J|A|C|S
ARTICLES
Published on Web 01/21/2003

HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information

Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin*

Contribution from the Department of NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, 3584CH, Utrecht, The Netherlands

Received May 17, 2002; E-mail: abonvin@nmr.chem.uu.nl

Abstract: The structure determination of protein−protein complexes is a rather tedious and lengthy process, by both NMR and X-ray crystallography. Several methods based on docking to study protein complexes have also been well developed over the past few years. Most of these approaches are not driven by experimental data but are based on a combination of energetics and shape complementarity. Here, we present an approach called HADDOCK (High Ambiguity Driven protein−protein Docking) that makes use of biochemical and/or biophysical interaction data such as chemical shift perturbation data resulting from NMR titration experiments or mutagenesis data. This information is introduced as Ambiguous Interaction Restraints (AIRs) to drive the docking process. An AIR is defined as an ambiguous distance between all residues shown to be involved in the interaction. The accuracy of our approach is demonstrated with three molecular complexes. For two of these complexes, for which both the complex and the free protein structures have been solved, NMR titration data were available. Mutagenesis data were used in the last example. In all cases, the best structures generated by HADDOCK, that is, the structures with the lowest intermolecular energies, were the closest to the published structure of the respective complexes (within 2.0 Å backbone RMSD).

For a better understanding of the biological function of a protein, knowledge of its three-dimensional structure is crucial. Solving protein structures is mainly achieved by two different methods: X-ray crystallography and nuclear magnetic resonance (NMR). From the statistics of the protein data bank (PDB)[1] (http://www.rcsb.org/pdb/), approximately 13 500 X-ray structures and 2225 NMR structures have been solved and deposited at this date. Most of the proteins achieve their function by interacting with other proteins and forming an active complex. Although many methods are available to study protein complexes at different levels (two-hybrid screening, fluorescence studies, resonance energy transfer, etc.), only few of these techniques provide high-resolution information at an atomic level. X-ray and NMR encounter difficulties in dealing with structures of complexes. Indeed, by X-ray, the dynamic of the complex formation makes the crystallization difficult, while the size limitation in NMR is a major problem when considering high molecular weight complexes. The traditional NMR approach to solving protein−protein complexes requires the collection of intermolecular nuclear Overhauser effect (NOE) distances, which is typically a lengthy and difficult process. In addition, intermolecular NOEs often involve side chain protons, requiring thus a rather complete assignment of all NMR signals. Because of these limitations, the number of protein−protein complexes solved and deposited in the PDB is rather low (643 by X-ray and 84 by NMR) compared with the number of free form structures. NMR, however, is very powerful in mapping protein−protein interfaces by titration experiments (reviewed in ref 2). Such experiments, which can be performed at the stage of backbone assignment already, easily allow us to identify amino acids involved in the complex formation but do not provide any information about the orientation of one protein with respect to its counterpart. Because of that, this information has rarely been directly used as a structural restraint in a structure calculation process. Next to these experimental approaches, theoretical methods to study protein complexes at a structural level based on docking are now emerging that have been well developed during the past few years. There are now a number of programs performing "ab initio" protein−protein docking (for review, see refs 3 and 4). Most of these programs use the same approach: one protein is fixed in space and the second one is rotated and translated around the first one. For each new configuration, a score is calculated on the basis of various terms such as surface complementarities, electrostatic interactions, van der Waals repulsion, and so forth. The drawback of these methods is that the search through the entire conformational space of the complex geometry makes the calculation heavy, rarely resulting in an unique solution. Recently, NMR data have been used in combination with docking methods in different ways to generate protein−protein complexes. Diamagnetic

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(2) Zuiderweg, E. R. *Biochemistry* **2002**, *41*, 1−7.
(3) Smith, G. R.; Sternberg, M. J. *Curr. Opin. Struct. Biol.* **2002**, *12*, 28−35.
(4) Camacho, C. J.; Vajda, S. *Curr. Opin. Struct. Biol.* **2002**, *12*, 36−40.

chemical shift changes and intermolecular pseudocontact shifts have been combined with restrained rigid-body molecular dynamics to solve the structure of the paramagnetic plastocyanin−cytochrome f complex.[5] Intermolecular NOEs and residual dipolar couplings (RDCs) have been combined to solve the structure of the EIN−HPr complex.[6] Intermolecular NOEs can very accurately define the interface. Their collection, however, is generally a tedious process. In addition, RDC data can be very useful to determine the relative orientation of the two proteins. Morelli et al. used the program BIGGER,[7] which makes use of an NMR filter on the basis of chemical shift perturbation data.[8] This approach allows the use of NMR titration data to rank the possible solutions, but the docking is not directly driven by these data. Recently, Fahmy et al. developed a new docking program, TreeDock, where the docking is oriented on the basis of anchors points which can be in principle derived from NMR chemical shift perturbation or mutagenesis data.[9] This program performs a rigid body docking, and the solutions are ranked in function of their Lennard−Jones potentials. McCoy et al. used chemical shift perturbation data in combination with RDCs to develop a new docking approach.[10] In that case, the RDC data are first introduced to orient the complex, and then the solutions are optimized by back calculating chemical shift perturbation with the SHIFTS software[11] and comparing them with the experimental data. During complex formation, usually, some structural rearrangements occur. By NMR titration, it is possible to check such rearrangements for backbone atoms, but no information is available on the side chain rearrangements that occur frequently at the interface, especially in the case of hydrophobic interfaces. It is therefore important, when docking two proteins, to consider the best orientation of their side chains leading to the minimum energy and the best side chain contacts. For this, the side chains at the interface should be free to adapt their conformation.

Here, we present a new high ambiguity driven docking approach (HADDOCK) that makes use of biochemical and/or biophysical interaction data such as, for example, chemical shift perturbation data obtained from NMR titration experiments or mutagenesis data. The information on the interacting residues is introduced as ambiguous interaction restraints (AIRs) to drive the docking. After calculation, the structures are ranked according to their intermolecular energy, that is, sum of electrostatic, van der Waals, and AIR energy terms. We should note that ambiguous distance restraints have first been introduced to solve symmetric dimer structures by NMR[12] and are now commonly used in protein structure determination and automated NOE assignment methods.[13] We demonstrate the usefulness of the AIRs and the accuracy of our docking approach for three different molecular complexes: the N-terminus domain of Enzyme I (EIN) in complex with the histidine-containing phosphocarrier protein (HPr), the Enzyme IIA^{glucose} (E2A) in complex with HPr, and the HIV protein gp120 in complex with the protein CD4. The structures of the first two complexes have been solved by NMR,[14,15] and their respective free forms are available from X-ray and/or NMR.[16−19] The NMR titration data of each protein upon complex formation to its partner are available.[20−22] For the gp120−CD4 complex, however, only the X-ray structures[23,24] of the individual partners of a complex were used. Instead of NMR titration data, mutagenesis data[25,26] were used to define ambiguous interaction restraints. In all three cases, starting from the complex or the free state structures, we found that the best solutions generated by HADDOCK, that is, the structures with the lowest intermolecular energy term, were those that are the closest in terms of backbone root-mean-square deviations at the interface (iRMSD) (between 0.8 and 2 Å) to the published structure of the respective complexes.

Results

Ambiguous Interaction Restraints (AIRs). The ambiguous interaction restraints are derived from any kind of experimental information available concerning residues that are involved in the intermolecular interaction. We distinguish here between "active" and "passive" residues. In the case of NMR titration data, the active residues correspond to all residues showing a significant chemical shift perturbation upon complex formation as well as a high solvent accessibility in the free form protein (>50% relative accessibility as calculated with NACCESS[27]). The threshold to define significant chemical shift perturbations will differ for each protein complex under study and needs some optimization by the user. In our examples, we used as starting point the residues that the authors of the original papers[20−22] defined as significantly perturbed in the complex. These perturbed residues that do not satisfy the high solvent accessibility criterion should be subsequently removed from the active residue list. In the case of mutagenesis data, the active residues are those that have been shown by mutations to alleviate complex formation and are also solvent exposed. The passive residues correspond to the residues that show a less significant chemical shift perturbation and/or that are surface neighbors of the active residues and have a high solvent accessibility (>50%).

(5) Ubbink, M.; Ejdeback, M.; Karlsson, B. G.; Bendall, D. S. *Structure* **1998**, *6*, 323−335.
(6) Clore, G. M. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9021−9025.
(7) Palma, P. N.; Krippahl, L.; Wampler, J. E.; Moura, J. J. *Proteins* **2000**, *39*, 372−384.
(8) Morelli, X. J.; Palma, P. N.; Guerlesquin, F.; Rigby, A. C. *Protein Sci.* **2001**, *10*, 2131−2137.
(9) Fahmy, A.; Wagner, G. *J. Am. Chem. Soc.* **2002**, *124*, 1241−1250.
(10) McCoy, M. A.; Wyss, D. F. *J. Am. Chem. Soc.* **2002**, *124*, 2104−2105.
(11) Xu, X. P.; Case, D. A. *J. Biomol. NMR* **2001**, *21*, 321−333.
(12) Nilges, M. *Proteins* **1993**, *17*, 297−309.
(13) Nilges, M.; Donoghue, S. I. *Prog. Nucl. Magn. Reson. Spectrosc.* **1998**, *32*, 107−139.
(14) Garrett, D. S.; Seok, Y. J.; Peterkofsky, A.; Gronenborn, A. M.; Clore, G. M. *Nat. Struct. Biol.* **1999**, *6*, 166−173.
(15) Wang, G.; Louis, J. M.; Sondej, M.; Seok, Y. J.; Peterkofsky, A.; Clore, G. M. *EMBO J.* **2000**, *19*, 5635−5649.
(16) Liao, D. I.; Silverton, E.; Seok, Y. J.; Lee, B. R.; Peterkofsky, A.; Davies, D. R. *Structure* **1996**, *4*, 861−872.
(17) Jia, Z.; Quail, J. W.; Waygood, E. B.; Delbaere, L. T. *J. Biol. Chem.* **1993**, *268*, 22490−22501.
(18) van Nuland, N. A.; Hangyi, I. W.; van Schaik, R. C.; Berendsen, H. J.; van Gunsteren, W. F.; Scheek, R. M.; Robillard, G. T. *J. Mol. Biol.* **1994**, *237*, 544−559.
(19) Worthylake, D.; Meadow, N. D.; Roseman, S.; Liao, D. I.; Herzberg, O.; Remington, S. J. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 10382−10386.
(20) van Nuland, N. A.; Boelens, R.; Scheek, R. M.; Robillard, G. T. *J. Mol. Biol.* **1995**, *246*, 180−193.
(21) Garrett, D. S.; Seok, Y. J.; Peterkofsky, A.; Clore, G. M.; Gronenborn, A. M. *Biochemistry* **1997**, *36*, 4393−4398.
(22) Chen, Y.; Reizer, J.; Saier, M. H., Jr.; Fairbrother, W. J.; Wright, P. E. *Biochemistry* **1993**, *32*, 32−37.
(23) Kwong, P. D.; Wyatt, R.; Robinson, J.; Sweet, R. W.; Sodroski, J.; Hendrickson, W. A. *Nature* **1998**, *393*, 648−659.
(24) Kwong, P. D.; Wyatt, R.; Majeed, S.; Robinson, J.; Sweet, R. W.; Sodroski, J.; Hendrickson, W. A. *Structure Fold. Des.* **2000**, *8*, 1329−1339.
(25) Olshevsky, U.; Helseth, E.; Furman, C.; Li, J.; Haseltine, W.; Sodroski, J. *J. Virol.* **1990**, *64*, 5701−5707.
(26) Moebius, U.; Clayton, L. K.; Abraham, S.; Harrison, S. C.; Reinherz, E. L. *J. Exp. Med.* **1992**, *176*, 507−517.
(27) Hubbard, S. J.; Thornton, J. M. *NACCESS*; Department of Biochemistry and Molecular Biology, University College London, 1993.

An AIR is defined as an ambiguous intermolecular distance (d_{iAB}) with a maximum value of 3 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:

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

where N_{atoms} indicates all atoms of a given residue and N_{res} the sum of active and passive residues for a given protein. In this way, the passive residues do not have direct AIRs to the partner protein but can satisfy the partner protein active restraints. A $1/r^6$ sum averaging is used, not by analogy to NOE restraints, but because this mimics 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 3 Å limit represents a compromise between hydrogen−hydrogen and heavy atom−heavy atom minimum van der Waals distances. The use of ambiguous interaction restraints allows HADDOCK to search through all the possible configurations around the interacting site defined by the biochemical and/or biophysical data such as NMR chemical shift perturbation data or mutagenesis data and to find the most favorable pair of interacting amino acids among the active and passive residues.

Docking Protocol. Our HADDOCK (high ambiguity driven protein−protein docking) has been implemented in CNS[28] for structure calculations and makes use of python scripts derived from ARIA[29] for automation (see Material and Methods). The docking protocol, which requires the PDB files of the free proteins and ambiguous interaction restraints, consists of three stages: (i) randomization of orientations and rigid body energy minimization (EM), (ii) semirigid simulated annealing in torsion angle space (TAD-SA), and (iii) final refinement in Cartesian space with explicit solvent.

The three stages are detailed in the Material and Methods section. During the TAD simulated annealing and the water refinement, the amino acids at the interface (side chains and backbone) are allowed to move to optimize the interface packing. The interface amino acids allowed to move are defined by the active and passive amino acids used in the AIRs ± 2 sequential amino acids. Although no real significant structural changes occur during the water refinement stage, it is useful for the improvement of the energetics of the interface. This is important for a proper scoring of the resulting conformations. The final structures are clustered using the pairwise backbone RMSD at the interface and analyzed according to their average interaction energies (sum of E_{elec}, E_{vdw}, E_{AIR}) and their average buried surface area. The entire docking procedure is performed automatically by HADDOCK and is followed by the cluster analysis (for more details, see Material and Methods). For the EIN−HPr complex (247 and 85 amino acids, 25 AIRs requiring 105 000 distance evaluations), the entire run required 2 days on 10 1.3 GHz AMD processors. The three docking stages required 10 s, 1.5 h, and 1 h per structure for the rigid body minimization, the semirigid TAD-SA, and the final water refinement, respectively.

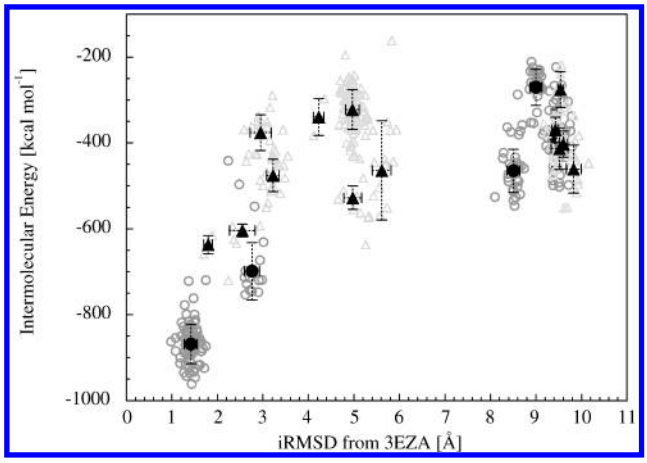

Figure 1. Intermolecular energies versus iRMSDs for the EIN−HPr complex. The energies are calculated as the sum of $E_{elec} + E_{vdw} + E_{AIR}$ after water refinement. iRMSD corresponds to the backbone RMSD at the interface from the pdb structure (3EZA). (Open circles) Single conformations (200) and (filled circles) cluster averages when starting from the complex conformation. (Open triangles) Single conformations (200) and (filled triangles) cluster averages when starting from the free form structures.

Validation of the HADDOCK Approach. HADDOCK was tested for three protein−protein complexes using chemical shift perturbation data in two cases and mutagenesis data in the third to define the ambiguous interaction restraints. As a first test, we performed the docking with ambiguous interaction restraints on the EIN−HPr complex,[14] starting from the structure of the complex. The coordinates of the two proteins in the structure of the complex were separated into two distinct pdb files. Although the structures of the two proteins and in particular of their interface were already in the geometry of the complex, the side chains and backbone atoms at the interface were still allowed to move during the TAD simulated annealing and the water refinement process. On the basis of the NMR titration data,[20,21] 24 amino acids of EIN and 19 amino acids of HPr showing significant chemical shift perturbation were first identified. The solvent accessibility of these amino acids was calculated, and only those that are exposed at the surface of the protein were further selected for the active ambiguous interaction restraints. At the end, 16 amino acids of EIN (E67, E68, K69, A71, I72, D82, E83, E84, G110, Q111, S113, A114, E116, E117, L118, and Y122) and 9 amino acids of HPr (H15, T16, R17, Q21, K24, K49, Q51, T52, and G54) were used as active AIRs. By displaying these amino acids on the free form structures, we defined five passive amino acids for EIN (M78, L79, L115, L123 and R126) and three for HPr (A20, L47and F48) (see Supporting Information). The interface residues that were allowed to move during the TAD simulated annealing and the water refinement process consisted of residues 65 to 74, 76 to 86 and 108 to 128 for EIN and 13 to 26 and 45 to 56 for HPr. Figure 1 (circles) shows the intermolecular energy as a function of the iRMSD (backbone RMSD at the interface) from the target, that is, the NMR structure, for the 200 calculated structures after water refinement. Five clusters were obtained. Their average intermolecular energies are, respectively, −868, −698, −465, −270, and −388 kcal mol^{-1}, and the average iRMSDs from the target are 1.4, 2.7, 8.5, 9.0, and 9.5 Å. For

(28) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54* (5), 905−921.
(29) Linge, J. P.; O'Donoghue, S. I.; Nilges, M. *Methods Enzymol.* **2001**, *339*, 71−90.

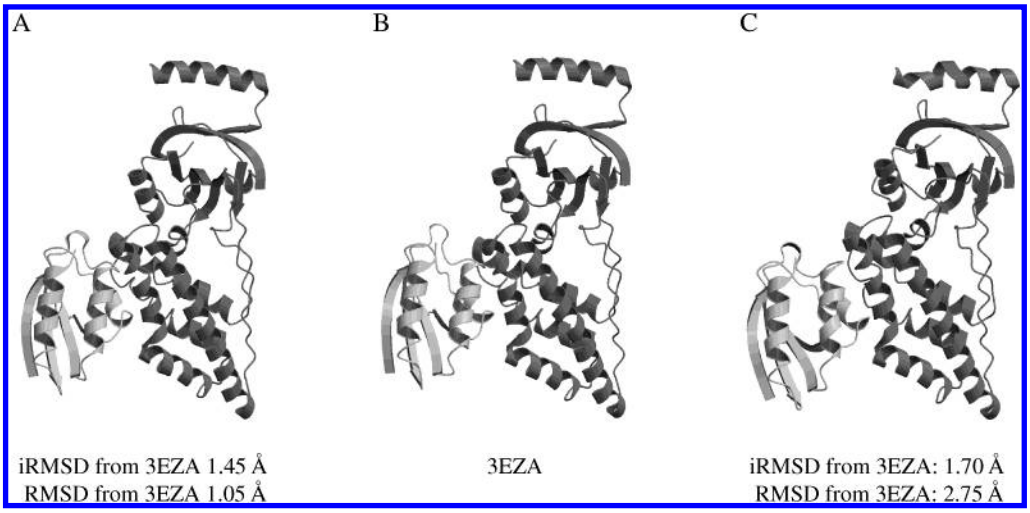

Figure 2. Comparison of the EIN−HPr solutions generated by HADDOCK with the reference structure. (A) Best solution of lowest energy cluster when starting from the structures of the complex. (B) Reference structure (PDB/3EZA). (C) Best solution of lowest energy cluster when starting from the free form structures. iRMSD corresponds to the backbone RMSD at the interface from the reference structure. These figures have been generated with the programs Molscript[35] and Raster3D.[36] HPr is represented in light gray.

reference, the published NMR structure that has however not been optimized within our chosen force field and parameters has an intermolecular energy of −370 kcal mol^{-1}. Cluster 1 has the lowest intermolecular energy as well as the lowest iRMSD from the target. This result demonstrates a nice correlation between the intermolecular energy of our solutions and the iRMSD between these solutions and the target. The best solution of Cluster 1 (the lowest in energy) has an intermolecular energy of −961 kcal mol^{-1} and an iRMSD of 1.45 Å (the backbone RMSD on both proteins is 1.05 Å) from the reference structure (Figure 2A).

Next, HADDOCK was run, starting from the protein structures in the free form.[16,17] The backbone iRMSDs between the free and bound form of EIN and HPr are 0.95 and 0.55 Å, respectively. The resulting intermolecular energies as a function of the iRMSD from the target for the 200 calculated structures after water refinement are shown in Figure 1 (triangles). After analysis, 13 clusters were obtained with average energies between −637 and −275 kcal mol^{-1} and average iRMSDs from the target between 1.80 and 9.85 Å. Again, in this case, the lowest intermolecular energy cluster corresponds to the lowest iRMSD from the target. The best solution of this cluster has an intermolecular energy of −658 kcal mol^{-1} and an iRMSD from the target of 1.70 Å (the backbone RMSD on both proteins is 2.75 Å) (Figure 2C). These results demonstrate that HADDOCK could generate the correct docking solution starting from the free form protein structures and that, again, the lowest intermolecular energy cluster is the closest one to the published NMR structure. Among all protein−protein complexes available in the PDB, the average buried interface area is 1600 ± 400 Å^2.[30] In our case, the best solutions have a buried interface area of 2064 Å^2 when starting from the complex form and 1798 Å^2 when starting from the free form proteins, while the buried surface area of the NMR structure of the complex is 1996 Å^2.

As a second test, the structure of the E2A−HPr complex[15] starting from the free form protein structures[17,19] was calculated with HADDOCK. The backbone iRMSDs between the free and bound form of E2A and HPr are 0.35 and 0.05 Å, respectively.

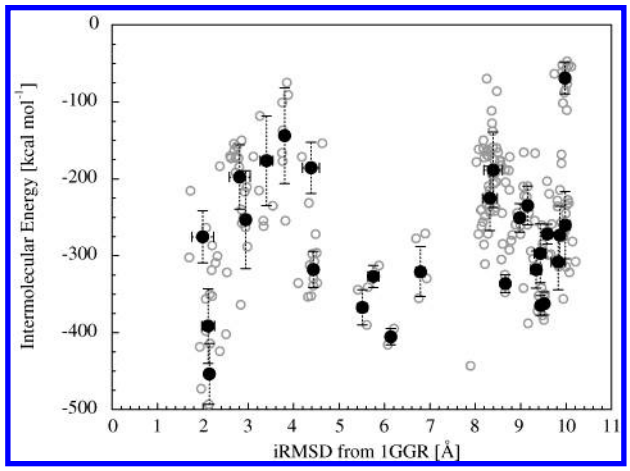

Figure 3. Intermolecular energies versus iRMSDs for the E2A−HPr complex. Energies and iRMSDs as defined in Figure 1. The pdb code of the reference structure is 1GGR. (Open circles) Single conformations (200) and (filled circles) cluster averages when starting from the free form structures.

The intermolecular energy of the complex is −207 kcal mol^{-1}, and its buried surface area is 1434 Å^2. We defined the AIRs as previously described, selecting 11 active (D38, V40, I45, V46, K69, F71, S78, E80, D94, V96 and S141) and 4 passive amino acids (V39, G68, E72 and E86) for E2A and 9 active (H15, T16, R17, A20, F48, Q51, T52, G54 and T56) and 1 passive (N12) amino acids for HPr (see Supporting Information). The flexible interface consisted of amino acids 36 to 48, 66 to 82, 84 to 88, 92 to 98, and 139 to 143 for E2A and 10 to 22 and 46 to 58 for HPr. The resulting intermolecular energies as a function of the iRMSD from target for the 200 calculated structures after water refinement are shown in Figure 3. Clusters (27) were obtained with average intermolecular energies between −453 and −69 kcal mol^{-1} and average iRMSDs from the published structure between 2.0 and 9.9 Å. Again, the lowest energy cluster is the one that is the closest to the reference structure. Its best solution has an intermolecular energy of −493 kcal mol^{-1}, an iRMSD from the published structure of 2.10 Å (the backbone RMSD on both proteins is 2.30 Å), and a buried surface area of 1404 Å^2 (Figure 4).

(30) Lo Conte, L.; Chothia, C.; Janin, J. *J. Mol. Biol.* **1999**, *285*, 2177−2198.

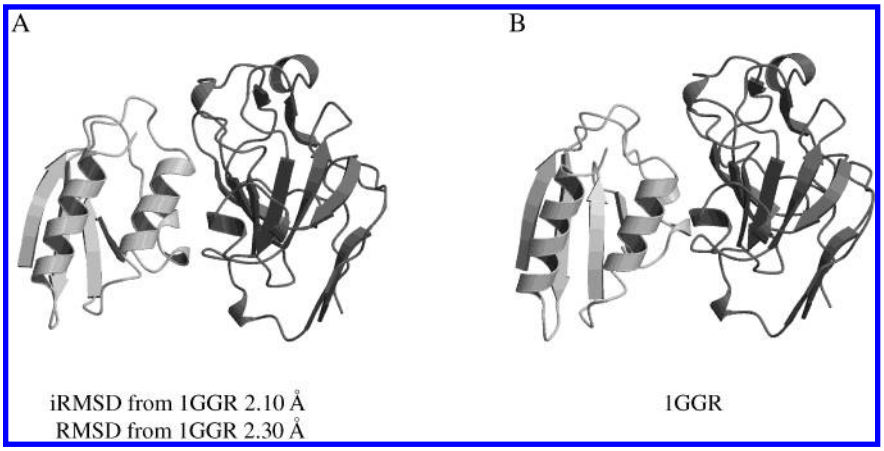

Figure 4. Comparison of the E2A−HPr best HADDOCK solution with the reference structure. (A) Best solution of lowest energy cluster. (B) Reference structure (PDB/1GGR). iRMSD as defined in Figure 2. HPr is represented in light gray.

We finally tested the feasibility of using data from mutagenesis studies to define ambiguous interaction restraints to drive the docking process. For this, docking was performed on the gp120−CD4 complex.[24] The intermolecular energy of the complex solved by crystallography is −283 kcal mol^{-1} and the buried surface area is 1990 Å^2. To speed-up the calculation, the C-terminus domain of CD4 that does not interact with gp120 was removed and only residues 90 to 492 of gp120 and residues 1 to 97 of CD4 were used. The separated forms of the complex were used as the starting point. Mutagenesis data have revealed that residues D368, E370, W427, and D457 of gp120 and residues K29, K35, F43, L44, K46, G47, and R59 of CD4 were important for the binding.[25,26] These amino acids have been used as active residues in the AIRs. In addition, 19 amino acids for gp120 (I109, N280, A281, K362, S365, G367, I371, N425, K429, V430, T455, G459, I467, R469, G471, G472, G473, D474, and R476) and 10 amino acids for CD4 (H27, Q33, I34, Q40, S42, T45, P48, N52, D53, and D56) were selected as passive residues. The flexible interface consisted of amino acids 107 to 111, 278 to 283, 360 to 373, 423 to 432, 453 to 461, and 465 to 478 for gp120 and 26 to 62 for CD4. The resulting intermolecular energies as a function of the iRMSD from target for the 200 calculated structures after water refinement are shown in Figure 5. Clusters (8) were obtained with average intermolecular energies between −407 and −139 kcal·mol^{-1} and average backbone iRMSDs from the target between 0.9 and 11.5 Å. Again, a nice correlation between the intermolecular energy and the iRMSD from the target for the clusters is observed. The best solution from the lowest energy cluster has an intermolecular energy of −445 kcal mol^{-1}, an iRMSD from the published structure of 0.80 Å (the backbone RMSD on both proteins is 0.80 Å), and a buried surface area of 2148 Å^2 (Figure 6). These results nicely demonstrate that biochemical interaction data such as mutagenesis data can also be used to define highly ambiguous restraints to drive the docking with HADDOCK.

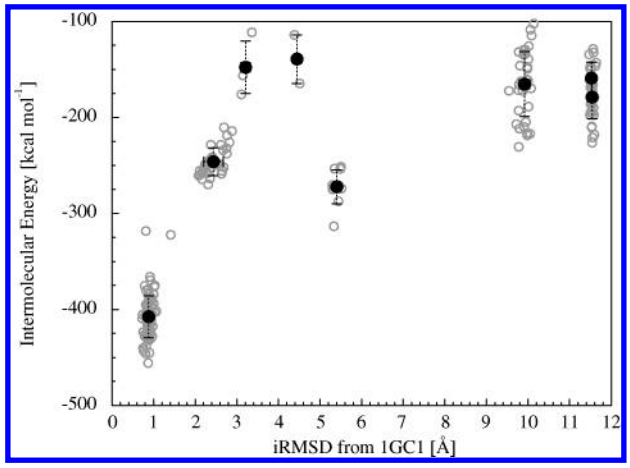

Figure 5. Intermolecular energies versus iRMSD for the gp120−CD4 complex. Energies and iRMSDs as defined in Figure 1. The pdb code of the reference structure is 1GC1. (Open circles) Single conformations (200) and (filled circles) cluster averages when starting from the complex conformation.

Discussion

We have developed an approach, HADDOCK, that allows rapid and accurate docking of protein complexes based on the use of biochemical or biophysical information. This information, which is introduced as ambiguous interaction restraints, is sufficient to drive the docking process. It is important to note that to reduce considerably the ambiguity, information about the interfaces of both proteins is needed. On the basis of the intermolecular energy, the lowest energy clusters generated by HADDOCK were in all cases the closest to the published structure. The fact that the side chains at the interface are allowed to move increases the accuracy of our scoring compared with classical rigid body docking. Indeed, simulated annealing and water refinement do not improve much the iRMSD from the target, but by allowing the side chains to reorient and adopt better conformations, a better scoring of the solutions (a good correlation between the intermolecular energy and the iRMSD from the target) is obtained. The AIR restraints that we have used in the three examples contain, in principle, no information on the relative orientation of the two partners in the complex. Indeed, 180° rotated solutions are obtained that have quite low intermolecular energies (see for example Figure 3). The discrimination between orientations must therefore come mainly from the van der Waals and electrostatic energy terms. This is made possible because of some degree of asymmetry at the interface both in shape complementarities and in the distribution of hydrophobic and hydrophilic residues. One should thus realize that a correct scoring of solutions will depend on the nature of the interface and that, without additional experimental information, the scoring might not be as effective in the case of complexes lacking some kind of asymmetry in their interface.

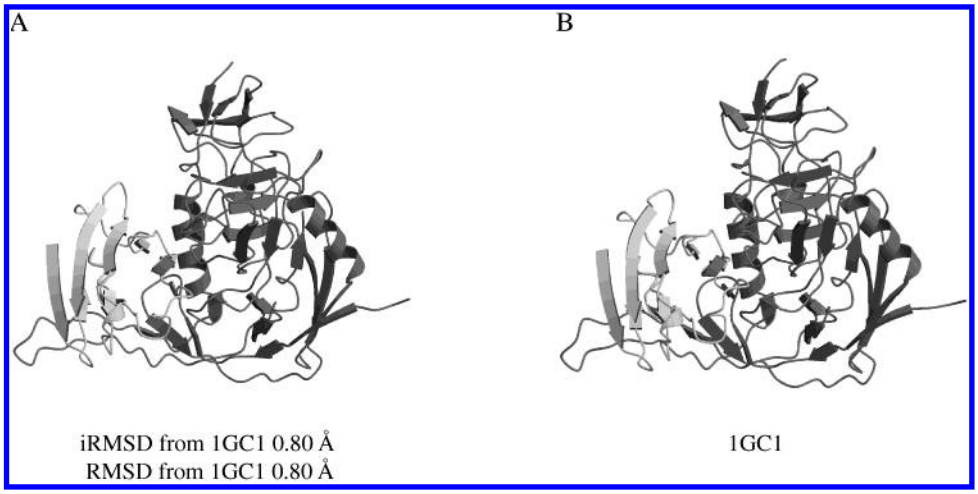

Figure 6. Comparison of the gp120−CD4 best HADDOCK solution with the reference structure. (A) Best solution of lowest energy cluster. (B) Reference structure (PDB/1GC1). iRMSD as defined in Figure 2. CD4 is represented in light gray.

The power of our approach has been demonstrated using chemical shift perturbation data and mutagenesis data, but any kind of data that provides information on the interaction interface could in principle be used to drive the docking and to improve the validity of the solutions. This could be additional NMR restraint such as intermolecular NOEs, RDCs, but also, other types of biochemical or biophysical interaction data could be considered. In this work, the ambiguous interaction restraints (AIRs) were defined with a conservative fixed distance of 3.0 Å. This value could be optimized by differentiating the strength of restraints, depending on a scaling of the distance as a function of the chemical shift perturbation in hertz and/or the type of amino acid. Though this may provide a more accurate and precise scoring, our results show that meaningful structures are already produced with simple and conservative restraints, demonstrating the robustness of our approach. It is also clear in the case of chemical shift perturbation data that better experimental data can be obtained. By using $^{15}N-^{13}C$ double-labeled proteins, side chain information can be collected that will allow a more precise definition of the side chains atoms that are implicated in the interaction. This information could be important to refine the ambiguous interaction restraints (AIRs) and thus the accuracy of HADDOCK.

Material and Methods

Structural Coordinates. The coordinates of all proteins in free and bound form were obtained from the protein data bank (PDB).[1] The accession number for the EIN−HPr complex,[14] the free EIN[16], and the free HPr[17] are, respectively, 3EZA, 1ZYM, and 1POH. The accession number of the E2A−HPr complex[15] and the free form of E2A[19] are, respectively, 1GGR and 1F3G. The accession number of the gp120−CD4 complex[24] is 1GC1.

Docking Protocol. Our HADDOCK (high ambiguity driven docking) approach consists of a collection of python scripts derived from ARIA[29] and makes use of CNS[28] for structure calculation. The python scripts take care of setting up the system from the PDB files of the free proteins, of carrying and monitoring the structure calculations, and of sorting and analyzing the docking solutions. Inter- and intramolecular energies are evaluated using full electrostatic and van der Waals energy terms with an 8.5 Å distance cutoff using the OPLS nonbonded parameters[31] from a modified version of the parallhdg5.2.pro parameter file[32] (Marc Williams, University College London, personal communication). The docking protocol consists of three stages: (i) randomization of orientations and rigid body energy minimization (EM), (ii) semirigid simulated annealing in torsion angle space (TAD-SA), and (iii) final refinement in Cartesian space with explicit solvent.

In the randomization stage, the two partner proteins are positioned at 150 Å from each other in space and each protein is randomly rotated around its center of mass. Rigid body EM is then performed: first, four cycles of orientational optimization are performed in which each protein in turn is allowed to rotate to minimize the intermolecular energy function. Then both translations and rotations are allowed, and the two proteins are docked by rigid body EM. Typically 1000 complex conformations are calculated at this stage. The best 200 solutions in terms of intermolecular energies are then refined. The second stage consists of three simulated annealing refinements. In the first simulated annealing (1000 steps from 2000 to 50 K with 8 fs time steps), the two proteins are considered as rigid bodies and their respective orientation is optimized. In the second simulated annealing (4000 steps from 2000 to 50 K with 4 fs time steps), the side chains at the interface are allowed to move. In the third simulated annealing (1000 steps from 500 to 50 K with 2 fs time steps), both side chains and backbone at the interface are allowed to move to allow for some conformational rearrangements. The resulting structures are then subjected to 200 steps of steepest descent EM. The final stage consists of a gentle refinement in an 8 Å shell of TIP3P water molecules.[33] A 2 fs time step is used for the integration of the equation of motions. The system is first heated to 300 K (500 steps at 100, 200, and 300 K) with position restraints (k_{pos} = 5 kcal mol^{-1} A^{-2}) on all atoms except for the flexible side chains at the interface. MD steps (5000) are then performed at 300 K with position restraints only on noninterface heavy atoms (k_{pos} = 1 kcal mol^{-1} A^{-2}). During the final cooling stage (1000 MD steps at 300, 200, and 100 K), the position restraints are limited to backbone atoms outside the interface. The final structures are clustered using the pairwise backbone RMSD at the interface. A cluster is defined as an ensemble of at least two conformations displaying an iRMSD (backbone RMSD at the interface) smaller than 1.0 Å. The resulting clusters are analyzed and ranked according to their average interaction energies (sum of E_{elec}, E_{vdw}, E_{ACS}) and their average buried surface area.

The HADDOCK package will be made available upon request. In a similar manner as the ARIA program,[34] HADDOCK can be set up via

(31) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.
(32) Linge, J. P.; Nilges, M. *J Biomol NMR* **1999**, *13*, 51−59.
(33) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1992**, *79*, 926−935.
(34) Nilges, M.; Macias, M. J.; O'Donoghue, S. I.; Oschkinat, H. *J. Mol. Biol.* **1997**, *269*, 408−422.
(35) Kraulis, P. J. *J. Appl. Crystallogr.* **1991**, *24*, 946−950.
(36) Merrit, E. A.; Murphy, M. E. P. *Acta Crystallogr.* **1994**, *D50*, 869−873.

a Web browser interface that makes it user-friendly. All the parameters that we used in our examples are set up as default parameters but can be modified by the user to possibly optimize the protocols for a particular problem.

Acknowledgment. This work was supported by a grant from the European community (5th Framework program NMRQUAL Contract Number QLG2-CT-2000-01313) and a "Jonge Chemici" grant from The Netherlands Organization for Scientific Research (NWO) to Dr. A. M. J. J. Bonvin. Financial support from the Center for Biomedical Genetics is also acknowledged. The authors are grateful to Shang-Te Hsu for useful discussions.

Supporting Information Available: Sets of ambiguous interaction restraints (AIRs) used in the docking of the three complexes. This material is available free of charge via the Internet at http://pubs.acs.org.

JA026939X