.8)	-99 (6)	-16.4 (1.1)	1733 (98)	201 (40)	-
4	8	20.7 (0.5)	-62 (6)	-16.9 (1.2)	1563 (112)	329 (57)	-

[a]Characteristics of all resulting clusters with at least 5 structures for the different docking runs. Averages (standard deviation in brackets) were calculated over the best 5 structures of each cluster. N_{str}, number of structures in cluster (out of 200 water refined structures); Score, HADDOCK score, defined as a weighted sum of different terms (see Methods); Dfire, statistical energy function based on the reference state of distance-scaled, finite, ideal gases [24]; BSA, buried surface area. **Bold** designates clusters selected as 'best'; selection is based on HADDOCK score.

[b]Lowest distance in best 5 structures from cluster between Cu and Cys SG in Cu-accepting protein.

[c]Cluster number when sorted by HADDOCK score; as discussed in the text, for cox17 – sco1 and cox11 – cox11 two clusters were obtained that scored almost equally. These are referred to in the text as cluster 1 and cluster 2, respectively.

with sco1 [25]). In both clusters, it is involved in a salt bridge, the difference being that in cluster 1 it contacts sco1 Glu48 and in cluster 2 it contacts Asp123. Another marked difference is found with respect to cox17 Phe61. In cluster 1, Phe61 is involved in contacts with Arg114, Ile135 and Tyr137, whereas in cluster 2 it contacts Asp43, Val44 and Glu47. Clearly, these data provide testable hypotheses for mutation experiments in order to select between the two models.

The copper transfer seems to be more efficient in cluster 2 since 21 out of 56 structures in this cluster have the Cu ion within 3Å of both sco1 Cys41 and sco1 Cys45, whereas in cluster 1 this is the case for only 1 out of 33 structures. This might indicate that cluster 2 is a more

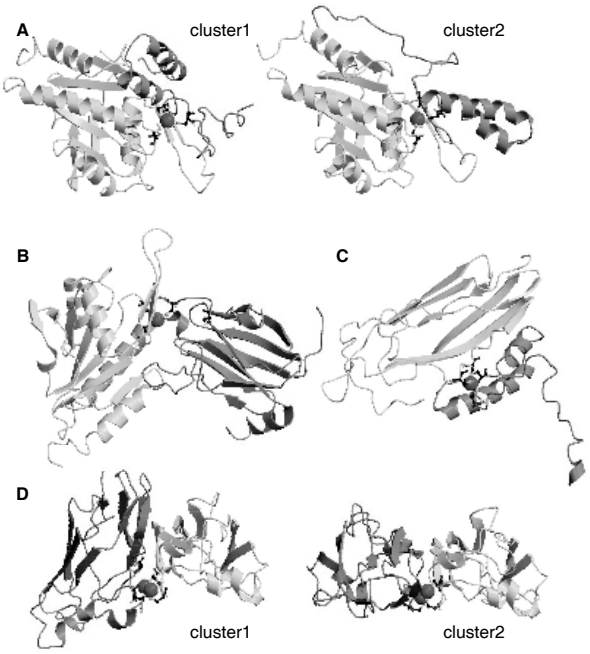


***Figure 4.** HADDOCK models of the complexes along the human cytochrome c oxidase Cu-delivery pathway: **(A)** cox17-sco1, **(B)** sco1-cox2, **(C)** cox17-cox11 and **(D)** cox11-cox11. Color coding: cox17 (A and C) in orange; cox2 (B) in orange; sco1 (A and B) and cox11 (C) in white. In (D), one of the cox11 monomers is orange and the other monomer is white. Note that for cox17-sco1 and cox11-cox11, as discussed in the text, there are two clusters of solutions with identical scores.

likely orientation. We analyzed our models to gain some insight into the copper transfer mechanism. In cluster 1 we find that the copper is preferably attaching to sco1 Cys41 (the average distance copper - Cys41 is lower than the distance copper – Cys45, see Table 4A); in cluster 2 there is no difference between Cys41 and Cys45 in this respect. When we redo the water-refinement of the original docking run without copper based restraints (data not shown) clustering results are similar but now cluster 1 clearly scores better than cluster 2 (note that this is consistent with the difference in score according to FastContact, Dfire and Probe, as mentioned above). Without these restraints the copper is not transferred to sco1 in most structures, but in those were it is, again Cys41 is the accepting cysteine. Note that in a docking run in which we did not use the Cu restraints at all (data not shown) we also find that copper is accepted by Cys41 (if transferred, which is much less often the case).

Sco1 – cox2

The CuA site in cox2 contains a binuclear Cu cluster and the transfer mechanism is unknown. In the present system we modelled the transfer of one copper ion, considering a mechanism of sequential insertion of the two copper ions (note that for modelling the transfer of the second copper the structure of cox2-Cu_1 should be known). The Cu ion in cox2 is buried below the surface of the protein and the coordinating cysteines have a low relative accessibility in the starting models (Cys196, 7 +/- 0.3% and Cys200, 4 +/- 1%, as calculated over the ensemble of starting structures using NACCESS [26]). However, in many docking solutions Cys200 becomes more exposed, as can be seen in Figure 5: the 5 best structures of the selected cluster the accessibility of Cys196 is still low, 2 +/-1% while that of Cys200 has now increased to 33 +/-19%. Based on our docking results, we predict that Cys200 is the cysteine that first binds the Cu ion upon interaction with sco1.

Table 4. Average interatomic distances between the sulfur atoms of copper binding cysteines and the copper ion calculated over the ensemble of the five best structures.

4A1. Cox17 (Cu) - Sco1 (apo) cluster1

Cox17	Sco1	Cu	Dist (Å)
C23 Sγ	C41 Sγ		5.1 (2.3)
C23 Sγ	C45 Sγ		8.7 (1.2)
C26 Sγ	C41 Sγ		4.1 (1.6)
C26 Sγ	C45 Sγ		6.9 (0.6)
C23 Sγ		Cu	2.3 (0.2)
C26 Sγ		Cu	2.2 (0.04)
	C41 Sγ	Cu	3.6 (1.9)
	C45 Sγ	Cu	7.0 (0.7)

4A2. Cox17 (Cu) - Sco1 (apo) cluster2

Cox17	Sco1	Cu	Dist (Å)
C23 Sγ	C41 Sγ		5.1 (1.7)
C23 Sγ	C45 Sγ		4.4 (1.5)
C26 Sγ	C41 Sγ		3.8 (1.2)
C26 Sγ	C45 Sγ		5.1 (0.9)
C23 Sγ		Cu	2.3 (0.2)
C26 Sγ		Cu	2.2 (0.04)
	C41 Sγ	Cu	3.3 (1.7)
	C45 Sγ	Cu	3.6 (1.7)

4B. Sco1 (Cu) – Cox2

Sco1	Cox2	Cu	Dist (Å)
C41 Sγ	C196 Sγ		13.6 (1.7)
C41 Sγ	C200 Sγ		6.2 (2.8)
C45 Sγ	C196 Sγ		14.6 (1.6)
C45 Sγ	C200 Sγ		7.8 (2.8)
C41 Sγ		Cu	2.2 (0.1)
C45 Sγ		Cu	2.3 (0.2)
	C196 Sγ	Cu	13.1 (2.1)
	C200 Sγ	Cu	6.1 (3.3)

4C. Cox17 (Cu) – Cox11

Cox17	Cox11	Cu	Dist (Å)
C23 Sγ	C86 Sγ		3.4 (0.2)
C23 Sγ	C88 Sγ		3.6 (0.2)
C26 Sγ	C86 Sγ		7.9 (1.0)
C26 Sγ	C88 Sγ		4.9 (0.7)
C23 Sγ		Cu	2.2 (0.03)
C26 Sγ		Cu	6.5 (0.4)
	C86 Sγ	Cu	2.2 (0.1)
	C88 Sγ	Cu	2.1 (0.04)

4D1. Cox11 – Cox11 (cluster1) [a]

Cox11	Cox11	Cu	Dist (Å)
C86 Sγ	C86 Sγ		3.2 (0.1)
C86 Sγ	C88 Sγ		5.2 (2.4)
C88 Sγ	C88 Sγ		4.0 (1.3)
C86 Sγ		Cu	2.1 (0.04)
C86 Sγ		Cu	2.2 (0.05)
C86 Sγ		Cu	4.1 (0.9)
C88 Sγ		Cu	2.0 (0.03)
C88 Sγ		Cu	2.1 (0.04)
C88 Sγ		Cu	3.0 (1.6)
		Cu-Cu	2.9 (0.03)

[a] Since cox11 – cox11 is a homodimer complex containing two copper atoms, for each copper atom four distances to cysteines can be measured. We list here the three lowest distances as well as the copper – copper distance; these can be directly compared to the EXAFS data (see Methods).

4D2. Cox11 – Cox11 (cluster2) [a]

Cox11	Cox11	Cu	Dist (Å)
C86 Sγ	C86 Sγ		3.2 (0.2)
C86 Sγ	C88 Sγ		3.6 (1.1)
C88 Sγ	C88 Sγ		3.2 (0.2)
C86 Sγ		Cu	2.1 (0.04)
C86 Sγ		Cu	2.2 (0.04)
C86 Sγ		Cu	2.9 (0.8)
C88 Sγ		Cu	2.1 (0.08)
C88 Sγ		Cu	2.2 (0.09)
C88 Sγ		Cu	2.3 (0.08)
		Cu-Cu	2.7 (0.1)

[a] See footnote for Table 4D1.

Table 5. Intermolecular Contact statistics[a]

5A1. Cox17 (Cu) - Sco1 (cluster1)

				M M	S S	M S	NB
Sco1		Cox17					
Lys	51	Asp	34	0	3	0	0
Cox17		Sco1					
Leu	20	Pro	121	0	0	0	3
Lys	21	Ile	129	0	0	0	3
Pro	22	Ser	118	0	0	0	3
Cys	23	Asp	131	0	2	0	5
Cys	24	Asp	131	0	0	3	0
Cys	26	His	132	0	0	1	3
Cys	26	Val	44	0	0	0	5
Pro	27	Pro	42	0	0	0	4
Lys	30	Glu	47	0	3	0	1
Lys	30	Glu	48	0	3	0	3
Lys	30	Val	44	0	0	2	4
Lys	31	Asp	43	0	4	0	3
Arg	33	Gln	151	0	0	4	0
Arg	33	Glu	48	0	4	0	2
His	52	Gln	151	0	3	0	3
Met	56	Gln	151	0	0	0	3
Met	56	Ile	134	0	0	0	3
Met	56	Ile	135	0	0	0	4
Leu	59	Thr	133	0	0	0	4
Leu	59	Tyr	116	0	0	0	4
Phe	61	Arg	114	0	0	0	3
Phe	61	Ile	135	0	0	0	5
Phe	61	Tyr	137	0	0	0	3

5A2. Cox17 (Cu) - Sco1 (cluster2)

				M M	S S	M S	NB
Sco1		Cox17					
Cys	41	Cys	26	0	3	0	0
Tyr	116	Glu	16	0	3	0	0
Cox17		Sco1					
Pro	19	Tyr	116	0	0	0	3
Lys	21	Asp	131	0	4	0	1
Lys	21	Thr	133	0	0	0	5
Lys	21	Tyr	116	0	0	1	3
Pro	22	Ile	134	0	0	0	4
Pro	22	Ile	135	0	0	0	3
Cys	23	His	132	0	1	0	3
Cys	23	Ile	134	0	0	0	4
Cys	26	His	132	0	1	0	4
Pro	27	Asp	131	0	0	0	3
Lys	30	Ile	129	0	0	2	5
Lys	31	Ser	118	0	4	0	2
Arg	33	Asp	123	0	4	0	3
Asp	34	Pro	121	0	0	0	3
Met	56	Pro	42	0	0	0	3
Leu	59	Gln	151	0	0	0	3
Phe	61	Asp	43	0	0	0	4
Phe	61	Glu	47	0	0	0	4
Phe	61	Val	44	0	0	0	3

[a]Contacts as defined using Dimplot (see Methods); numbers in table are number of structures for which the respective contact is present (out of 5 structures). MM, main chain – main chain hydrogen bonds; MS, main chain – side chain hydrogen bonds; SS, side chain – side chain hydrogen bonds and NB, non bonded contacts. For the hydrogen bonds, the first residue designates the donor and the second the acceptor.

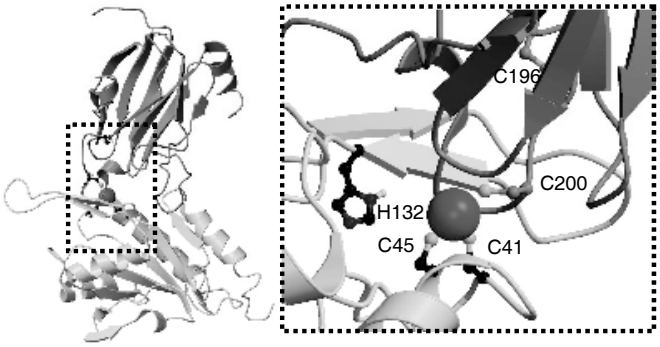


Figure 5. Detailed view of the sco1 – cox2 interface in one of the 5 best structures of the selected cluster where copper is (partially) transferred towards cox2. Note that His132 no longer coordinates the copper in this structure.

5B. Sco1 (Cu) – Cox2

				MM	SS	MS	NB
Cox2		Sco1					
Gln	103	Asp	43	0	0	3	0
Gln	103	Val	44	0	0	0	4
Trp	104	Val	44	0	0	0	4
Leu	135	Tyr	116	0	0	0	3
Ile	199	Ile	129	0	0	0	4
Ala	202	Cys	41	0	0	2	3
Ala	202	His	132	3	0	1	4
Ala	202	Phe	38	0	0	0	3
Asn	203	Asp	131	0	3	0	2
His	204	Asp	131	0	0	3	2
Ser	205	Asp	131	0	3	4	2
Phe	206	Asp	131	0	0	1	3
Phe	206	Thr	133	0	0	0	4
Phe	206	Tyr	116	0	0	0	3

5C. Cox17 (Cu) – Cox11

				MM	SS	MS	NB
Cox17		Cox11					
Lys	21	Phe	87	0	0	3	1
Cys	23	Cys	86	0	3	1	5
Cys	23	Cys	88	0	0	0	4
Cys	26	Cys	88	0	0	1	5
Cys	26	Met	102	0	0	0	3
Pro	27	Glu	90	0	0	0	4
Lys	30	Met	102	0	1	0	4
Lys	30	Pro	103	0	0	0	3
Lys	31	Glu	98	0	4	0	0
Lys	31	Glu	99	0	0	4	0
Lys	31	Val	100	0	0	0	3
Met	56	Phe	105	0	0	0	4
Met	56	Pro	103	0	0	0	3
Leu	59	Cys	86	0	0	0	3
Leu	59	Gln	85	0	0	1	3
Phe	61	Ile	84	0	0	0	4
Phe	61	Phe	105	0	0	0	5
Ile	63	Phe	105	0	0	0	4

5D1. Cox11 – Cox11 (cluster1)

				MM	SS	MS	NB
Cox11		Cox11					
His	3	Arg	93	0	2	0	3
Gln	38	Glu	98	0	4	0	0
Gln	39	Glu	98	0	4	0	0
Leu	51	Leu	94	0	0	0	3
Leu	51	Val	100	0	0	0	3
Phe	53	Glu	99	0	0	0	4
Phe	53	Val	100	0	0	0	4
Ile	84	Glu	90	0	0	0	5
Cys	86	Cys	86	0	1	2	5
Cys	86	Cys	88	0	5	1	3
Cys	86	Phe	87	0	0	3	4
Phe	87	Cys	88	0	0	3	4
Cys	88	Cys	88	0	2	0	5
Glu	90	Phe	105	0	0	4	5
Glu	90	Val	104	0	0	0	3
Gln	92	Phe	105	0	0	0	3
Val	100	Pro	103	0	0	0	4
Asp	101	Asp	101	0	0	0	3
Met	102	Met	102	0	0	0	4
Met	102	Pro	103	0	0	0	5

5D2. Cox11 – Cox11 (cluster2)

				MM	SS	MS	NB
Cox11		Cox11					
Gly	2	Val	100	0	0	0	5
His	3	Glu	98	0	0	0	3
His	3	Glu	99	0	0	1	3
His	3	Leu	94	0	0	0	3
Ser	5	Glu	98	0	4	5	4
Asp	6	Glu	98	0	0	5	1
Leu	51	Phe	53	0	0	0	5
Leu	51	Pro	103	0	0	0	5
Phe	53	Phe	53	0	0	0	3
Ile	84	Cys	88	0	0	0	4
Ile	84	Glu	90	0	0	0	3
Ile	84	Met	102	0	0	0	4
Gln	85	Gln	92	0	3	0	0
Gln	85	Glu	90	0	4	5	5
Cys	86	Cys	86	0	2	1	4
Cys	86	Cys	88	0	3	0	5
Cys	88	Cys	88	0	2	0	4
Asp	101	Phe	105	0	0	0	4
Met	102	Phe	105	0	0	0	5
Pro	103	Phe	105	0	0	0	4
Pro	103	Pro	103	0	0	0	5

A closer look at the position of the Cu ion in the docking results shows that the copper in some models is still close to the starting cysteine on sco1 and in others it is transferred to cox2 Cys200 (see Figure 5). In the latter case, sco1 His132 is no longer coordinating the copper ion. The distance copper travels can be assessed by fitting both the cox2 and the sco1 starting structure to the modeled complex and calculating the distance from the copper in the docking model to the copper in the initial sco1 structure and its location in the cox2 structure. For the best 5 structures of the cluster, the first distance is either 1-2Å, which is comparable to the precision of the cluster, or around 5Å, in the two cases were Cu is transferred to cox2 Cys200. In the latter case, the Cu ion is approximately halfway the path is has to travel from the sco1 to the cox2 site which is approximately 11Å.

Cox17 – cox11

It has been suggested previously that Glu90 and Glu91 in cox11, which are conserved in eukaryotes but not in prokaryotes, could be involved in complex formation with cox17 [15] (note that cox17 is not present in prokaryotes). Although the selected structures do not form intermolecular contacts involving Glu91 (in fact in the starting structure Glu91 is already forming two intramolecular contacts *via* its two OE atoms to two backbone amides), Glu90 indeed is involved in contacts with cox17; these are however not well defined within the cluster since contacts to various lysines of cox17 are observed (two times Lys21, once Lys17 and once Lys30).

With respect to the predicted path of the copper transfer we find that Cu first detaches from Cys26 of cox17. On the cox11 side, the Cu ion is in most cases bound to both Cys86 and Cys88 (107 out of 125 structures in the top ranking cluster have the Cu within 3Å of both cysteines), indicating that the transfer of copper in this case has almost fully occurred. This means that our model for the cox17-cox11 transfer complex represents a stage in the pathway further towards the final Cu-loaded accepting structure as compared to the cox17-sco1 complex. This obviously has to do with the fact that the two accepting Cys on cox11 are more easily accessible: the relative accessibilities of Cys86 and Cys88 in cox11 (calculated using NACCESS over the ensemble of starting structures) are 55 +/- 10% and 32 +/- 9%, respectively, whereas these are only 16 +/-7% and 3 +/- 3% for Cys41 and Cys45 in sco1.

Cox11 – cox11

In this case docking results in two clusters with comparable scores (both HADDOCK as well as other scores (data not shown)). The two clusters have a somewhat different orientation, but involve in both cases a hydrophobic interface consisting of residues Met102, Pro103 and Phe105. Pro103 is absolutely conserved in eukaryotes but less so in prokaryotes, which matches with the fact that only in eukaryotes cox11 is known to form homodimers. Both clusters reasonably fit the EXAFS data that were used as restraints, although cluster 2 shows a somewhat better agreement: cluster 2 indeed has the best fit to the EXAFS data also when compared with other clusters which are not selected based on their score (data not shown).

Interestingly, in both clusters the two cox11 molecules are oriented anti-parallel; the angle between the two long axes is between 20 – 40 degrees in cluster1 and between 10 – 15 degrees in cluster2. Taking into account that the N-terminal extension of cox11 is attached to a transmembrane helix, this orientation means that the two copper ions and their surrounding site face the membrane, as shown in Figure 6.

It is thought that copper is inserted into cox1 co-translationally [16]. Based on our model we speculate that copper could be brought into the membrane by virtue of the cox11 homodimer inserting into the membrane. Indeed, there are some hydrophobic residues (the absolutely conserved Phe87 and Phe89 as well as Phe81 and Phe106 which are absolutely conserved in eukaryotes as well) around the two Cu ions (see Figure 6) that might interact with the lipid

hydrocarbon tails. This would mean that dimerization of cox11 could help in delivering Cu to cox1.

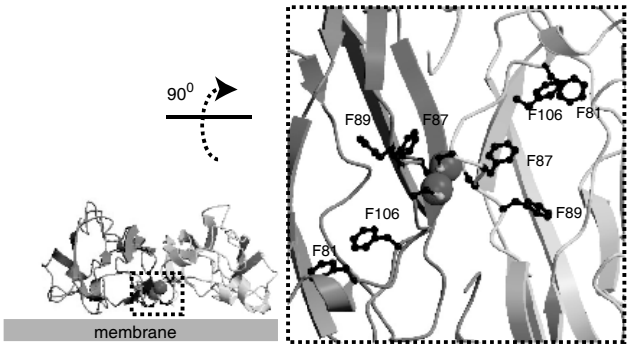


Figure 6. Resulting model for the cox11 – cox11 homodimer as seen from the membrane side. Note that, as discussed in the text, there are two clusters of solutions with related but somewhat different orientations; here we show the best scoring model from cluster 2. Based on our docking model we predict that the dimer contains a protein or membrane interaction site; possible interacting phenylalanines are shown in ball-and-stick (see Results for details).

Predicting interaction surfaces for the dimer (as performed before for the monomer structures; see Methods) indicates a putative binding site at the membrane facing site around the Cu (data not shown); this region might be involved in membrane interaction or alternatively might be a protein interaction site. The observation that cox11 from *S. Pombe* exists as a fusion protein with the partner protein being homologous to *S. cerevisiae* Rsm22 (which has been proposed to have a role in mitochondrial translation), suggests that cox11 function may occur during mitochondrial translation of the cox1 mRNA [16]. The interaction between cox11 and Rsm22 might however be indirect [27].

Computational aspects

The selection of correct solutions (scoring) is a difficult problem in docking. In the present case, the fact that we know that the Cu ion should be transferred imposes some additional restraints on the solutions, and reduces the search space. For the cox11 homodimer we could in addition use EXAFS data. We validated our docking approach by repeating the docking of the cox11 homodimer without the EXAFS-derived restraints. The clusters obtained in this way are similar to the ones obtained with inclusion of the EXAFS data (data not shown), which supports our docking approach.

It is interesting to analyze the influence of the starting structure used in the ensemble docking on the final result. For sco1-cox2 and cox17-sco1 we find that whether the Cu ion is transferred towards the accepting partner or not depends very much on the specific starting structure used for the docking. For the sco1-cox2 complex, the sco1 starting models have somewhat different orientations for the long loop 8, but there is no clear reason why some orientations would facilitate the transfer. What can be seen when transfer takes place is that Phe38, Val44, His132 and Ile134 have a somewhat different arrangement in the starting structures; it is however not clear why this arrangement should be favorable (note that His132 is involved in Cu coordination). We also find for the cox17-cox11 complex a large influence of the starting conformation. As mentioned above, in this complex the transfer is rather efficient. We observe that those models in which no transfer takes place are mostly derived from one specific starting conformation of cox11 (i.e. the accepting partner in this case); in this particular structure the two cysteines are somewhat further away from each other.

These observations all indicate that docking from ensembles of starting conformations is worthwhile doing since it increases the chance of obtaining relevant conformations even at the price of a larger interaction space to be searched. This is in line with previous studies [28-33].

Conclusions and perspectives

We modelled several complexes of the human proteins involved in copper delivery to cytochrome c oxidase: cox17 - sco1, sco1 - cox2, cox17 – cox11 and cox11 – cox11. Details of these interactions have not been obtained before, due to experimental difficulties in obtaining the structures of these complexes. Our models provide new information such as details on the copper transfer from the donating to the accepting protein in the case of cox17 - sco1, sco1 – cox2 and cox17 – cox11. The cox11 – cox11 homodimer suggests a possible membrane or protein interaction site, which might be involved in copper delivery to cox1. Finally, our models provide testable hypotheses for mutagenesis experiments, which are currently being carried out.

Methods

Structural coordinates

Models for the human proteins along the Cu-delivery pathway were generated using Modeller 6 [34] using the templates listed in Table 1. Out of 1000 generated models, the 10 best based on the objective score were used for docking. In the case of the cox17-sco1 complex we performed two docking runs, using either apo-sco1 or holo-sco1 (the latter without the Cu ion). Since the latter gave better convergence the results we show are from this docking run.

WHATIF [35] was used for determination of the protonation state of the histidines. Cu was introduced with charge +1 and for the appropriate cysteines (see Table 2), the SH atom was removed and the charge on atom SG was set to -0.68 instead of -0.45 to compensate for the introduction of Cu+1 (note that the charge of the removed SH atom is 0.27 and the charge of the CB atom is 0.18). In this way, the total net charge of two cysteines plus copper is zero. All non-polar hydrogens were removed as well.

Interface prediction

WHISCY [22] and ProMate [23] were used for interface predictions and combined for consensus scoring using WHISCYMATE [22]. Multiple sequence alignments were constructed by first finding homologues of the human proteins using blastp against the nr database using default settings (Word Size 3; Expect 10; Blosum62 matrix; Gap Costs: existence 11, extension 1) [36], except that a restriction to only eukaryote sequences was used since the copper-delivery pathway might be different in prokaryotes as compared to eukaryotes (for example, cox17 is only found in eukaryotes [37]). Subsequently, multiple sequence alignment was performed using ClustalW 1.83 [38]. ProMate predictions were obtained making use of the web interface of ProMate (http://bioportal.weizmann.ac.il/promate/), using default settings. For WHISCYMATE, a residue was predicted if its ProMate score was higher than or equal to 98.52 or its WHISCY score higher than or equal to 0.371 or if its ProMate and WHISCY scores were both higher than or equal to 55.42 and 0.107, respectively. Interface predictions were used to generate AIRs as discussed before [22]: predicted residues were designated active residues and their surface neigbours passive residues. We also used PPI-Pred [39] for interface prediction; PPI-Pred output consists of a most likely binding site location and two other possible binding site. In each case, the PPI-Pred predictions were overlapping with the WHISCYMATE predictions, although the overlap did not always occur with the most likely PPI-Pred predicted binding site.

Copper-based restraints
When appropriate, we defined (unambiguous) restraints between the Cu-ion and the S atoms of coordinating cysteines on both the delivering (holo) and accepting (apo) protein partner. The restraint distance was 2.4 +/- 0.1Å; residues involved in those restraints are shown in Table 2. For the cox11 homodimer, restraints were based on EXAFS data which indicate the presence of three S atoms at 2.23Å in the first coordination shell of the Cu, and a second Cu at 2.7Å [15]. In addition, we used C2-symmetry[40] and NCS-restraints for this homodimer.

Docking protocol
HADDOCK [14] uses ambiguous interaction restraints (AIRs) defined from any information about the interface. In our case, active residues are based on interface predictions, as mentioned above. The definition of ambiguous interaction restraints has been described in detail before [14]. The docking protocol consists of three consecutive stages [14]:

(i) randomization of orientations followed by rigid body energy minimization (EM);
(ii) semi-flexible simulated annealing in torsion angle space (TAD-SA), which consists of (ii-a) a rigid body Molecular Dynamics search and first simulated annealing, (ii-b) a second semi-flexible simulated annealing during which side chains at the interface are free to move, and (ii-c) a third semi-flexible simulated annealing during which both side chains and backbone at the interface are free to move; and
(iii) final refinement in Cartesian space with explicit solvent.

The docking was performed starting from models for the human proteins along the Cu-delivery pathway (see above). For each protein, 10 models were used, resulting in 100 different combinations. In the initial rigid body docking phase, 1000 structures were generated (10 for each combination) and the best 200 in terms of total intermolecular energy were further submitted to the semi-flexible simulated annealing and final water refinement. Flexible segments were defined as default in HADDOCK (using active and passive residues +/- 2 consecutive residues). In addition, for cox17 the first 22 disordered residues were treated as fully flexible.

Non-bonded intermolecular interactions were calculated with an 8.5Å cut-off using the OPLS parameters [41]. The dielectric constant epsilon was set to 10 in the vacuum part of the protocol and to 1 for the explicit solvent refinement. To account for the presence of false positives in the interface predictions, for each docking trial, 50 % of the restraints based on the interface predictions were randomly discarded. In order to keep the secondary structure elements intact during the simulated annealing refinement, hydrogen bond and dihedral angle restraints were introduced. We used DSSP [42] to identify secondary structure elements in the top ranking homology model: dihedral angle restraints were defined as the measured dihedral angle +/- a 20^0 range and for helices a distance restraint was defined between each i,i+4 O-N and O-HN pair (with upper/lower bound of 2.3Å/3.5Å and 1.7Å/2.5Å, respectively).

Scoring of the structures was performed using the HADDOCK score with standard weights [43]. In addition, we also calculated various scores with DFIRE (a statistical energy function based on the reference state of distance-scaled, finite, ideal gases [24]), FastContact (an estimate of the binding free energy based on a statistically determined desolvation contact potential and Coulomb electrostatics with a distance-dependent dielectric constant [44]) and Probe (which uses small-probe contact dots to assess the goodness-of-fit [45]); the latter was calculated using the MolProbity package [46]. The final selection of clusters was based on HADDOCK score.

The docking solutions were clustered based on positional RMSDs using a 7.5 Å cut-off; only clusters with at least 5 members were analyzed. The RMSDs were calculated on the interface backbone atoms of the smallest component after superposition on the interface backbone atoms of the largest component). This can be termed “ligand interface RMSD”.

Analysis of intermolecular contacts

Intermolecular contacts (hydrogen bonds and non-bonded contacts) were analyzed with DIMPLOT which is part of the LIGPLOT software [47] using the default settings (3.9 Å heavy-atoms distance cut-off for non-bonded contacts; 2.7 Å and 3.35 Å proton-acceptor and donor-acceptor distance cut-offs respectively with minimum 90^0 angles (D-H-A, H-A-AA, D-A-AA) for hydrogen bonds. A contact is defined to be present if it is found in at least 3 of the 5 best structures.

Acknowledgements

This work has been performed under the Project HPC-EUROPA (RII3-CT-2003-506079), with the support of the European Community – Research Infrastructure Action under the FP6 "Structuring the European Research Area" Programme and was also supported by a "Jonge Chemici" grant (grant number 700.50.512) from the Netherlands Organization for Scientific Research (N.W.O.) to AB.

References

1. Carr HS, Winge DR. Assembly of cytochrome c oxidase within the mitochondrion. Accounts Chem Res (2003), **36**, 309-316.
2. Cobine PA, Pierrel F, Winge DR. Copper trafficking to the mitochondrion and assembly of copper metalloenzymes. Biochimica et Biophysica Acta, in press (2006).
3. Lode A, Kuschel M *et al.* Mitochondrial copper metabolism in yeast: interaction between Sco1p and Cox2p. Febs Lett (2000), **485**, 19-24.
4. Hiser L, Di Valentin M *et al.* Cox11p is required for stable formation of the Cu-B and magnesium centers of cytochrome c oxidase. J Biol Chem (2000), **275**, 619-623.
5. Papadopoulou LC, Sue CM *et al.* Fatal infantile cardioencephalomyopathy with COX deficiency and mutations in SCO2, a COX assembly gene. Nat Genet (1999), **23**, 333-337.
6. Sue CM, Karadimas C *et al.* Differential features of patients with mutations in two COX assembly genes, SURF-1 and SCO2. Ann Neurol (2000), **47**, 589-595.
7. Valnot I, Osmond S *et al.* Mutations of the SCO1 gene in mitochondrial cytochrome c oxidase deficiency with neonatal-onset hepatic failure and encephalopathy. Am J Hum Genet (2000), **67**, 1104-1109.
8. van Dijk ADJ, Boelens R, Bonvin AMJJ. Data-driven docking for the study of biomolecular complexes. Febs J (2005), **272**, 293-312.
9. Bonvin AM. Flexible protein-protein docking. Curr Opin Struc Biol (2006), **16**, 194-200.
10. Gray JJ. High-resolution protein-protein docking. Curr Opin Struc Biol (2006), **16**, 183-193.
11. Bertini I, Cavallaro G, Rosato A. A structural model for the adduct between cytochrome c and cytochrome c oxidase. J Biol Inorg Chem (2005), **10**, 613-624.
12. Banci L, Bertini I *et al.* The delivery of copper for thylakoid import observed by NMR. PNAS (2006), **103**, 8320-8325.
13. Arnesano F, Banci L *et al.* A docking approach to the study of copper trafficking proteins: Interaction between metallochaperones and soluble domains of copper ATPases. Structure (2004), **12**, 669-676.
14. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc (2003), **125**, 1731-1737.
15. Banci L, Bertini I *et al.* Solution structure of Cox11, a novel type of beta-immunoglobulin-like fold involved in Cu-B site formation of cytochrome C oxidase. J Biol Chem (2004), **279**, 34833-34839.
16. Carr HS, Maxfield AB *et al.* Functional analysis of the domains in Cox11. J Biol Chem (2005), **280**, 22664-22669.
17. Banci L, Bertini I *et al.* A hint for the function of human Sco1 from different structures. PNAS (2006), **103**, 8595-8600.
18. Arnesano F, Balatri E *et al.* Folding studies of Cox17 reveal an important interplay of cysteine oxidation and copper binding. Structure (2005), **13**, 713-722.
19. Yoshikawa S, Shinzawa-Itoh K *et al.* Redox-coupled crystal structural changes in bovine heart cytochrome c oxidase. Science (1998), **280**, 1723-1729.
20. Kraulis PJ. MOLSCRIPT: A program to produce both detailled and schematic plots of protein structures. J Appl Cryst (1991), **24**, 946-950.
21. Merrit EA, Murphy MEP. Raster3D version 2.0: A program for photorealistic molecular graphics. Acta Cryst (1994), **D50**, 869-873.

22. de Vries SJ, van Dijk ADJ, Bonvin AMJJ. WHISCY: What information does surface conservation yield? Application to data-driven docking. Proteins (2006), **63**, 479-489.
23. Neuvirth H, Raz R, Schreiber G. ProMate: A structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol (2004), **338**, 181-199.
24. Zhou HY, Zhou YQ. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci (2002), **11**, 2714-2726.
25. Punter FA, Glerum DM. Mutagenesis reveals a specific role for Cox17p in copper transport to cytochrome oxidase. J Biol Chem (2003), **278**, 30875-30880.
26. Hubbard SJ, Thornton JM. NACCESS. London: Department of Biochemistry and Molecular Biology, University College; 1993.
27. Khalimonchuk O, Ostermann K, Rodel G. Evidence for the association of yeast mitochondrial ribosomes with Cox11p, a protein required for the Cu-B site formation of cytochrome c oxidase. Curr Genet (2005), **47**, 223-233.
28. Smith GR, Sternberg MJE, Bates PA. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. J Mol Biol (2005), **347**, 1077-1101.
29. Grunberg R, Leckner J, Nilges M. Complementarity of structure ensembles in protein-protein binding. Structure (2004), **12**, 2125-2136.
30. Dominguez C, Bonvin AMJJ *et al.* Structural model of the UbcH5B/CNOT4 complex revealed by combining NMR, mutagenesis, and docking approaches. Structure (2004), **12**, 633-644.
31. van Dijk ADJ, de Vries SJ *et al.* Data-driven docking: HADDOCK's adventures in CAPRI. Proteins (2005), **60**, 232-238.
32. Smith GR, Fitzjohn PW *et al.* Incorporation of flexibility into rigid-body docking: Applications in rounds 3-5 of CAPRI. Proteins (2005), **60**, 263-268.
33. Camacho CJ. Modeling side-chains using molecular dynamics improve recognition of binding region in CAPRI targets. Proteins (2005), **60**, 245-251.
34. Sali A, Blundell TL. Comparative Protein Modeling by Satisfaction of Spatial Restraints. J Mol Biol (1993), **234**, 779-815.
35. Vriend G. What If - a Molecular Modeling and Drug Design Program. J Mol Graphics (1990), **8**, 52-&.
36. Altschul SF, Madden TL *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res (1997), **25**, 3389-3402.
37. Banci L, Bertini I *et al.* A copper(I) protein possibly involved in the assembly of Cu-A center of bacterial cytochrome c oxidase. PNAS (2005), **102**, 3994-3999.
38. Thompson JD, Higgins DG, Gibson TJ. Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Res (1994), **22**, 4673-4680.
39. Bradford JR, Westhead DR. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics (2005), **21**, 1487-1494.
40. Nilges M. A Calculation Strategy for the Structure Determination of Symmetrical Dimers by H-1-Nmr. Proteins-Structure Function and Genetics (1993), **17**, 297-309.
41. Jorgensen WL, Tirado-rives J. The OPLS Potential functions for proteins. Energy minimizations for crystals of cyclin peptides and crambin. J Am Chem Soc (1988), **110**, 1657-1666.
42. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers (1983), **22**, 2577-2637.
43. van Dijk ADJ, Kaptein R *et al.* Combining NMR relaxation with chemical shift perturbation data to drive protein-protein docking. J Biomol Nmr (2006), **34**, 237-244.
44. Camacho CJ, Zhang C. FastContact: rapid estimate of contact and binding free energies. Bioinformatics (2005), **21**, 2534-2536.
45. Word JM, Lovell SC *et al.* Visualizing and quantifying molecular goodness-of-fit: Small-probe contact dots with explicit hydrogen atoms. J Mol Biol (1999), **285**, 1711-1733.
46. Davis IW, Murray LW *et al.* MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res (2004), **32**, W615-W619.
47. Wallace AC, Laskowski RA, Thornton JM. Ligplot - a Program to Generate Schematic Diagrams of Protein Ligand Interactions. Protein Eng (1995), **8**, 127-134.

Appendix

Supporting Material Chapter 3

Supporting Table 1. The results of ROTDIF analysis of ^{15}N relaxation data for each Ub_2 structure, fitting data for both domains simultaneously and for each domain separately. The meaning of all parameters is the same as in Table 4 (text). Also shown, for comparison (bottom row), are the Euler angles for an average orientation (over each ensemble) of the rotational diffusion tensor predicted from hydrodynamic calculations using HYDRONMR (de la Torre et al., *J. Mag. Reson.* **2000**, *147*, 138-146), which uses the "shell modeling" strategy representing hydration effects by a shell covering the surface of the protein. These calculations were performed for each Ub_2 structure; a shell thickness of 3.5 Å was used, representing the sum of the thickness of the hydration shell and the average atomic vanderWaals radius.

CSP

	D+P					Dist					Prox					
Str #	α	β	γ	χ^2	χ^2/df	α	β	γ	χ^2	χ^2/df	α	β	γ	χ^2	χ^2/df	$\chi^{2(}$D+P)/ [χ^2(D)+χ^2(P)]
1	-19	90	-3	211.4	3.1	-18	94	-30	107.3	3.4	-12	94	5	72.7	2.3	1.17
2	-12	91	7	413.6	6.0	-25	108	-47	136.1	4.3	10	88	8	105.8	3.4	1.71
3	-16	91	3	385.2	5.6	-24	104	-67	146.1	4.6	6	89	4	122.4	3.9	1.44
4	-16	92	2	349.0	5.1	-23	104	-60	142.4	4.5	-1	91	8	105.7	3.4	1.41
5	-14	93	4	356.9	5.2	-19	107	-40	177.9	5.6	-7	91	7	106.5	3.4	1.25
6	-16	93	-2	358.6	5.2	-25	113	-85	167.3	5.2	-8	89	4	72.4	2.3	1.50
7	-17	93	5	295.9	4.3	-23	101	3	144.2	4.5	1	92	10	91.7	3.0	1.25
8	-19	88	1	302.0	4.4	-24	100	-53	148.0	4.6	-12	88	9	81.1	2.6	1.32
9	-20	85	1	265.8	3.9	-26	95	-25	131.3	4.1	-11	87	15	62.5	2.0	1.37
10	-13	86	8	362.4	5.3	-21	101	-85	159.0	5.0	-2	87	7	81.5	2.6	1.51
mean	-16	90	3	330.1	4.8	-23	103	-49	145.9	4.6	-4	90	8	90.2	2.9	1.39
std	3	3	4	60.7	0.9	3	6	27	19.6	0.6	8	2	3	19.3	0.6	0.16
Hydro NMR	-3 (1)	90 (2)	-2 (8)													

VEAN

	D+P					Dist					Prox					
Str #	α	β	γ	χ^2	χ^2/df	α	β	γ	χ^2	χ^2/df	α	β	γ	χ^2	χ^2/df	$\chi^{2(}$D+P)/ [χ^2(D)+χ^2(P)]
1	13	84	188	238.3	3.5	11	80	184	125.6	3.9	10	84	190	77.4	2.5	1.17
2	14	84	185	321.4	4.7	14	76	170	119.6	3.7	11	85	187	126.8	4.1	1.30
3	12	83	178	401.8	5.8	14	73	103	167.0	5.2	9	88	173	160.3	5.2	1.23
4	8	83	188	286.5	4.2	3	72	194	138.0	4.3	17	87	189	56.7	1.8	1.47
5	20	86	156	390.7	5.7	17	82	165	144.6	4.5	19	92	159	187.8	6.1	1.18
6	14	85	188	323.5	4.7	16	76	140	119.5	3.7	6	84	185	115.8	3.7	1.37
7	15	85	188	346.1	5.0	17	80	183	132.6	4.1	-2	84	185	143.3	4.6	1.25
8	21	86	177	319.3	4.6	15	78	181	95.7	3.0	28	96	171	148.9	4.8	1.31
9	17	82	186	343.3	5.0	18	78	150	112.8	3.5	12	84	181	155.1	5.0	1.28
10	16	87	184	329.4	4.8	12	77	129	186.3	5.8	17	84	190	99.0	3.2	1.15
mean	15	85	182	330.0	4.8	14	77	160	134.2	4.2	13	87	181	127.1	4.1	1.27
std	4	2	10	46.9	0.7	4	3	29	26.6	0.8	8	4	10	40.3	1.3	0.10
Hydro NMR	1 (1)	90 (2)	183 (9)													

Appendix

SANI

	D+P					Dist					Prox					
Str #	α	β	γ	χ^2	χ^2/df	α	β	γ	χ^2	χ^2/df	α	β	γ	χ^2	χ^2/df	$\chi^{2(D+P)}/[\chi^2(D)+\chi^2(P)]$
1	19	89	175	255.0	3.7	19	78	97	126.1	3.9	17	87	180	64.9	2.1	1.33
2	15	88	176	194.7	2.8	14	81	174	114.3	3.6	12	90	177	51.4	1.7	1.17
3	14	82	181	248.4	3.6	15	74	181	110.7	3.5	11	84	176	68.6	2.2	1.39
4	17	89	177	218.2	3.2	16	80	173	110.2	3.4	12	90	182	61.6	2.0	1.27
5	18	83	187	289.1	4.2	20	74	123	134.4	4.2	13	82	180	64.4	2.1	1.45
6	22	91	170	219.9	3.2	21	84	146	106.2	3.3	18	90	185	62.6	2.0	1.30
7	19	86	181	223.5	3.2	18	77	177	110.4	3.5	18	89	178	59.5	1.9	1.32
8	20	90	181	227.6	3.3	19	82	181	113.8	3.6	14	91	176	66.5	2.1	1.26
9	20	86	176	202.8	2.9	19	78	171	115.8	3.6	16	88	177	50.7	1.6	1.22
10	17	84	174	243.4	3.5	17	77	161	112.9	3.5	13	86	173	61.4	2.0	1.40
mean	18	87	178	232.2	3.4	18	79	158	115.5	3.6	14	88	178	61.2	2.0	1.31
std	2	3	5	27.6	0.4	2	3	28	8.5	0.3	3	3	3	6.0	0.2	0.08
Hydro NMR	1 (1)	90 (2)	173 (9)													

VEAN-SANI

	D+P					Dist					Prox					
Str #	α	β	γ	χ^2	χ^2/df	α	β	γ	χ^2	χ^2/df	α	β	γ	χ^2	χ^2/df	$\chi^{2(D+P)}/[\chi^2(D)+\chi^2(P)]$
1	97	88	121	237.8	3.4	104	85	112	99.2	3.1	93	85	118	82.0	2.6	1.31
2	98	92	120	217.3	3.2	101	90	114	104.9	3.3	96	87	119	79.7	2.6	1.18
3	98	86	120	237.8	3.4	105	81	93	104.4	3.3	95	85	121	85.5	2.8	1.25
4	98	92	123	262.7	3.8	102	91	107	130.2	4.1	94	87	118	98.1	3.2	1.15
5	96	91	117	226.5	3.3	103	88	110	111.0	3.5	96	88	116	69.4	2.2	1.26
6	97	91	121	230.2	3.3	104	86	107	115.7	3.6	91	92	116	71.5	2.3	1.23
7	99	89	114	221.7	3.2	103	85	97	104.1	3.3	96	86	121	77.3	2.5	1.22
8	94	91	122	217.7	3.2	99	90	109	104.3	3.3	92	85	123	72.1	2.3	1.23
9	100	92	123	225.5	3.3	104	90	114	111.4	3.5	95	88	126	98.2	3.2	1.08
10	96	90	121	225.0	3.3	100	87	124	105.7	3.3	94	86	118	78.8	2.5	1.22
mean	97	90	120	230.2	3.3	103	87	109	109.1	3.4	94	87	120	81.3	2.6	1.21
std	2	2	3	13.4	0.2	2	3	9	8.8	0.3	2	2	3	10.2	0.3	0.07
Hydro NMR	87 (1)	80 (1)	111 (5)													

Crystal

	D+P					Dist					Prox					
	α	β	γ	χ^2	χ^2/df	α	β	γ	χ^2	χ^2/df	α	β	γ	χ^2	χ^2/df	$\chi^{2(D+P)}/[\chi^2(D)+\chi^2(P)]$
	101	87	163	329.2	4.8	113	79	130	105.4	3.3	85	84	173	88.7	2.9	1.70
Hydro NMR	90	82	147													

Supporting Table 2. The intervening angles (in degrees) between the corresponding axes of the overall rotational diffusion tensor of Ub_2 derived from fitting both domains simultaneously ("Ub_2") and from the structures of each individual Ub domain ("Dist" or "Prox").

CSP

str	Ub_2 *vs.* Dist			Ub_2 *vs.* Prox			Dist *vs.* Prox		
#	x-x	y-y	z-z	x-x	y-y	z-z	x-x	y-y	z-z
1	27	27	4	9	10	8	35	35	6
2	55	51	21	1	22	22	56	54	40
3	70	68	15	1	22	22	72	67	33
4	62	61	14	6	16	15	68	65	25
5	45	43	15	3	8	7	49	45	20
6	83	80	22	7	10	9	90	82	29
7	8	7	10	4	18	18	7	25	25
8	55	53	13	8	11	7	63	61	17
9	28	26	12	15	17	9	40	42	17
10	94	91	17	2	11	11	93	89	24
mean	52.7	50.7	14.3	5.6	14.5	12.8	57.3	56.5	23.6
std	26.6	25.9	5.2	4.4	5.2	6.0	26.0	20.3	9.4

VEAN

str	Ub_2 *vs.* Dist			Ub_2 *vs.* Prox			Dist *vs.* Prox		
#	x-x	y-y	z-z	x-x	y-y	z-z	x-x	y-y	z-z
1	6	5	4	2	3	3	7	6	4
2	17	15	8	2	4	3	19	17	9
3	75	74	10	7	6	6	70	69	16
4	13	6	12	6	8	10	18	11	20
5	9	10	5	7	3	6	11	8	10
6	48	48	9	4	9	8	44	44	13
7	7	5	5	5	17	17	3	19	19
8	8	7	10	11	10	12	19	17	22
9	36	36	4	6	8	5	31	30	8
10	56	56	11	7	6	3	62	62	9
mean	27.5	26.2	7.8	5.7	7.4	7.3	28.4	28.3	13
std	24.7	25.4	3.0	2.7	4.2	4.6	23.1	22.6	6.0

SANI

str	Ub_2 *vs.* Dist			Ub_2 *vs.* Prox			Dist *vs.* Prox		
#	x-x	y-y	z-z	x-x	y-y	z-z	x-x	y-y	z-z
1	78	78	11	5	5	3	83	83	9
2	7	3	7	2	3	4	10	3	9
3	8	1	8	6	6	4	12	7	11
4	10	4	9	5	7	5	13	9	11
5	64	63	9	8	9	5	56	55	11
6	25	24	7	15	16	4	39	39	7
7	10	4	9	4	3	3	12	1	12
8	8	1	8	5	8	6	11	7	10
9	9	5	8	2	4	4	12	6	10
10	15	13	7	3	4	4	15	12	10
mean	23	20	8	6	7	4	26	22	10
std	25.9	28.0	1.3	3.7	3.8	1.0	25.3	27.6	1.4

VEAN-SANI

str	Ub_2 *vs.* Dist			Ub_2 *vs.* Prox			Dist *vs.* Prox		
#	x-x	y-y	z-z	x-x	y-y	z-z	x-x	y-y	z-z
1	11	9	8	4	6	5	11	7	11
2	7	6	4	1	5	5	6	7	6
3	27	26	9	2	2	3	29	27	11
4	17	16	4	5	8	6	12	13	9
5	10	7	8	2	3	3	9	6	7
6	16	14	9	8	5	6	17	9	14
7	17	17	6	7	8	4	24	23	7
8	14	13	5	2	6	6	14	16	9
9	10	9	4	4	7	6	14	13	9
10	6	3	5	3	5	4	8	7	6
mean	13.5	12	6.2	3.8	5.5	4.8	14.4	12.8	8.9
std	6.2	6.7	2.1	2.3	2.0	1.2	7.2	7.3	2.6

Crystal

	Ub_2 *vs.* Dist			Ub_2 *vs.* Prox			Dist *vs.* Prox		
	x-x	y-y	z-z	x-x	y-y	z-z	x-x	y-y	z-z
	34	32	14	9	18	16	42	44	28

Supporting Table 3. Analysis of intermolecular contacts and hydrogen bonds for the solution and the crystal (PDB entry 1AAR) structures of Ub_2.[a]

Donor → acceptor	Hydrogen bonds			Non-bonded contacts
	M-M	S-S	M-S	
Common contacts for the crystal and solution structures of Ub_2				
Ile44P – Val70D [M]	-	-	-	10
Val70P – Ile44D [M]	-	-	-	10
Ala46P – Thr9D	-	-	-	1
Gly47P – Thr9D	-	-	-	2
Tyr59P -- Gly76D	-	-	-	5
His68P – Leu8D	-	-	-	1
Arg54P → Gly75D	-	-	1	-
Arg72D → Glu51P	-	3	-	-
Gln49P → Leu71D [M]	8	-	-	-
Gln49D → Leu71P [M]	4	-	-	-
Leu71P → Gly47D [M]	5	-	-	-
Leu71D → Gly47P [M]	7	-	-	-
Contacts present only in the solution structure of Ub_2				
Leu8P – Ile44D [M]	-	-	-	5
Ile44P – Leu8D [M]	-	-	-	4
Leu8P – Gly47D [M]	-	-	-	5
Gly47P – Leu8D [M]	-	-	-	9
Lys48P – Leu73D [M]	-	-	-	4
Leu73P – Lys48D [M]	-	-	-	4
Lys48P – Val70D [M]	-	-	-	4
Val70P – Lys48D [M]	-	-	-	4
Gln49P – Val70D [M]	-	-	-	6
Val70P – Gln49D [M]	-	-	-	5
Gln49P – Arg72D [M]	-	-	-	10
Arg72P – Gln49D [M]	-	-	-	6
Glu51P – Arg72D [M]	-	-	-	5
Arg72P – Glu51D [M]	-	-	-	4
Glu51P – Arg74D [M]	-	-	-	5
Arg74P – Glu51D [M]	-	-	-	7
Gly75D → Glu51P	-	-	4	-
Tyr59P → Gly76D	-	-	4	-
Gln49P → Arg72D	-	4	-	-
Glu51P → Arg72D [M]	-	4	-	-
Arg72P → Glu51D [M]	-	9	-	-
Glu51P → Arg74D [M]	-	5	-	-
Arg74P → Glu51D [M]	-	9	-	-

[a] Intermolecular contacts were analyzed with Dimplot [43] and are reported if present in at least 4 of the ten best structures (for the common contacts, all occurrences are reported). The occurrence of main-chain-main-chain (M-M), side-chain-side-chain (S-S) and main-chain-side-chain (M-S) hydrogen bonds is reported. D and P correspond to distal and proximal domain, respectively. The number of occurrences of a given interaction over the ensemble of ten best structures is reported for the solution structure. [M] indicates a contact which is mirrored, being present both in the distal-proximal and proximal-distal form.

Supporting Material Chapter 5

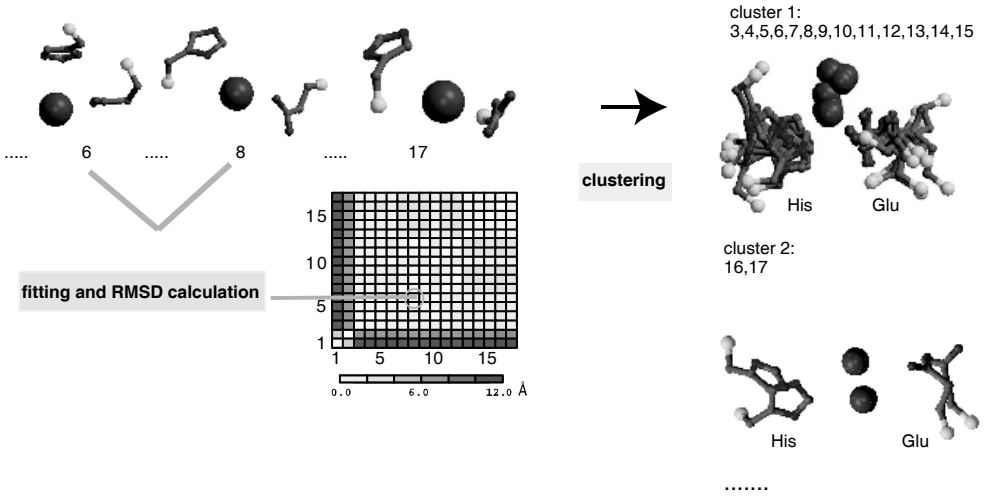


Figure 5. Clustering of water-mediated contacts.
Each type of water-mediated contact is clustered separately (see Methods); here the Glu – His contact is used as an example. Clustering of all water-mediated Glu – His contacts is performed on the basis of positional RMSD, calculated after least-square positional fitting on the coordinates of the water oxygen, its contacting heavy atoms within 5Å on both chains and their respective first bonded partner (total of five atoms). Since several atoms of a given side-chain can make contacts with the water oxygen atom within 5Å, various combinations of atoms were tested for the calculation of the RMSD matrix and the one resulting in the best clustering (most populated first cluster) was selected. In the example shown in the figure, the fitting is performed using the water oxygen, His NE, His CD, Glu OE1 and Glu CD. This happens to be the combination of atoms that gives the best clustering (with 13 out of 17 instances in the first cluster).

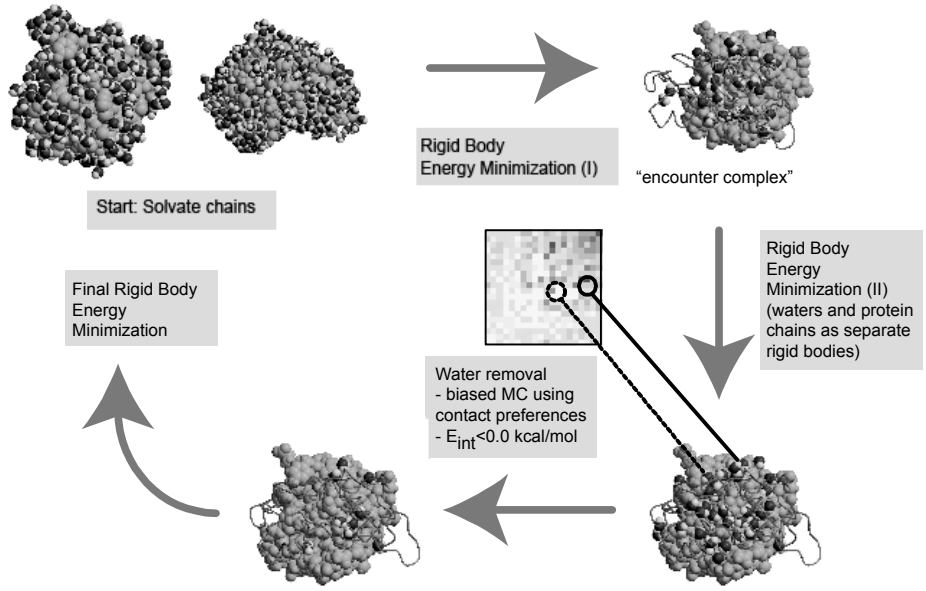


Figure 6. Overview of our solvated docking protocol.

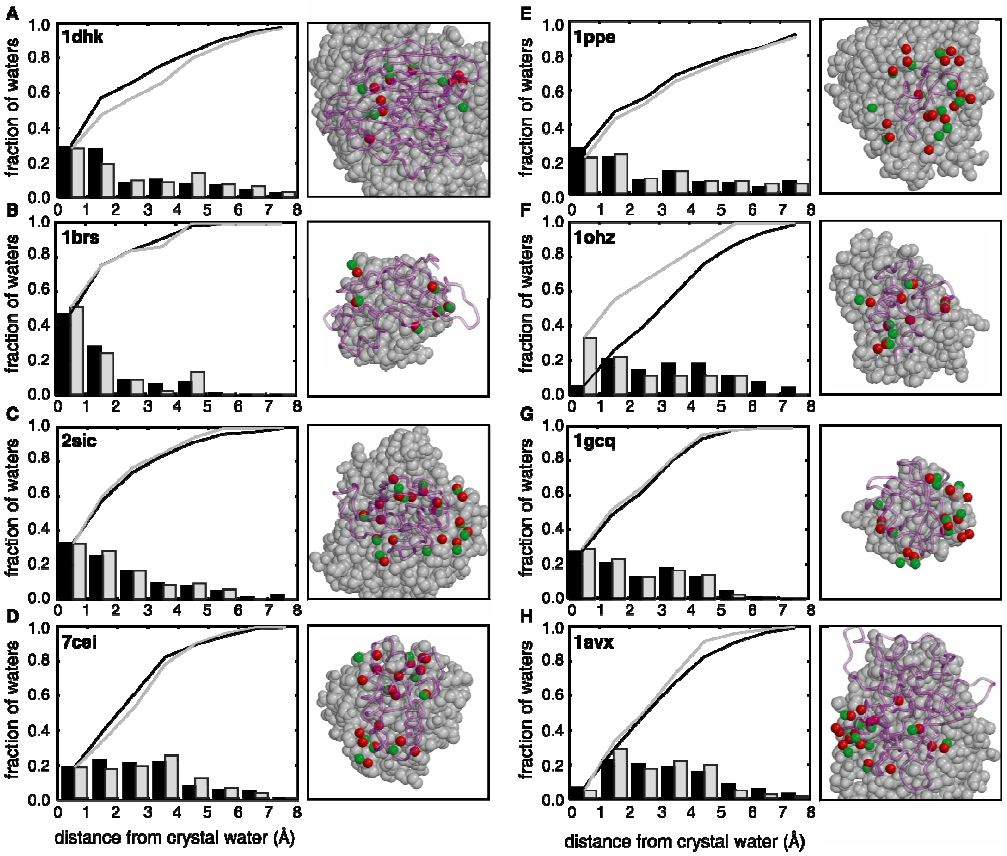


Figure 7. Accuracy of predicted water molecules in bound rigid body solvated docking. For each of the studied complexes (A-H) the cumulative fraction of the closest distances between modelled waters and crystal waters for all acceptable structures (dark grey) and for the top 10 acceptable structures (light grey) out of the top 200 ranked structures is shown in the left panel. The best-scoring acceptable solvated docking solution is shown in the right panel, together with its predicted waters (red) and the corresponding crystal waters (green).

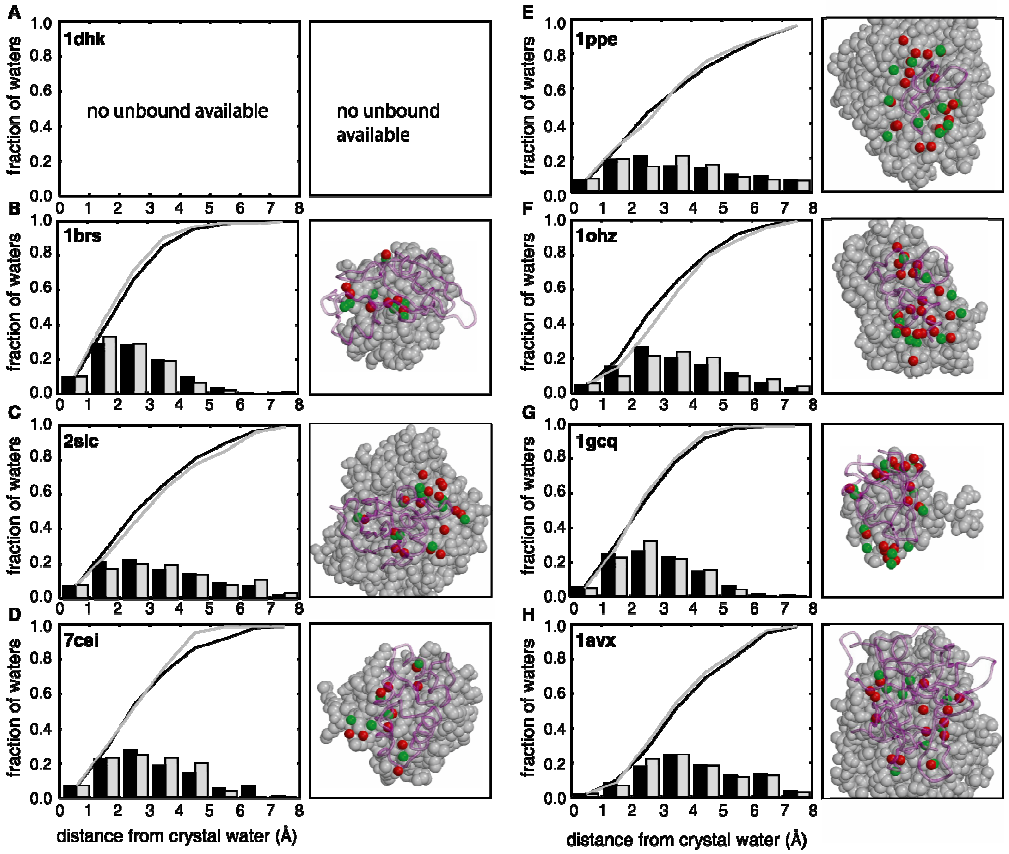


Figure 8. Accuracy of predicted water molecules in unbound rigid body solvated docking. For each of the studied complexes (A-H) the cumulative fraction of the closest distances between modelled waters and crystal waters for all acceptable structures (dark grey) and for the top 10 acceptable structures (light grey) out of the top 200 ranked structures is shown in the left panel. The best-scoring acceptable solvated docking solution is shown in the right panel, together with its predicted waters (red) and the corresponding crystal waters (green).

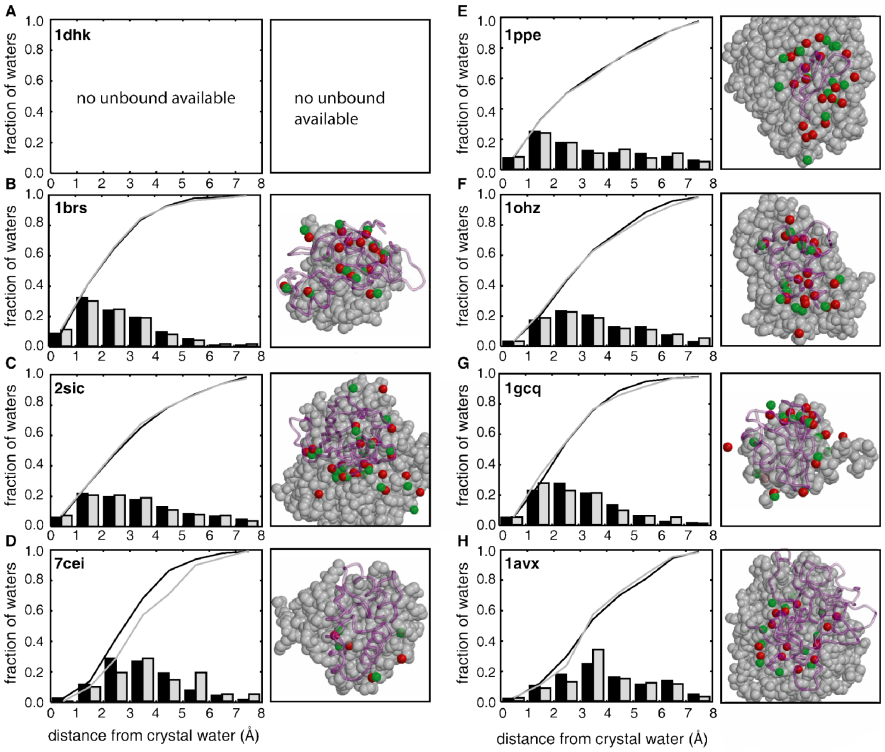


Figure 9. Accuracy of predicted water molecules in unbound solvated docking after semi-flexible refinement. For each of the studied complexes (A-H) the cumulative fraction of the closest distances between modelled waters and crystal waters for all acceptable structures (dark grey) and for the top 10 acceptable structures (light grey) out of the top 200 ranked structures is shown in the left panel. The best-scoring acceptable solvated docking solution is shown in the right panel, together with its predicted waters (red) and the corresponding crystal waters (green).

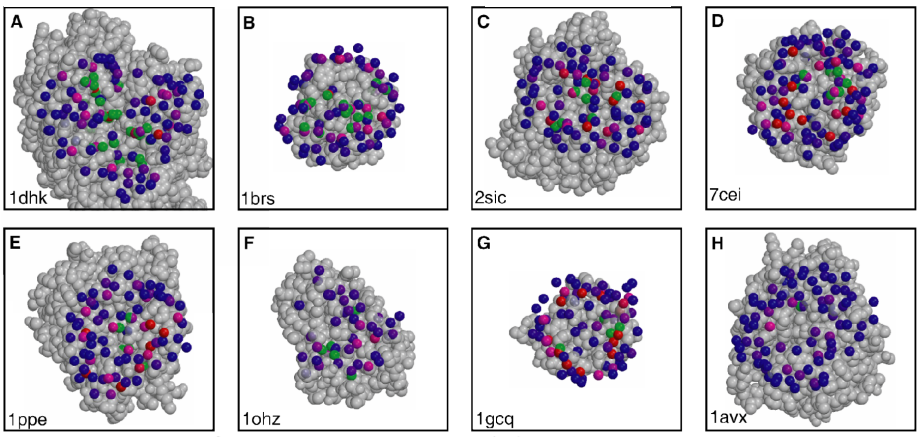


Figure 10. Recovery of interfacial water molecules in bound rigid body solvated docking. For each of the studied complexes (A-H), a cpk representation of the largest component is shown together with its associated crystal waters (transparent green) and cluster representatives from all predicted waters in the acceptable solutions. The latter are color-coded according to the fraction of acceptable structures in which they are observed, from blue 0% to red 40% (maximal observed fraction). Waters from all acceptable solutions were clustered based on pairwise distances using a 2.5Å cut-off.

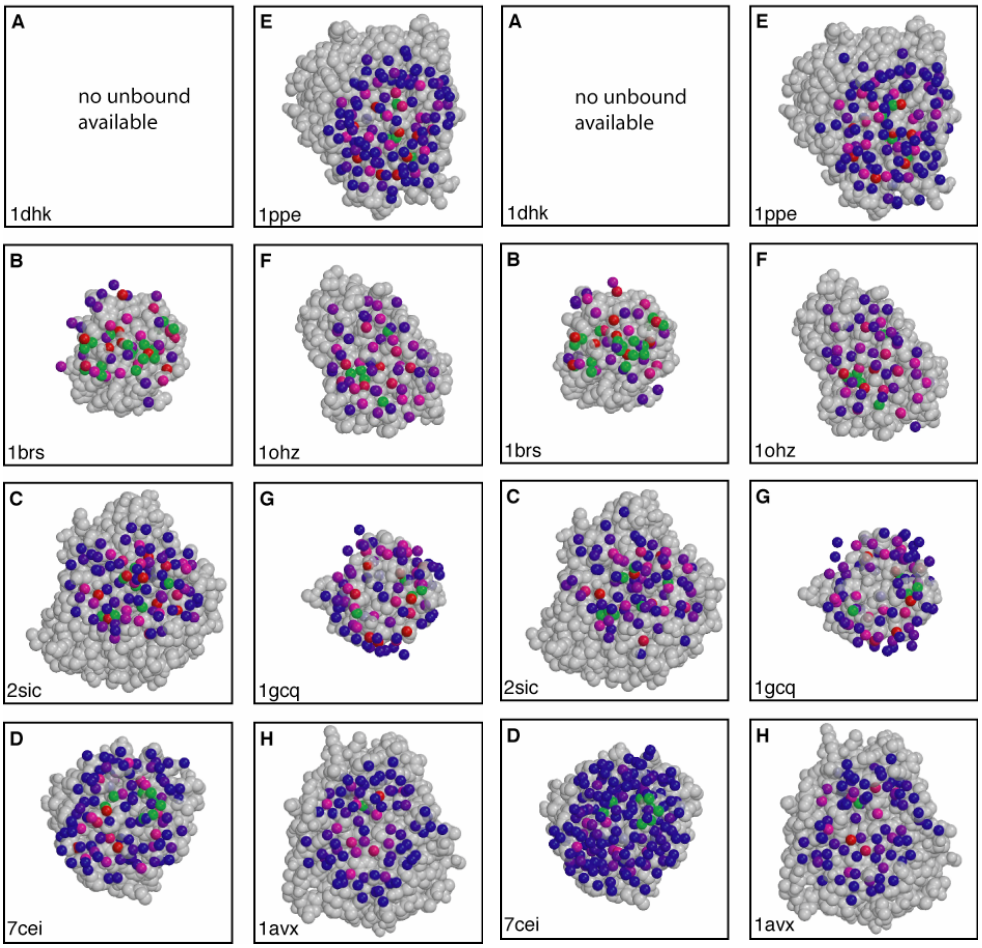


Figure 11 (left) and Figure 12 (right). Recovery of interfacial water molecules in unbound rigid body solvated docking before (figure 11) and after (figure 12) semi-flexible refinement. For each of the studied complexes (A-H), a cpk representation of the largest component is shown together with its associated crystal waters (transparent green) together and cluster representatives from all predicted waters in the acceptable solutions. The latter are color-coded according to the fraction of acceptable structures in which they are observed, from blue 0% to red 40% (maximal observed fraction). Waters from all acceptable solutions were clustered based on pairwise distances using a 2.5Å cut-off.

Table 5. List of PDB and chain id's of the protein-protein complexes used in the analysis of water-mediated contacts: Keskin dataset (R<=2.0)

1a0m	A	B	1aoh	A	B	1cqx	A	B	1ijx	C	D
1a2p	B	C	1b8d	A	L	1dci	A	B	1j7d	A	B
1a8k	A	C	1bb3	A	B	1evr	F	J			
1ae9	A	B	1bxk	A	B	1fs5	A	B			
1aik	N	C	1byo	A	B	1g4y	B	R			

Table 6. Contact statistics

For each amino acid – amino acid contact, the corresponding number of contacts found in the dataset is listed, followed by the number of main chain and side chain water mediated contacts.

ALA	ALA	47	7	2	ASN	ASP	35	3	9	CYS	LEU	21	0	1
ALA	ARG	56	1	11	ASN	CYS	31	0	7	CYS	LYS	6	2	0
ALA	ASN	37	2	9	ASN	GLN	31	0	8	CYS	MET	6	0	0
ALA	ASP	62	6	14	ASN	GLU	26	1	5	CYS	PHE	10	0	0
ALA	CYS	13	0	2	ASN	GLY	36	6	4	CYS	PRO	10	2	0
ALA	GLN	30	2	3	ASN	HIS	13	1	2	CYS	SER	12	4	3
ALA	GLU	71	7	12	ASN	ILE	43	1	2	CYS	THR	15	0	0
ALA	GLY	55	15	1	ASN	LEU	66	1	5	CYS	TRP	4	1	0
ALA	HIS	33	2	6	ASN	LYS	29	1	7	CYS	TYR	16	0	1
ALA	ILE	76	5	8	ASN	MET	22	0	4	CYS	VAL	14	0	2
ALA	LEU	112	3	9	ASN	PHE	21	0	1	GLN	GLN	22	0	6
ALA	LYS	61	7	6	ASN	PRO	27	0	3	GLN	GLU	52	3	16
ALA	MET	46	2	9	ASN	SER	48	2	11	GLN	GLY	32	3	5
ALA	PHE	41	0	3	ASN	THR	28	0	7	GLN	HIS	8	1	1
ALA	PRO	37	5	4	ASN	TRP	7	0	0	GLN	ILE	32	2	3
ALA	SER	95	6	10	ASN	TYR	30	0	5	GLN	LEU	64	3	9
ALA	THR	58	2	2	ASN	VAL	42	1	10	GLN	LYS	29	2	8
ALA	TRP	20	2	1	ASP	ASP	36	5	10	GLN	MET	18	0	4
ALA	TYR	63	0	9	ASP	CYS	13	1	2	GLN	PHE	15	0	0
ALA	VAL	63	0	5	ASP	GLN	30	0	9	GLN	PRO	14	4	4
ARG	ARG	23	2	13	ASP	GLU	51	1	22	GLN	SER	30	4	6
ARG	ASN	42	1	10	ASP	GLY	48	5	1	GLN	THR	23	0	3
ARG	ASP	54	2	19	ASP	HIS	20	4	10	GLN	TRP	13	0	0
ARG	CYS	10	0	2	ASP	ILE	50	1	4	GLN	TYR	19	0	9
ARG	GLN	16	1	4	ASP	LEU	100	3	11	GLN	VAL	26	1	5
ARG	GLU	44	0	15	ASP	LYS	54	4	21	GLU	GLU	32	2	8
ARG	GLY	34	2	7	ASP	MET	42	2	6	GLU	GLY	40	7	4
ARG	HIS	21	2	5	ASP	PHE	26	0	8	GLU	HIS	26	2	17
ARG	ILE	34	1	7	ASP	PRO	31	0	4	GLU	ILE	51	3	10
ARG	LEU	72	3	14	ASP	SER	67	9	9	GLU	LEU	97	3	19
ARG	LYS	31	0	11	ASP	THR	43	1	10	GLU	LYS	58	1	22
ARG	MET	20	0	6	ASP	TRP	14	0	2	GLU	MET	24	0	6
ARG	PHE	22	0	2	ASP	TYR	46	2	11	GLU	PHE	15	0	4
ARG	PRO	29	1	9	ASP	VAL	62	4	15	GLU	PRO	34	0	10
ARG	SER	63	4	20	CYS	CYS	8	0	1	GLU	SER	30	0	14
ARG	THR	32	0	4	CYS	GLN	7	0	0	GLU	THR	47	2	11
ARG	TRP	16	0	3	CYS	GLU	11	1	1	GLU	TRP	6	0	1
ARG	TYR	43	0	12	CYS	GLY	10	4	0	GLU	TYR	26	2	6
ARG	VAL	46	4	9	CYS	HIS	4	0	0	GLU	VAL	48	0	13
ASN	ASN	24	2	9	CYS	ILE	11	1	0	GLY	GLY	32	6	0

GLY	HIS	8	1	1	ILE	THR	48	2	1	PHE	PHE	5	0	0
GLY	ILE	73	3	4	ILE	TRP	14	0	0	PHE	PRO	15	0	0
GLY	LEU	76	9	5	ILE	TYR	41	0	1	PHE	SER	15	0	1
GLY	LYS	23	6	6	ILE	VAL	59	2	3	PHE	THR	36	0	1
GLY	MET	19	4	4	LEU	LEU	94	3	8	PHE	TRP	3	0	0
GLY	PHE	20	2	2	LEU	LYS	74	1	12	PHE	TYR	25	0	3
GLY	PRO	44	6	5	LEU	MET	57	7	8	PHE	VAL	43	0	5
GLY	SER	36	8	4	LEU	PHE	59	0	2	PRO	PRO	10	1	2
GLY	THR	42	3	7	LEU	PRO	49	0	6	PRO	SER	31	1	5
GLY	TRP	9	0	1	LEU	SER	105	2	11	PRO	THR	38	1	8
GLY	TYR	27	2	8	LEU	THR	94	2	10	PRO	TRP	15	0	2
GLY	VAL	53	7	6	LEU	TRP	19	0	2	PRO	TYR	29	4	12
HIS	HIS	7	1	3	LEU	TYR	44	0	4	PRO	VAL	45	1	8
HIS	ILE	27	1	4	LEU	VAL	98	0	10	SER	SER	55	3	4
HIS	LEU	20	0	2	LYS	LYS	24	2	12	SER	THR	51	5	12
HIS	LYS	22	0	6	LYS	MET	24	0	5	SER	TRP	23	0	0
HIS	MET	13	1	4	LYS	PHE	18	0	5	SER	TYR	52	1	6
HIS	PHE	9	0	0	LYS	PRO	15	1	2	SER	VAL	75	7	10
HIS	PRO	12	0	2	LYS	SER	46	2	17	THR	THR	22	1	1
HIS	SER	30	2	10	LYS	THR	37	2	10	THR	TRP	14	0	0
HIS	THR	21	3	6	LYS	TRP	9	1	1	THR	TYR	25	4	1
HIS	TRP	5	0	2	LYS	TYR	21	0	5	THR	VAL	38	0	5
HIS	TYR	24	0	4	LYS	VAL	38	3	7	TRP	TRP	3	0	0
HIS	VAL	13	0	0	MET	MET	15	3	3	TRP	TYR	4	0	1
ILE	ILE	34	1	2	MET	PHE	7	0	0	TRP	VAL	14	0	2
ILE	LEU	110	4	8	MET	PRO	13	0	0	TYR	TYR	26	2	5
ILE	LYS	30	1	6	MET	SER	24	0	3	TYR	VAL	46	3	7
ILE	MET	27	0	1	MET	THR	29	2	8	VAL	VAL	42	2	5
ILE	PHE	38	1	0	MET	TRP	8	1	1					
ILE	PRO	40	3	1	MET	TYR	22	0	5					
ILE	SER	55	2	4	MET	VAL	35	2	4					

Clustering of water-mediated contacts

RMSD-based clustering of water-mediated contacts was performed using a cut-off of 1.5 Å. For each amino acid – amino acid pair, the contacting atoms that resulted in the best clustering (most populated first cluster) were selected. Contacts involving two close waters that would fall into the same cluster, were counted only once. Table 7 reports the number of clusters observed and the fraction of amino-acid contacts falling in the first cluster. Note that the lower the number of clusters, and the larger the size of the first cluster, the better the clustering.

Table 7. Clustering statistics of water-mediated side chain contacts.[a]

Number of clusters	1	2	3	>3
number of amino acid-amino acid combinations	79	38	8	7
Fraction in first cluster[b]	0-25%	25-50%	50-75%	75-100%
number of amino acid-amino acid combinations	10	69	26	27

[a] The total number of amino acid combinations (max. 210) for which clustering could be performed is 132. The remaining amino acid combinations that do not cluster correspond to main chain contacts or to amino acid pairs with zero or only one water-mediated contact.

[b] Fraction of all observed water-mediated contacts within the first cluster.

Note: Tables 8 and 9 are placed on page 124.

Color Figures chapter 2

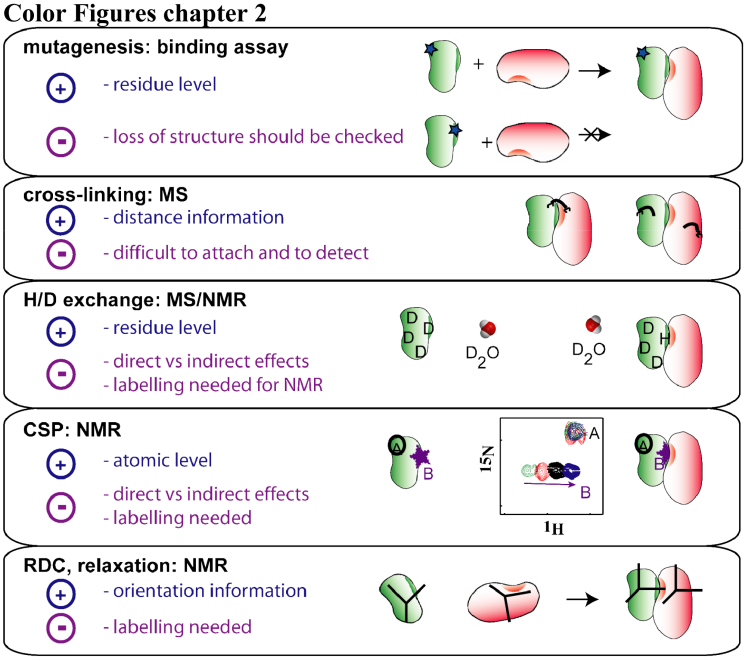


Figure 1.

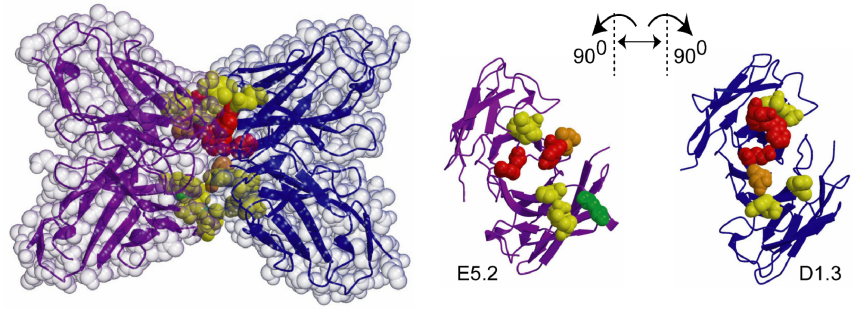


Figure 2.

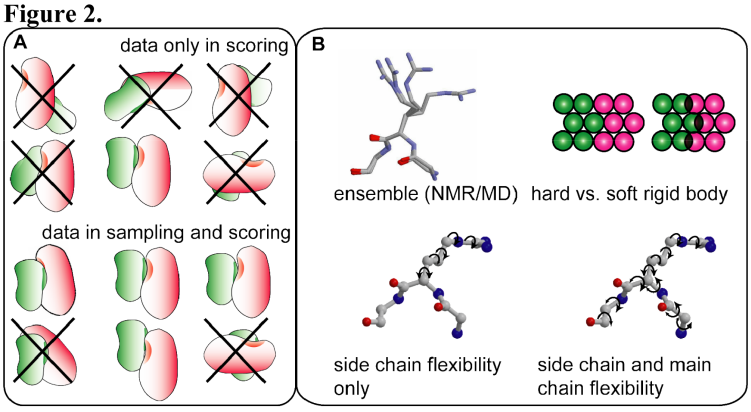


Figure 3.

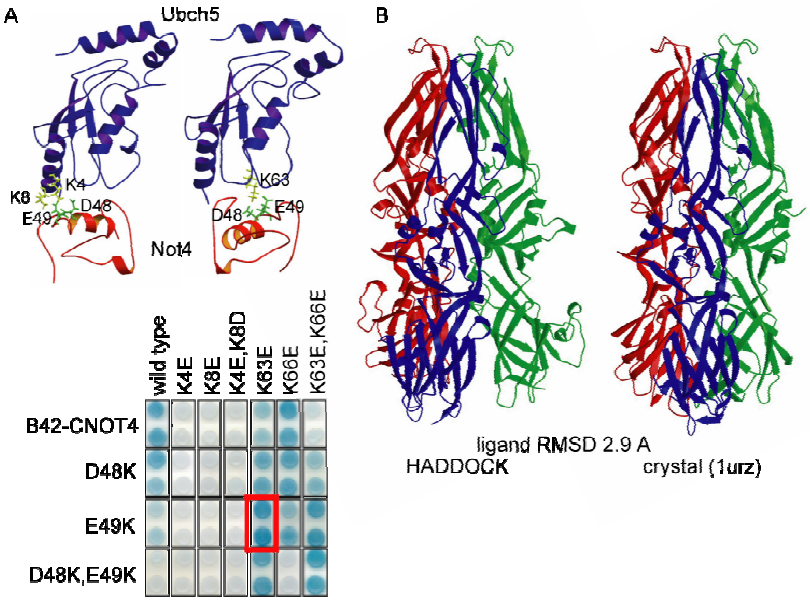


Figure 4.

Color Figures Chapter 3.

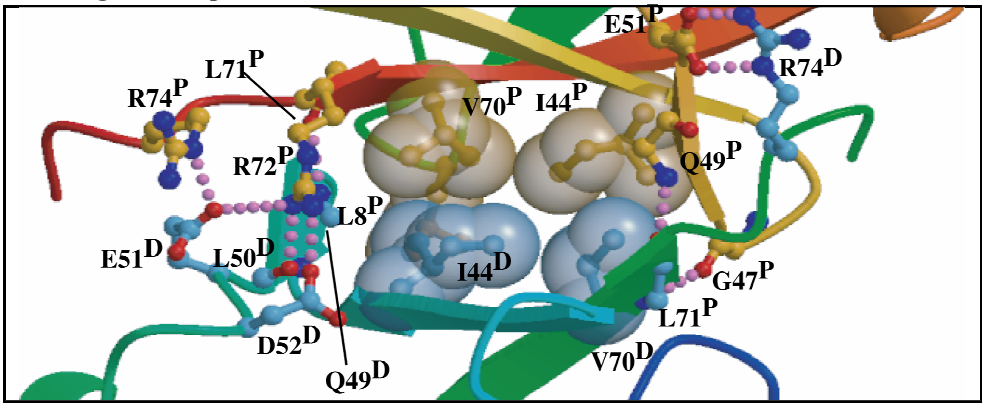


Figure 7.

Table 8. Bound solvated and unsolvated docking results.[a]

		Top 200 <4 Å[b]	Top RMSD[c]	Best Rank[d]	Best RMSD[e]
1dhk	- R	**200**	**0.2**	**1**	**0.1**
	- S	**200**	**0.2**	**1**	**0.1**
1brs	- R	**200**	**0.1**	**1**	**0.1**
	- S	**200**	**0.1**	**1**	**0.1**
2sic	- R	**200**	0.3	**1**	**0.2**
	- S	**200**	0.6	**1**	**0.2**
7cei	- R	**200**	0.4	**1**	**0.3**
	- S	**200**	0.8	**1**	**0.3**
1ppe	- R	**101**	**0.2**	**1**	**0.2**
	- S	**200**	**0.2**	**1**	**0.2**
1ohz	- R	168	0.1	1	0.1
	- S	43	7.7	9	0.3
1gcq	- R	182	**5.3**	**9**	**0.2**
	- S	125	**5.3**	**4**	**0.2**
1avx	- R	**181**	**0.3**	**1**	**0.2**
	- S	**198**	**0.3**	**1**	**0.2**
1kpf	- R	**200**	**0.2**	**1**	**0.2**
	- S	**200**	**0.2**	**1**	**0.2**
1ggr	- R	200	**0.4**	**1**	**0.4**
	- S	189	**0.4**	**1**	**0.4**

[a] Results from reference (R) unsolvated and solvated (S) protein – protein docking for the various test-cases (see Table 2). **Bold** indicates cases where solvated docking performs equal to or better than the unsolvated docking.
[b] The number of structures below the indicated interface-RMSD (i-RMSD) value is reported (<4Å, acceptable). The i-RMSD is calculated over the backbone atoms of all residues making contacts across the interface within a 10Å cut-off.
[c] i-RMSD of highest scoring solution.
[d] Rank of best-ranked structure below 1Å i-RMSD.
[e] i-RMSD value of lowest i-RMSD structure.

Table 9. Recovery of water molecules in bound solvated docking.[a]

	<#waters>[b]	recovery[c] #	$f_{recover}$ (%)[d]
1dhk	8.6 (6.0)	24 /25	12 (7)
1brs	6.1 (4.5)	18 /18	12 (8)
2sic	8.8 (3.8)	7/8	19 (14)
7cei	10.1 (3.8)	8/8	13 (9)
1ppe	8.5 (4.0)	6/6	11 (4)
1ohz	7.6 (3.8)	5/5	19 (7)
1gcq	10.3 (3.6)	4/4	33 (25)
1avx	5.4 (5.0)	1/1	8

[a] Solvated docking results for the acceptable solutions out of the top 200 models for the various test-cases (see Table 2).
[b] Average number (standard deviation) of water molecules per structure.
[c] Number of fully buried crystal waters recovered (i.e. within 2.0Å of a modeled water) / total number of buried crystal waters (see Table 2).
[d] Average fraction (standard deviation) of acceptable structures in which a fully buried water is recovered.

Color Figures Chapter 5.

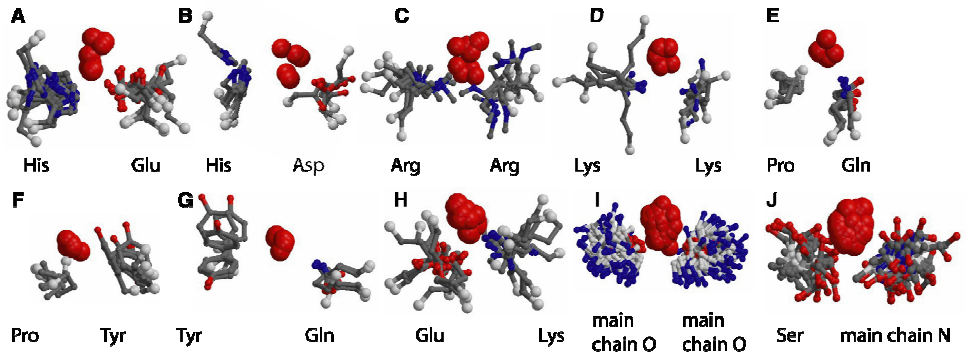


Figure 2.

For figures 3 and 4, see figures 7A, 7B, 8B and 8F (figures 3A-3D) and figures 10A, 10B, 11B, 11F (figures 4A-4D), respectively, in "Supplementary Material" for chapter 5 (above).

Color Figures Chapter 6.

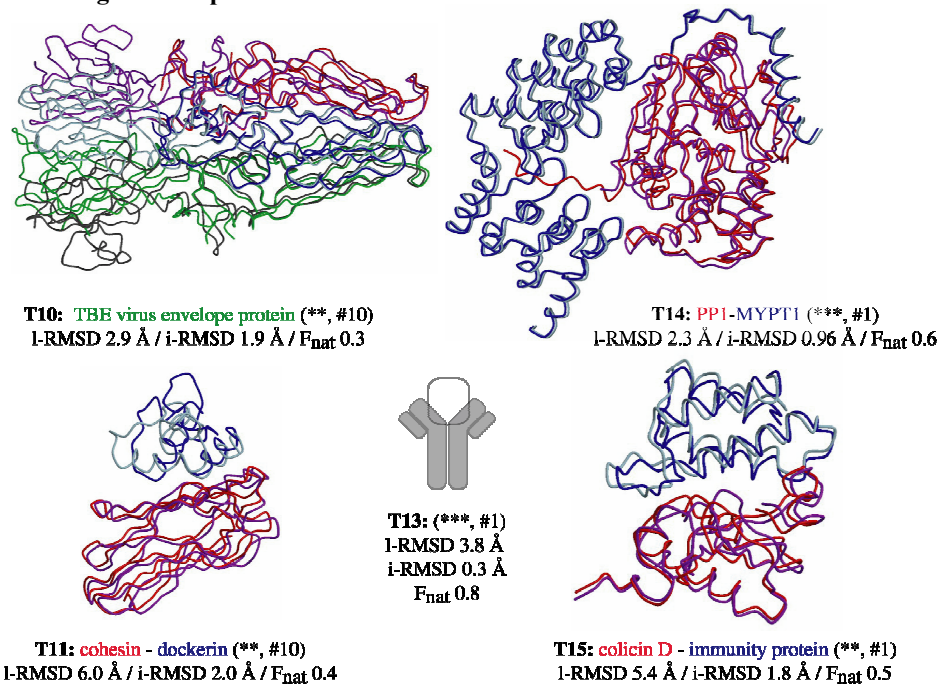


Figure 1.

Color Figures Chapter 7.

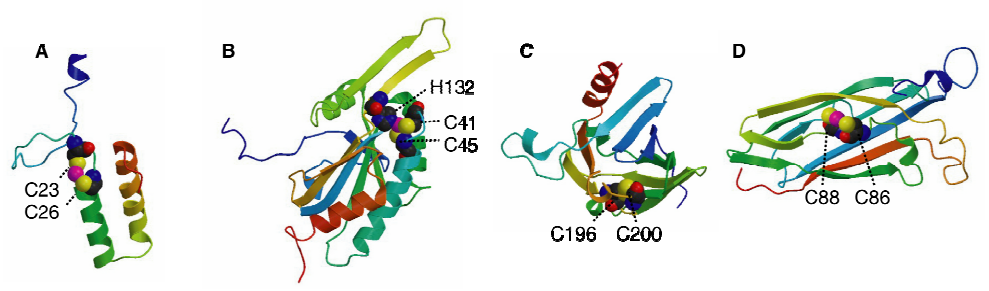


Figure 2.

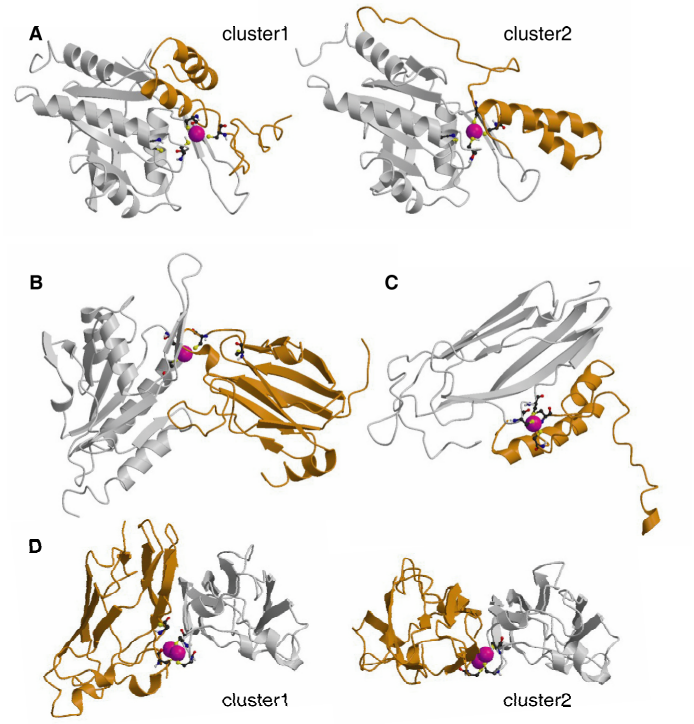


Figure 4.

Summary

Proteins play a major role in biology by interacting with each other and with other biomolecules such as lipids, nucleic acids and various smaller ligands. The study of these interactions is of fundamental importance to understand cellular processes, and this could be a key towards understanding mechanisms of diseases and possible development of drugs.

As a result, considerable attention is given in the current post-genomic era to the study of protein interactions. As described in the general introduction of this thesis (Chapter 1), various methods have been used in order to find out the partners with which a given protein interacts ("interaction mapping"). From the point of view of structural biology, a key task is to unravel these interactions at atomic details: the structure of a biological molecule can explain its function and provides a starting point for further experiments. However, traditional structure determination methods (X-ray crystallography and NMR) encounter serious experimental difficulties in dealing with structures of protein complexes. Therefore, complementary computational methods are needed – both to predict structures of systems that cannot be approached experimentally and also to cope with the large number of expected complexes.

This thesis describes the modeling of biomolecular complexes by data-driven docking, a computational method that, based on the known structures of the constituents of a complex and any information about their interface, derives a model for the structure of the complex. Various experimental data such as obtained by mutagenesis or from a variety of NMR experiments can give information about protein – protein interfaces in a relatively easy way (compared to a full structure determination). Chapter 2 explains the basics of the docking methodology and reviews several approaches towards the use of experimental information in docking. It also introduces the data-driven docking method HADDOCK which we used in all the modelling described in this thesis.

Chapters 3 and 4 focus on the use of NMR information in data-driven docking. Chemical shift perturbation (CSP) data are obtained by recording HSQC spectra of one ^{15}N-labelled partner in the complex in the absence and presence of increasing amounts of the partner protein. Changes in chemical shifts allow assessment of the interface in the complex. These data can be combined with Residual Dipolar Couplings (RDCs, chapter 3) or diffusion anisotropy data (chapter 4). Both RDCs and diffusion anisotropy data provide orientational restraints and are complementary to the information provided by the CSP data.

In Chapter 3 this methodology is applied towards the determination of the solution structure of the ubiquitin dimer. Poly-ubiquitin chains are the signal by which proteins are recognized by the proteasome, resulting in their degradation. The smallest ubiquitin chain consists of a dimer. A crystal structure of such a dimer was already available; this structure, however, does not fit the NMR data as well as the solution structure which we describe in this chapter. The difference between the two structures is a slight rotation (~ 20 degrees) of the two ubiquitin moieties, indicating a subtle difference between solution and crystal structures.

In Chapter 4 the usefulness of diffusion anisotropy data in modelling biomolecular complexes is assessed by generating synthetic datasets and comparing docking results under various assumptions with regard to the diffusion anisotropy data (e.g. various amounts of noise and various degrees of anisotropy).

In Chapter 5 we turn towards the docking methodology itself. In all current docking approaches, modelling takes place in vacuum, completely ignoring the possible role of water in biomolecular recognition. In order to explicitly introduce water into the modelling process, we first investigate its possible structural role at protein-protein interfaces by analyzing a set of complexes, and then describe a novel solvated docking protocol that accounts for water in the docking process. It is based on the concept of the "first encounter complex" in which a

water layer is present in between the two molecules in the first recognition step. Our approach thus mimics the pathway from the initial encounter complex towards the final assembly in which most water molecules have been expelled from the interface. In our protocol, the two protein chains are solvated prior to docking, and a large fraction of the water is subsequently removed in the first rigid body docking stage based on water-mediated contact probabilities derived from a set of high-resolution crystal structures. We demonstrate the feasibility of this approach on complexes with known structures, both with dry and wet interfaces, and show that the docking results in general improve. Solvated docking is thus a viable tool in the study of biomolecular complexes. In addition, it adds considerable information to existing docking approaches, since it allows to predict structural water molecules at an interface, which can be particularly valuable for drug design for example.

The last two chapters are directed towards validation and application of our docking methodology. Chapter 6 describes the results of our participation to the blind docking experiment CAPRI (Critical Assessment of Predicted Interactions). In CAPRI, participants must make blind predictions of the structure of a complex within a limited time; these are then compared by independent assessors to the yet unpublished experimental structure. All major groups worldwide that develop methods to predict the structure of complexes participate. CAPRI is instrumental in stimulating new developments in the docking field. Our participation to CAPRI was quite successful: for many of the targets we managed to generate medium or even high quality results, and our overall performance places us in the top of the field. Chapter 6 discusses our approach and also gives a detailed look at the impact of flexibility on our docking results. Flexibility is an important issue in docking since the free and bound conformations of a protein can differ. Our analysis shows that HADDOCK, which is quite unique in its treatment of both side chain and backbone flexibility, is able to deal with small to moderate conformational changes.

In the final chapter, HADDOCK is applied to the study of complexes along the cytochrome c oxidase copper-delivery pathway. This is an important process since the correct assembly of cytochrome c oxidase (which generates the proton gradient that drives synthesis of ATP, the major biological energy carrier) depends on the copper-delivery pathway. So far, structural details about interactions between the proteins involved in this pathway (cox17, sco1, cox2 and cox11) are unknown. Here, models for the structures of the human complexes cox17-sco1, sco1-cox2, cox17-cox11 and cox11-cox11 were generated, based on information from interface prediction methods. The resulting models shed light on the path that copper travels during its way towards cytochrome c oxidase and provide new starting points for experimental studies.

Samenvatting

Eiwitten vervullen hun uitermate belangrijke rol in biologische processen door interacties aan te gaan met elkaar en met andere biomoleculen, zoals lipiden, nucleinezuren en allerlei kleine liganden. Het bestuderen van deze interacties is nodig om cellulaire processen te begrijpen. Daarnaast kan het een belangrijke bijdrage leveren aan het ontrafelen van de mechanismes van ziektes en het ontwikkelen van medicijnen.

In het huidige post-genomic tijdperk wordt dan ook veel aandacht besteed aan het bestuderen van interacties van eiwitten. Er zijn, zoals beschreven in de algemene introductie van dit proefschrift (hoofdstuk 1), verschillende methodes om te bepalen met welke partners een eiwit interacties aangaat ('interactie inventarisatie'). Een belangrijke taak voor structuur-biologie is om deze interacties op atomair niveau te begrijpen: de structuur van een biomolecuul kan zijn functie verklaren en geeft ideeën voor nieuwe experimenten. Vaak is het echter zeer lastig om de structuur van eiwit-complexen te bepalen met behulp van de traditionele structuurbepalingsmethoden (röntgen-kristallografie en NMR). Daarom zijn er aanvullende computer-berekeningsmethoden nodig om structuren van complexen te berekenen die experimenteel niet bepaald kunnen worden, en meer in het algemeen ook vanwege het grote aantal interacties tussen eiwitten.

Dit proefschrift beschrijft het modelleren van biomoleculaire complexen door middel van data-gestuurde dokking, een computer-berekeningsmethode die gebruik maakt van de bekende structuren van de partners in een interactie en beschikbare informatie over het grensvlak tussen de verschillende partners. Er zijn allerlei experimentele methoden (zoals mutagenese of verschillende NMR experimenten) die relatief gemakkelijk informatie verschaffen over het grensvlak tussen de verschillende componenten van een complex (relatief gemakkelijk vergeleken met een volledige experimentele structuurbepaling van het complex). Hoofdstuk 2 beschrijft de dokking-methodologie en geeft een overzicht van verschillende manieren om experimentele informatie in dokking te gebruiken. Het bespreekt ook de data-gestuurde dokking methode HADDOCK die in al het modelleerwerk in dit proefschrift wordt gebruikt.

Hoofdstuk 3 en 4 behandelen het gebruik van NMR-gegevens in data-gestuurde dokking. Veranderingen in chemische verschuivingen (*chemical shift perturbations, CSP*) zijn het resultaat van een NMR-experiment waarin eenvoudige spectra worden gemeten voor een ^{15}N-gelabeld component van een complex in afwezigheid en aanwezigheid van een toenemende hoeveelheid ongelabeld partner-eiwit. Veranderingen in chemische verschuivingen vertellen dan iets over de locatie van het grensvlak op het gelabelde eiwit. Deze data kunnen gecombineerd worden met Residuele Dipolaire Koppelingen (*Residual Dipolar Couplings, RDCs*, hoofdstuk 3) of diffusie anisotropie data (hoofdstuk 4). Zowel RDCs als diffusie anisotropie data bevatten informatie over de orientatie van de partners in het complex ten opzichte van elkaar en vullen de informatie in CSP data aan.

In hoofdstuk 3 wordt deze methodologie gebruikt om de structuur van de ubiquitine dimeer in oplossing te bepalen. Ketens van ubiquitine moleculen zijn het signaal waardoor eiwitten worden herkend door het proteasoom, wat leidt tot hun afbraak. De kleinste ubiquitine-keten is een dimeer. Er was al een kristalstructuur voor zo'n dimeer beschikbaar, maar die komt niet zo goed overeen met de beschikbare NMR data als de structuur in oplossing die in dit hoofdstuk wordt beschreven. Het verschil tussen de twee structuren is een rotatie van zo'n 20 graden van de twee ubiquitin moleculen ten opzichte van elkaar; dit laat zien dat er subtiele verschillen kunnen zijn tussen de structuur van een complex in kristalvorm en in oplossing.

In Hoofdstuk 4 wordt het nut van diffusie anisotropie data in data-gestuurde dokking onderzocht door verschillende synthetische datasets (met verschillende anisotropie en

verschillende hoeveelheid ruis) te genereren en de resultaten van de dokking met elkaar te vergelijken.

In hoofdstuk 5 wordt een nieuwe dokking-methode voorgesteld. In alle bestaande dokking-methodes vindt dokking plaats in vacuüm, zonder de mogelijke rol van water in biomoleculaire herkenning in ogenschouw te nemen. Om water expliciet mee te nemen in het modelleren van biomoleculaire complexen wordt eerst een analyse uitgevoerd van een aantal complexen om de structuur-gerelateerde rol van water aan biomoleculaire grensvlakken te bepalen. Daarna wordt een nieuwe methode beschreven die water introduceert tijdens het dokking-proces. Deze methode is gebaseerd op het concept van 'eerste ontmoetings-complex' waarin een waterlaag aanwezig is tussen de twee moleculen; we simuleren het proces vanaf dit ontmoetings-complex tot het uiteindelijke complex waar het meeste water verdwenen is. In ons protocol worden de twee eiwitketens gesolvateerd voorafgaande aan de dokking en wordt vervolgens tijdens de dokking het meeste water verwijderd op basis van water-gemedieerde contact-waarschijnlijkheden die worden bepaald aan de hand van een set hoge resolutie kristal structuren. We laten zien dat deze gesolvateerde-dokking-methode werkt voor zowel 'droge' als 'natte' grensvlakken, en dat deze methode leidt tot betere dokking-resultaten. Gesolvateerde dokking voegt informatie toe aan bestaande dokking-methodes omdat het water-posities voorspelt aan het grensvlak; dit kan belangrijk zijn voor bijvoorbeeld het ontwerpen van medicijnen.

De laatste twee hoofdstukken bespreken validatie en toepassing van onze data-gestuurde dokking-methode. Hoofdstuk 6 behandelt onze deelname in CAPRI (Critical Assessment of PRedicted Interations, Kritische Evaluatie van Voorspelde Interacties). Deelnemers aan CAPRI moeten binnen een bepaalde tijd een blinde voorspelling maken van de structuur van een complex; deze voorspelling wordt dan door een onafhankelijke jury vergeleken met de nog niet gepubliceerde experimentele structuur. Alle belangrijke groepen uit de hele wereld die methoden ontwikkelen om structuren van eiwit-complexen te modelleren doen mee aan CAPRI. Daardoor is CAPRI een belangrijke stimulans voor nieuwe ontwikkelingen in dokking. Onze deelname aan CAPRI was erg succesvol: voor de meeste complexen wisten we medium- of zelfs hoge-kwaliteit structuren te voorspellen, en we hoorden duidelijk bij de top van het veld. Dit hoofdstuk laat zien welke methodes we gebruikten en laat ook specifiek zien hoe belangrijk flexibiliteit is. Dit laatste is een belangrijk thema omdat in het algemeen de ongebonden en gebonden conformatie van een eiwit verschillend zijn. Onze analyse laat zien dat HADDOCK, dat bijzonder is omdat het zowel zijketen- als hoofdketen-flexibiliteit toelaat, in staat is om conformatie-veranderingen te modelleren, zolang ze niet al te groot zijn.

In het laatste hoofdstuk wordt HADDOCK gebruikt om eiwit-eiwit complexen te bestuderen die betrokken zijn bij het afleveren van koper aan cytochrome c oxidase. Dit is een belangrijk proces omdat het correcte functioneren van cytochrome c oxidase hier van afhangt, en cytochrome c oxidase genereert de proton-gradiënt die nodig is voor de synthese van ATP, de belangrijkste biologische energie-drager. Tot nu toe zijn er geen details bekend over de structuur van de interacties tussen de eiwitten die bij dit koper-afleveringsproces betrokken zijn (cox17, sco1, cox2 en cox11). In dit hoofdstuk modelleren we de complexen cox17 – sco1, sco1 – cox2, cox17 – cox11 en cox11 – cox11, door gebruik te maken van grensvlakpredictie-methodes en data-gestuurde dokking. Deze modellen geven inzicht in de route van koper naar cytochrome c oxidase en kunnen dienen als startpunt voor nieuwe experimenten.

Dankwoord

Na vier jaar promotie-onderzoek ben ik nu eindelijk toe aan het onderdeel van dit proefschrift dat het meest gelezen wordt. Ik heb me vier jaar lang bezig gehouden met computerberekeningen aan de structuur van eiwit-eiwit complexen, dat wil zeggen eiwitten die 'interacties' met elkaar aangaan. Hier wil ik iedereen bedanken met wie ik de afgelopen jaren een positieve interactie gehad heb. 'Iedereen' bedanken kan natuurlijk niet, ik zal een selectie moeten maken. Sorry als je er niet bij staat, maar dat betekent niet dat ik je ben vergeten.

Als ik terug kijk op de afgelopen jaren is er een duidelijke cesuur: het overlijden van mijn vader in oktober 2004. Naast het verdriet heeft het me ook laten zien dat veel dingen, inclusief onderzoek doen aan eiwit-eiwit interacties, relatief onbelangrijk zijn; en hoe belangrijk interacties met anderen, je familie, vrienden, vooral ook God, onze Vader.

Hanneke, jou wil ik als eerste bedanken voor je steun en voor de chemie tussen ons. Verder ben ik erg blij met de voorkant die jij ontworpen hebt. Mijn familie, ma, opa en oma, en mijn broers en zussen wil ik bedanken voor de goede relatie die we met elkaar hebben en omdat we altijd op elkaar kunnen rekenen. Ik vond de interesse voor de vorderingen van mijn onderzoek altijd erg leuk; degenen die hebben toegezegd het hele boekje te gaan lezen ga ik daar natuurlijk wel aan houden…..

Jan en Cees, jullie ben ik dankbaar omdat jullie de taak van paranimf willen vervullen. Ook maken jullie samen met Anne Marie, Margreet en Hanneke deel uit van de studiekring waar we naast het lezen van literatuur variërend van Calvijn (1568) tot Wieringa (2006) het ook vaak over de vorderingen en de (on)zin van promotie-onderzoek hebben gehad. Geerten, jou wil ik bedanken voor het spelen van Go, wat me naast de mooie avonden die dat opleverde ook heeft geholpen om scherp te blijven. De donderdag-avond hardloopclub is intussen een vast onderdeel van mijn tijdsbesteding: Khanh, Thuy-Nga en Noelle, jullie ook bedankt.

Veel vriendschappen heb ik de afgelopen jaren opgedaan in de Mattheüskerk. Het is onmogelijk om iedereen hier te noemen, maar vooral de bijbelkring van de afgelopen paar jaar heeft veel voor me betekend.

Uiteraard wil ik ook mijn collega's en iedereen met wie ik een interactie heb gehad binnen het terrein van mijn onderzoek bedanken. Alexandre en Rolf, bedankt voor jullie enthousiaste begeleiding. Ik heb het erg gewaardeerd dat er veel ruimte was voor eigen ideeën en tegelijk ook tijd voor discussie en goede adviezen. Many thanks to my present roommates, Devasish and Kostas; I wish you all the best. I want to thank Cyril, my former roommate, for his support when he was still in the lab. Hetzelfde geldt voor Henry, Monique, Aart, Eiso en alle andere voormalige collega's. Verder wil ik speciaal Mark en Sjoerd noemen; ik heb jullie als master-student begeleid, en ik vond het erg leuk dat jullie ook als AIO verder zijn gegaan. Gloria, thanks for discussions and support. Barbara, bedankt omdat je veel hebt gedaan wat betreft de administratieve dingen die bij een promotie horen. En ook Johan, Michiel, Mark, Mickael, Aurelien, Hans, Rob, Julija, Anding, Jeff, Hugo, Gert, Suat, Karine, Rainer en Tammo: bedankt.

Finally I want to thank those with whom I collaborated during those four years. Prof. Fushman provided much of the input for chapter 3; thanks for this. Prof. Bertini, prof. Banci, and especially Simone Ciofi-Baffoni: thanks for your collaboration and your kind hospitality when I stayed in Florence. Ook Erik Kalkhoven en Ellen Jeninga, bedankt voor jullie samenwerking. Tenslotte bedank ik de leescommissie, prof. Kaptein, prof. Gros, prof. Frenkel en prof. Janin.

List of publications

A.D.J. van Dijk, S. Ciofi-Baffoni, L. Banci, I. Bertini, R. Boelens and A.M.J.J. Bonvin, "Modelling protein-protein complexes along the cytochrome c oxidase Cu-delivery pathway", *manuscript in preparation.*

E. Jeninga, O. van Beekum, **A.D.J. van Dijk**, N. Hamers, B. Hendriks-Stegeman, A.M.J.J. Bonvin, R. Berger and E. Kalkhoven, "The Familial Partial Lipodystrophy-associated R425C mutation impairs PPARγ function on multiple levels", submitted to *Molecular Endocrinology*.

G. Fuentes, **A.D.J. van Dijk**, A.M.J.J. Bonvin, "NMR-based modelling and refinement of protein 3D structures and their complexes", contribution to "Molecular Modelling of Proteins", Humana Press, 2006.

A.D.J. van Dijk and A.M.J.J. Bonvin, "Solvated docking: introducing water into the modeling of biomolecular complexes", *Bioinformatics*, in press.

M. van Dijk, **A.D.J. van Dijk**, V. Hsu, R. Boelens and A.M.J.J. Bonvin, "Information driven protein-DNA docking using HADDOCK: it is a matter of flexibility", *Nucleic Acids Research* (2006), 34, 3317 – 3325.

A.D.J. van Dijk, R. Kaptein, R. Boelens and A.M.J.J. Bonvin, "Combining NMR relaxation with chemical shift perturbation data to drive protein-protein docking" *J. Biomol. NMR* (2006), 34, 237 – 244.

S.J. de Vries, **A.D.J. van Dijk** and A.M.J.J. Bonvin, "WHISCY: WHat Information does Surface Conservation Yield? Application to data-driven docking.", *Proteins: Struc. Funct. & Bioinformatics* (2006), 63, 479 – 489.

A.D.J. van Dijk, S.J. de Vries, C. Dominguez, H. Chen, H.-Z. Zhou and A.M.J.J. Bonvin, "Data-driven docking: HADDOCK's adventures in CAPRI", *Proteins: Struc. Funct. & Bioinformatics* (2005), 60, 232 – 238.

A.D.J. van Dijk, D. Fushman and A.M.J.J. Bonvin, "Various strategies of using residual dipolar couplings in NMR-driven protein docking: application to Lys48-linked di-ubiquitin and validation against ^{15}N-relaxation data", *Proteins: Struc. Funct. & Bioinformatics* (2005), 60, 367 – 381.

A.D.J. van Dijk, R. Boelens and A.M.J.J. Bonvin, "Data-driven docking for the study of biomolecular complexes", *FEBS Journal* (2005), 272, 293 – 312.

J.P.J.M van der Eerden, M.M.E. Snel, J. Makkinje, **A.D.J. van Dijk** and H.A. Rinia, "Striped phases in thin layers: simulation and observation", *J. Cryst. Growth* (2002), 237, 111-115.

A.D.J. van Dijk "Modelling of biomolecular complexes by data-driven docking"