J|A|C|S
ARTICLES

Structure Determination of Protein−Protein Complexes Using NMR Chemical Shifts: Case of an Endonuclease Colicin−Immunity Protein Complex

Rinaldo W. Montalvao,[†,‡] Andrea Cavalli,[†] Xavier Salvatella,[†] Tom L. Blundell,[‡] and Michele Vendruscolo*[,†]

Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K., and Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1GA, U.K.

Received July 8, 2008; E-mail: mv245@cam.ac.uk

Abstract: Nuclear magnetic resonance (NMR) spectroscopy provides a range of powerful techniques for determining the structures and the dynamics of proteins. The high-resolution determination of the structures of protein−protein complexes, however, is still a challenging problem for this approach, since it can normally provide only a limited amount of structural information at protein−protein interfaces. We present here the determination using NMR chemical shifts of the structure (PDB code 2K5X) of the cytotoxic endonuclease domain from bacterial toxin colicin (E9) in complex with its cognate immunity protein (Im9). In order to achieve this result, we introduce the CamDock method, which combines a flexible docking procedure with a refinement that exploits the structural information provided by chemical shifts. The results that we report thus indicate that chemical shifts can be used as structural restraints for the determination of the conformations of protein complexes that are difficult to obtain by more standard NMR approaches.

Introduction

The great majority of cellular processes, such as enzyme catalysis, signal transduction, gene expression regulation, and the immune response, depend on the formation of transient or permanent macromolecular complexes.[1−3] Structural information about these complexes at low to intermediate resolution can be obtained through a range of techniques, including mass spectrometry, cryo-electron microscopy, and small-angle X-ray scattering.[2] When structures at high resolution are instead required, X-ray crystallography is yet unrivaled.[4] There are, however, often cases in which it is possible to crystallize the molecular units individually but not as a complex or in which a crystal can be obtained for a complex but not in a biological relevant conformation. As crystallization is not required for solution nuclear magnetic resonance (NMR) spectroscopy, this method can be exploited in the study of protein−protein interactions in such situations.[5,6] In favorable cases, NOESY spectra can be measured to derive interproton distance restraints.[7] More generally, residual dipolar couplings,[8] especially when coupled with small-angle X-ray scattering measurements,[9,10] paramagnetic resonance enhancement,[11] and pseudocontact shifts,[6] can be used to derive structural information about the relative positions of the molecules comprising the complex. Also, chemical shifts have proved to be very useful, since they can be used to map the position of interfaces by monitoring the perturbations in the chemical shifts of a protein resulting from the addition of an unlabeled interacting partner.[6] In this context, it has been established that chemical shift measurements of microcrystalline samples obtained through solid-state NMR (SSNMR) spectroscopy provide values that in most cases are in good agreement with those that can be calculated from structures determined by X-ray crystallography.[12−15] These results show that SSNMR spectroscopy can provide specific information about crystal contacts, thus enabling new insights to be obtained about their effects with respect to interactions of more immediate biological relevance.

[†] Department of Chemistry, University of Cambridge.
[‡] Department of Biochemistry, University of Cambridge.

(1) Alberts, B. *Cell* **1998**, *92* (3), 291–294.
(2) Robinson, C. V.; Sali, A.; Baumeister, W. *Nature* **2007**, *450*, 973–982.
(3) Russell, R. B.; Alber, F.; Aloy, P.; Davis, F. P.; Korkin, D.; Pichaud, M.; Topf, M.; Sali, A. *Curr. Opin. Struct. Biol.* **2004**, *14* (3), 313–324.
(4) Blundell, T. L.; Johnson, L. *Protein Crystallography*; Academic Press: New York, 1976.
(5) Takeuchi, K.; Wagner, G. *Curr. Opin. Struct. Biol.* **2006**, *16* (1), 109–117.
(6) Zuiderweg, E. R. P. *Biochemistry* **2002**, *41* (1), 1–7.
(7) Garrett, D. S.; Seok, Y. J.; Peterkofsky, A.; Gronenborn, A. M.; Clore, G. M. *Nat. Struct. Biol.* **1999**, *6* (2), 166–173.
(8) Clore, G. M.; Schwieters, C. D. *J. Am. Chem. Soc.* **2003**, *125* (10), 2902–2912.
(9) Grishaev, A.; Tugarinov, V.; Kay, L. E.; Trewhella, J.; Bax, A. *J. Biomol. NMR* **2008**, *40* (2), 95–106.
(10) Grishaev, A.; Wu, J.; Trewhella, J.; Bax, A. *J. Am. Chem. Soc.* **2005**, *127* (47), 16621–16628.
(11) Tang, C.; Iwahara, J.; Clore, G. M. *Nature* **2006**, *444* (7117), 383–386.
(12) Loquet, A.; Laage, S.; Gardiennet, C.; Elena, B.; Emsley, L.; Bockmann, A.; Lesage, A. *J. Am. Chem. Soc.* **2008**, *130* (32), 10625–10632.
(13) Martin, R. W.; Zilm, K. W. *J. Magn. Reson.* **2003**, *165* (1), 162–174.
(14) Schmidt, H. L. F.; Sperling, L. J.; Gao, Y. G.; Wylie, B. J.; Boettcher, J. M.; Wilson, S. R.; Rienstra, C. A. *J. Phys. Chem. B* **2007**, *111* (51), 14362–14369.
(15) Zech, S. G.; Wand, A. J.; McDermott, A. E. *J. Am. Chem. Soc.* **2005**, *127* (24), 8618–8626.

10.1021/ja805258z CCC: $40.75
Published on Web 11/04/2008

Computational methods are also emerging as valuable tools for predicting the structures of protein complexes in the absence of direct experimental information.[2,16] Such developments are particularly timely, as, with the advent of structural genomics initiatives, a large number of putative protein−protein interactions are being identified.[17,18] There are several computer programs that have been developed for high-resolution protein−protein docking, including Hex,[19] ClusPro,[20] DOT,[21] Rosettadock,[22] PyDock,[23] and ZDOCK.[24] The state-of-the-art for protein−protein docking is assessed periodically through the Critical Assessment of Predicted Interactions (CAPRI).[25]

The combination of the use of experimental techniques and *ab initio* methods is developing into promising approaches for determining the structures of protein complexes.[2,26] When the structures of the proteins in their free states are known, it is possible to use chemical shifts in terms of ambiguous distance restraints in combination with electrostatic and van der Waals energy terms[27] to obtain the structures of the complexes, at least in cases where the structural rearrangement is limited and the chemical shift changes are localized in specific regions. HADDOCK[27] is a program capable of using chemical shift perturbation data from NMR tritation experiments to drive the docking process. The performance of HADDOCK during the CAPRI experiment confirms that the inclusion of biochemical and physical information represents a powerful approach for protein−protein docking.[28]

In this work we present the determination of the structure of a cytotoxic endonuclease domain from bacterial toxin colicin (E9) in complex with its cognate immunity protein (Im9). Since both E9 and Im9 undergo conformational changes upon binding,[29,30] this case is extremely challenging for computational docking procedures. In addition, this complex has also so far escaped determination by NMR spectroscopy, and the only structures available have been obtained by X-ray crystallography.[29,30] It is still very interesting, however, to investigate the structure by NMR methods, since the complex exhibits significant conformational heterogeneity in solution.[30,31] The structure (PDB code 2K5X) that we have determined in this work is of comparable accuracy to the X-ray one (PDB code 1EMV[30]), thus showing that NMR chemical shifts can provide key information in protein complex determination.

The method that we introduce, CamDock, is based on the recent recognition that the information provided by chemical shifts can be used to determine the structures of globular proteins at high resolution.[32−34] We thus show here that this approach can be extended to the determination of the structures of protein complexes. Since chemical shifts are very sensitive structural probes, they can readily identify the residues involved in interactions between proteins upon complex formation. Rather than translating chemical shifts in terms of ambiguous distance restraints, as done for example in HADDOCK,[27] we directly exploit here their dependence on a range of structural factors including torsion angles, electric field effects, ring currents, and the presence of hydrogen bonding,[35] thus increasing the amount of structural information that can be extracted from them.

The results that we present indicate that, although individual chemical shifts do not provide very detailed structural information about the geometry of the interfaces and about the relative orientations of the proteins, the simultaneous inclusion of a large number of chemical shifts in a protein docking protocol is capable of providing, at least in favorable cases, the correct conformations of protein−protein complexes.

Methods

CamDock. In its current implementation, CamDock starts from the structures in the free state of the two proteins to be docked. After a preprocessing step (see below), in which missing atoms in the initial structures are reconstructed, the CamDock procedure consists of two phases, the *ab initio* generation of candidate structures by the Chord program (see below) and the refinement through chemical shifts by the Cheshire program[32] (see below).

The use of chemical shifts in CamDock is different from that in HADDOCK,[27] since we do not transform the information provided by chemical shifts into ambiguous interaction restraints. Instead, we use the program SHIFTX[35] to calculate the chemical shifts corresponding to a given structure by considering a phenomenological approximation of the secondary and tertiary interactions contributing to the chemical shifts, including dihedral angles, ring current shifts, hydrogen bonding, and electric fields.

Preprocessing. The first step to produce the candidate structures is to process the structures of the binding partners in order to identify and rebuild missing atoms and residues. Particular attention is paid to the reconstruction of surface side chains, as they often play a critical role in the docking process; the ANDANTE program,[36] based on the "Penultimate Rotamer Library",[37] is used to find suitable side chain conformations.

(16) 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. *Science* **2004**, *303* (5666), 2026–2029.
(17) Gavin, A. C.; et al. *Nature* **2006**, *440* (7084), 631–636.
(18) Tarassov, K.; Messier, V.; Landry, C. R.; Radinovic, S.; Molina, M. M. S.; Shames, I.; Malitskaya, Y.; Vogel, J.; Bussey, H.; Michnick, S. W. *Science* **2008**, *320* (5882), 1465–1470.
(19) Ritchie, D. W.; Kemp, G. J. L. *Proteins* **2000**, *39* (2), 178–194.
(20) Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J. *Nucleic Acids Res.* **2004**, *32*, W96–W99.
(21) Mandell, J. G.; Roberts, V. A.; Pique, M. E.; Kotlovyi, V.; Mitchell, J. C.; Nelson, E.; Tsigelny, I.; Ten Eyck, L. F. *Protein Eng.* **2001**, *14* (2), 105–113.
(22) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. *J. Mol. Biol.* **2003**, *331* (1), 281–299.
(23) Cheng, T. M. K.; Blundell, T. L.; Fernandez-Recio, J. *Proteins* **2007**, *68* (2), 503–515.
(24) Chen, R.; Li, L.; Weng, Z. P. *Proteins* **2003**, *52* (1), 80–87.
(25) Lensink, M. F.; Mendez, R.; Wodak, S. J. *Proteins* **2007**, *69* (4), 704–718.
(26) Bonvin, A.; Boelens, R.; Kaptein, R. *Curr. Opin. Chem. Biol.* **2005**, *9* (5), 501–508.
(27) Dominguez, C.; Boelens, R.; Bonvin, A. *J. Am. Chem. Soc.* **2003**, *125* (7), 1731–1737.
(28) De Vries, S. J.; van Dijk, A. D. J.; Krzeminski, M.; van Dijk, M.; Thureau, A.; Hsu, V.; Wassenaar, T.; Bonvin, A. *Proteins* **2007**, *69* (4), 726–733.
(29) Kleanthous, C.; Kuhlmann, U. C.; Pommer, A. J.; Ferguson, N.; Radford, S. E.; Moore, G. R.; James, R.; Hemmings, A. M. *Nat. Struct. Biol.* **1999**, *6* (3), 243–252.
(30) Kuhlmann, U. C.; Pommer, A. J.; Moore, G. R.; James, R.; Kleanthous, C. *J. Mol. Biol.* **2000**, *301* (5), 1163–1178.
(31) Whittaker, S. B. M.; Czisch, M.; Wechselberger, R.; Kaptein, R.; Hemmings, A. M.; James, R.; Kleanthous, C.; Moore, G. R. *Protein Sci.* **2000**, *9* (4), 713–720.
(32) Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, (104), 9615–9620.
(33) Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G. H.; Eletsky, A.; Wu, Y. B.; Singarapu, K. K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C. H.; Szyperski, T.; Montelione, G. T.; Baker, D.; Bax, A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (12), 4685–4690.
(34) Wishart, D. S.; Arndt, D.; Berjanskii, M.; Tang, P.; Zhou, J.; Lin, G. *Nucleic Acids Res.* **2008**, *36*, W496–W502.
(35) Neal, S.; Nip, A. M.; Zhang, H. Y.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26* (3), 215–240.
(36) Smith, R. E.; Lovell, S. C.; Burke, D. F.; Montalvao, R. W.; Blundell, T. L. *Bioinformatics* **2007**, *23* (9), 1099–1105.
(37) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. *Proteins* **2000**, *40* (3), 389–408.

Chord. Chord, the computational method that we introduce here to produce candidate structures for protein−protein complexes, is based on the approach used by the HEX program,[19] which employs a spherical harmonics description of the protein surface. The use of spherical harmonics has several advantages over grid-based FFT docking correlation methods.[19] Most notably, rotations and translations can be carried out by operating on the initial expansion coefficients. This procedure not only results in a very quick search of the conformational space of the interacting partners but also allows control of the degree of resolution of the shapes. Additionally, it is possible to focus the search around specific regions in the receptor surface around a suspected docking site. Putative docking sites may be determined from experimental data or predicted from programs such as CRESCENDO.[38] These latter programs can suggest possible locations for the binding site by comparing the observed amino acid substitution patterns in the protein family with those directly predicted from the local structural environment.

In Chord, the functions describing receptor and ligand surfaces, $A(\mathbf{r})$, are described in terms of spherical harmonics as

$$A(\mathbf{r}) = \sum_{nim}^{N} a_{nim} R_{nl}(r)\, Y_m^l(\theta, \phi) \qquad N \geq n > l \geq |m| \geq 0 \tag{1}$$

where the position vectors $\mathbf{r}$ are represented in spherical coordinates, a_{nlm} are the expansion coefficients

$$a_{nim} = \int A(\mathbf{r})\, R_{nl}(r)\, Y_m^l(\theta, \phi)\, \mathrm{d}V \tag{2}$$

R_{nl} is the radial function, Y_m^l are real spherical harmonics, and N is the order of the expansion. Ritchie and Kemp have demonstrated that the protein surface or the electrostatic potential associated with it can be represented by carefully choosing the radial function R_{nl}.[19] For surface shape representation, the radial function assumes the form

$$R_{nl}(r) = \left[\left(\frac{2}{k^{3/2}}\right)\frac{(n-l-1)!}{\Gamma(n+1/2)}\right]^{1/2} \exp(-\rho/2)\rho^{l/2} L_{n-l-1}^{(l+1/2)} \tag{3}$$

where $\rho = r^2/k$ and $k = 20$. In eq 3, the square root term is a normalization factor, ρ is a scaled distance, and k is its scaling parameter.

The expansion coefficients can be calculated through eq 2, and an expansion of order N will generate $N(N + 1)(2N + 1)/6$ coefficients. As mentioned before, any translation or rotation needed for exploring the docking space is realized by simply rotating those coefficients.[19]

In Chord, the production of candidate structures is conducted in two distinct phases. During the first, an expansion of order $N = 16$ is used to obtain a low-resolution map of the protein surface. This relatively coarse-grained representation has the advantage of enabling a very fast search having just 1496 coefficients and the respective rotation matrices. In this phase, 10^9 candidate structures are produced and processed in a few minutes using a modern computer. As a consequence of the low-resolution of the spherical harmonics approximation for the surface of the proteins, however, steric clashes between the surface of the receptor and ligand are a common occurrence. In order to remove these candidate structures, Chord employs a surface correlation scoring function

$$E_{sc} = \omega \Delta_{SP} \tag{4}$$

where $\omega = 1.6$ is a weight and Δ_{SP} is the difference between the surface properties (derived from the differential geometry of the surface) of the two surfaces in contact for a particular candidate structure. The filtering process eliminates most of the unsuitable structure, and only the top 5000 candidate structures are moved to the next phase. Phase two makes use of an expansion of order $N = 32$ with 14440 coefficients and the corresponding rotation matrices. In this phase the search is refined by focusing around the regions determined in phase one by using a more refined surface representation and small steps in the angular and translational moves. It also employs a more sophisticated scoring function that combines the surface correlation function with a cavitation free energy,[39] which is also very fast to calculate, as it is a simple linear function of the molecular surface area

$$\Delta G_{cav} = \gamma_{MS}(A_{MS}^{complex} - A_{MS}^{receptor} - A_{MS}^{ligand}) \tag{5}$$

where $\gamma_{MS} = 0.069$ is a constant and A_{MS} are the molecular surface areas. The top 500 models are selected for the refinement step.

The overall Chord score is given by

$$E_{Chord} = E_{sc} + \Delta G_{cav} \tag{6}$$

The computer code for both phases can be straightforwardly parallelized, thus improving the performance. In addition, the small search space is also suitable for the inclusion of a scoring function based on experimental data or statistically derived from databases of protein−protein interfaces.

Chord for Flexible Docking. For complexes in which the component proteins undergo significant conformational changes upon docking, the Chord algorithm includes a description of the structural flexibility during the docking procedure. Instead of using a single structure for a protein, we consider an ensemble of conformations representing the possible variations of the structure upon docking. The construction of the ensemble is implemented by rebuilding the flexible regions using the program RAPPER,[40] creating an ensemble of structures that is then used for docking. Three options are currently included for identifying the flexibile regions: (1) the CamP method,[41] which identifies regions of low local stability, (2) the analysis of the regions exhibiting conformational variability in the family of homologous proteins,[42] (3) or the prediction of the protection factors from hydrogen exchange, which is carried out from the knowledge of the structure.[43] Typically, this procedure produces an initial ensemble of hundreds of putative structures, although this number could be increased for rebuilding highly flexible structures. In addition, the inclusion of side-chain remodeling in the putative protein−protein interfaces can also increase the number of structures present in the initial ensemble.

Chord produces a spherical harmonic representation for each model in the initial ensemble and applies the procedure described previously (see the section "Chord" above) to each one of them. Thus, while in the case of rigid docking only one structure is considered, in flexible docking all of the structures in the initial ensemble are analyzed. This procedure allows the CamDock program to access a relatively broad range of conformations and orientations for the proteins to be bound. By exploring an ensemble of structures representing the flexibility of the protein in the free state, we obtain a model for its structure in the bound state, at least in the cases in which an equilibrium shift mechanism applies,[44−46] or otherwise a better starting point for the subsequent refinement

(38) Chelliah, V.; Blundell, T. L.; Fernandez-Recio, J. *J. Mol. Biol.* **2006**, *357* (5), 1669–1682.
(39) Jackson, R. M.; Sternberg, M. J. E. *J. Mol. Biol.* **1995**, *250* (2), 258–275.
(40) DePristo, M. A.; de Bakker, P. I. W.; Lovell, S. C.; Blundell, T. L. *Proteins* **2003**, *51* (1), 41–55.
(41) Tartaglia, G. G.; Cavalli, A.; Vendruscolo, M. *Structure* **2007**, *15*, 139–143.
(42) Best, R. B.; Lindorff-Larsen, K.; DePristo, M. A.; Vendruscolo, M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (29), 10901–10906.
(43) Best, R. B.; Vendruscolo, M. *Structure* **2006**, *14* (1), 97–106.
(44) Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E. *Science* **2006**, *313* (5793), 1638–1642.
(45) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky, J. J.; Kay, L. E.; Kern, D. *Nature* **2005**, *438* (7064), 117–121.

of the structure of the complex. In the case of flexible docking, Chord produces by default 500 models for the complex for each conformation in the initial ensemble, thus generating a total of 50.000 models. Only the top 500 models are passed forward to the next phase; we have verified that this procedure reduces the computational cost in the chemical shift refinement without compromising the quality of the final results.

Chemical Shift Refinement. As explained above, Chord produces a set of possible solutions for the structure of the complex using a conformational sampling that takes into account both the flexibility of the component proteins and their relative arrangement. In the next step of the CamDock procedure, chemical shifts and molecular simulations are combined to obtain the final solution for the complex among all candidate structures. This step is analogous to the structural refinement driven by chemical shift information used in the Cheshire procedure.[32]

The CamDock approach is based on the observation that when a candidate structure is correct, the differences between the observed and the predicted chemical shifts will be minimal and their correlation will be maximal. In an ensemble, the candidate structure with the largest correlation should be the one with its configuration closest to the actual protein complex. In order to produce the energy function for refinement, Chord calculates the chemical shifts for the candidate structures by using PROSHIFT,[47] SPARTA,[48] or SHIFTX.[35] The correlations, r, between the chemical shifts of the candidate structures ("predicted") and of the complex ("observed") are then calculated for the N, Cα, Cβ, and Hα atoms, and the total correlation is defined as[32]

$$C = k_{H\alpha}(1 - r_{H\alpha}) + k_N(1 - r_N) + k_{C\alpha}(1 - r_{C\alpha}) + k_{C\beta}(1 - r_{C\beta}) \quad (7)$$

where the values for the weight constants are $k_{H\alpha} = 75$ and $k_N = k_{C\alpha} = k_{C\beta} = 25$. The chemical shift correlation C is capped at 15 to avoid results where the correlations between the calculated chemical shifts of the candidate structures and the experimental chemical shift of the complex are better than the error of the programs.

The ranking of the structures is done by defining a chemical-shift-based energy as a combination of a physicochemical term[32]

$$E_{FF} = E_{vdW} + E_{elec} + E_{EEF1} + E_{PMF} + E_{hb} \quad (8)$$

which includes contributions from van der Waals (E_{vdW}), electrostatic (E_{elec}), solvation (E_{EEF1}), potential of mean force (E_{PMF}), and hydrogen bonding interactions (E_{hb}), and the chemical shift total correlation[32] as

$$E_{Cheshire} = E_{FF} + \alpha(C_1 + C_2) \quad (9)$$

where $\alpha = 10$, and C_1 and C_2 are the correlations for the two proteins. The assessment of the quality of the solution is done by defining the Z score[32] as

$$Z = \frac{E_{Cheshire} - \mu}{\sigma} \quad (10)$$

where μ and σ are the energy and the standard deviations over all the candidate structures generated for a given protein complex.

Results

E9−Im9 Complex. Proteins in the colicin family are 60 kD α/β endonucleases produced by *E. coli* under stress conditions in order to reduce competition from related strains.[49] In order to neutralize the nuclease activity within the bacteria that produce them, colicins are prevented from binding DNA by forming a complex with cognate immunity proteins, which are smaller than 9.5 kD α-helical proteins.[29,30,49] The complexes are formed from endonuclease colicins and their immunity protein partners that have an extremely high dissociation constant, which is in the region of 10^{-16} M, making these complexes among the tightest known ones. Noncognate immunity proteins, by contrast, have dissociation constants smaller by 6 to 10 orders of magnitude, which provide insufficient specificity to prevent cell death.[50,51]

The study of the interaction between endonuclease colicins and immunity proteins has provided considerable insight into the mechanism of specificity of protein−protein interactions.[29,30,52] One particular aspect of the mechanism of inhibition of the enzymatic activity of endonucleases is that the binding of the immunity proteins does not take place at the active site of the endonucleases, which is often the case for ribonucleases, proteases and kinases, as well as for other endonucleases.[29,30] It has also been realized that the endonuclease colicins−immunity proteins binding involves a "dual recognition" mechanism, in which the conserved region of helix III provides the stability of the complex by binding tightly to the cognate DNase through hydrophobic and hydrogen bond interactions, and the variable region of helix II enables high specificity in the recognition.[53]

CamDock Method. The CamDock method, which is introduced in this work, is based on the combination of two programs, Chord (see Methods), which performs protein−protein docking, and Cheshire,[32] which uses NMR chemical shifts to determine the structures of proteins. CamDock is divided into two phases (see Methods): (1) generation of an ensemble of candidate structures of the complex, through Chord, and (2) refinement of the candidate structures by molecular simulations carried out to minimize an energy function based on chemical shifts, through Cheshire. This approach therefore exploits the complementary information provided by the protein−protein docking scoring functions adopted by Chord and by the chemical-shift-based energy function used by Cheshire.

Determination of the Structure of the E9−Im9 Complex. We applied CamDock to the determination of the structure of the complex formed by the DNase domain of the *E. coli* endonuclease colicin E9 and the immunity protein Im9 (Figures 1 and 2). This complex is very difficult to determine both by standard docking approaches, because there are significant conformational rearrangements upon binding, and by standard NMR methods, since it exhibits significant conformational heterogeneity in solution.[30,31] The overall Cα rmsd between the X-ray structures of E9 DNase in the free (PDB code 1FSJ[30]) and in the bound (PDB code 1EMV[30]) states is 0.96 Å. The Cα rmsd between the NMR structure of Im9 in the free state (PDB code 1IMQ[54]) and the X-ray structure in the bound state is 1.76 Å (Figure 3). In particular, Im9 experiences large conformational changes upon binding in a region corresponding to the interface loop (residues 54−62), as shown by the Cα rmsd per residue, which

(46) Vendruscolo, M.; Dobson, C. M. *Science* **2006**, *313* (5793), 1586–1587.
(47) Meiler, J. *J. Biomol. NMR* **2003**, *26* (1), 25–37.
(48) Shen, Y.; Bax, A. *J. Biomol. NMR* **2007**, *38* (4), 289–302.
(49) Cascales, E.; Buchanan, S. K.; Duche, D.; Kleanthous, C.; Lloubes, R.; Postle, K.; Riley, M.; Slatin, S.; Cavard, D. *Microbiol. Mol. Biol. Rev.* **2007**, *71* (1), 158–229.
(50) Wallis, R.; Leung, K. Y.; Pommer, A. J.; Videler, H.; Moore, G. R.; James, R.; Kleanthous, C. *Biochemistry* **1995**, *34* (42), 13751–13759.
(51) Wallis, R.; Moore, G. R.; James, R.; Kleanthous, C. *Biochemistry* **1995**, *34* (42), 13743–13750.
(52) Kleanthous, C.; Hemmings, A. M.; Moore, G. R.; James, R. *Mol. Microbiol.* **1998**, *28* (2), 227–233.
(53) Wallis, R.; Leung, K. Y.; Osborne, M. J.; James, R.; Moore, G. R.; Kleanthous, C. *Biochemistry* **1998**, *37* (2), 476–485.
(54) Osborne, M. J.; Breeze, A. L.; Lian, L. Y.; Reilly, A.; James, R.; Kleanthous, C.; Moore, G. R. *Biochemistry* **1996**, *35* (29), 9505–9512.

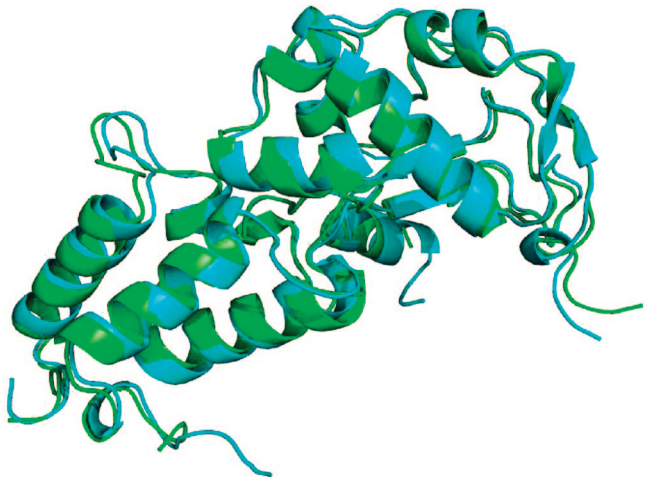

Figure 1. Comparison of the X-ray structure of the E9−Im9 complex (green, PDB code 1EMV) with the CamDock structure (cyan, PDB code 2K5X). The root-mean-square distance (rmsd) is 1.18 Å for Cα and 1.73 Å for all-atom.

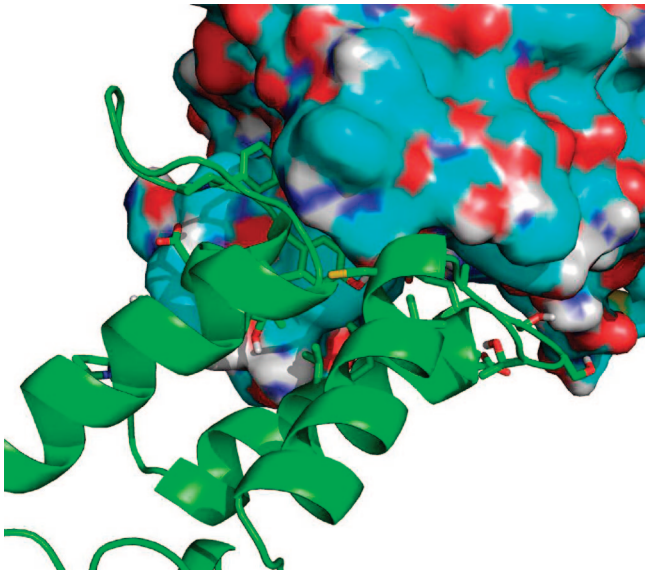

Figure 2. Representation of the structure of the E9−Im9 complex determined in this work using NMR chemical shift information (PDB code 2K5X). The structure of Im9 is shown as a green ribbon diagram, while the structure of E9 is shown as a surface.

extends up to about 5 Å, between the structure of Im9 in the bound and in the free states (Figure 4, blue line).

In order to take into account the conformational changes upon binding that take place in Im9, in the Chord phase of the CamDock procedure, starting from the free structure of Im9 (PDB code 1IMP), we generated an ensemble of conformations, called "seed" structures, to be used in the docking procedure. An example of a seed structure is provided in Figure 3 and analyzed in Figure 4. In this representative example, the seed structure is very similar to the structure in the free state, except in the interface loop (Figure 4, green line). By contrast, the same seed structure is rather dissimilar from the structure in the bound state, except in the interface loop (Figure 4, red line).

The ensemble of seed structures can be obtained by a variety of methods that are capable of accounting for their conformational flexibility (see Methods). In the case of the E9−Im9 complex discussed here, we derived spatial restraints from the 21 NMR structures of the Im9 protein (PDB code IMP); by

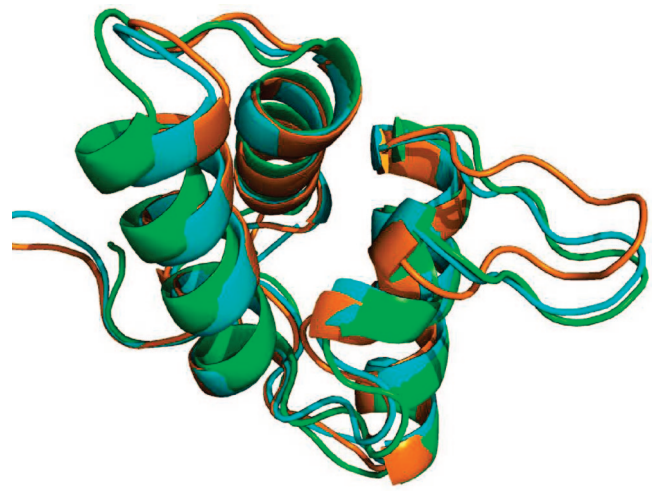

Figure 3. Comparison of the X-ray structure of the Im9 in the bound state with E9 (green, PDB code 1EMV) and in the free state (gold, PDB code 1IMP); the overall Cα rmsd between these two structures is 1.76 Å. One of the candidate structures generated by CamDock from the structure of the free state in order to take into account the flexibility upon docking (the "seed" structure) is shown in gold; the binding loop in this candidate structure is very close to that of the bound state.

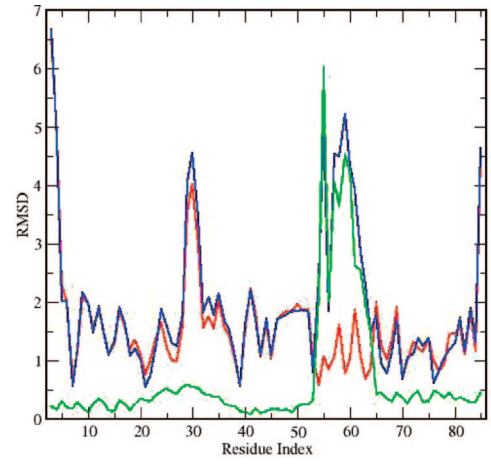

Figure 4. Cα rmsd analysis of the Im9 structures shown in Figure 3. (Blue line) Cα rmsd between the structures in the bound state with E9 (PDB code 1EMV) and in the free state (PDB code 1IMP). (Green line) Cα rmsd between the "seed" structure shown in Figure 3 and the structure in the free state (PDB code 1IMP). (Red line) Cα rmsd between the seed structure of Im9 and the structure in the bound state (PDB code 1EMV).

using the RAPPER procedure (see Methods), such restraints were then used to produce 100 models of the structure of Im9, which were subsequently used as input in the Chord procedure (Figure 5). By using such an ensemble, the Chord generation of candidate structures for the complex offers conformations close (1−5 Å Cα rmsd) to the X-ray structure of the complex (PDB code 1EMV) but also very different structures, with an Cα rmsd from the X-ray structure above 10−15 Å (Figure 6a); the structure of the minimal Chord score, E_{Chord} (eq 6), is at about 4 Å Cα rmsd. We then ranked these same candidate structures by the Cheshire score, E_{Cheshire} (eq 8), obtained by combining the Cheshire force field and the structural information provided by NMR chemical shifts (Figure 6b); we used the chemical shifts of the BMRB entries 4115 and 4352 (see Table 1). The resulting energy landscape is weakly funneled toward the X-ray structure of the complex, thus showing that the

Figure 5. Ensemble of candidate ("seed") conformations of Im9. The seed conformations, which are generated by CamDock from the structure of the free state by taking into account its flexibility, are used as models for the bound state to be used in the docking procedure.

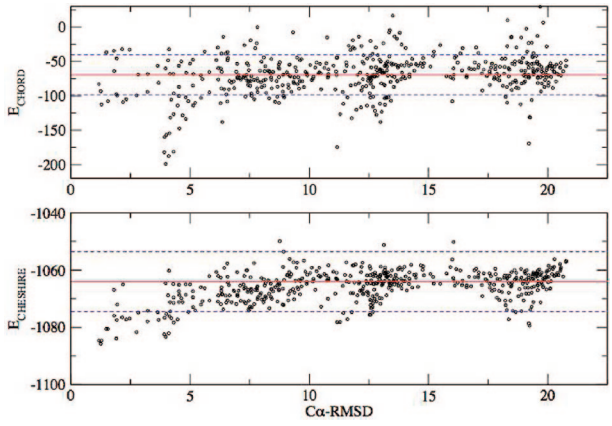

Figure 6. (a) Energy landscape for the E9−Im9 complex for the candidate structures generated by the Chord procedure before their chemical-shift-based refinement; these structures are ranked by a surface-complementarity energy function E_{Chord} (eq 6). (b) Energy landscape for the same candidate structures represented in panel a but as a function of the Cheshire score $E_{Cheshire}$ (eq 9), which combines the surface-complementarity energy function and the chemical shift score; the energy landscape of the $E_{Cheshire}$ score is weakly funneled toward the correct structure.

information provided by chemical shifts is key in identifying the correct conformation of the E9−Im9 complex (Figure 6b).

In the following step of the CamDock procedure, we used the chemical shift information to refine the candidate structures using molecular simulations with chemical shift restraints (see Methods). This procedure enables the determination of a final structure for the complex that has a 1.18 Å Cα rmsd (1.73 Å all-atom rmsd) from the X-ray structure (Figure 1); as a result of the better exploitation of the information provided by the chemical shifts, the corresponding energy landscape is more strongly funneled toward the correct conformation (Figure 7) than that before the refinement (Figure 6b). This type of accuracy in the structure is comparable to that achievable in the chemical-shift-based determination of the structures of the native states of proteins in solution.[32,33] As a further assessment of the quality of the NMR structure determined here, we present in Table 2 the comparison of the interface residues in the X-ray structure (PDB code 1EMV) and in the NMR structure determined here (PDB code 2K5X).

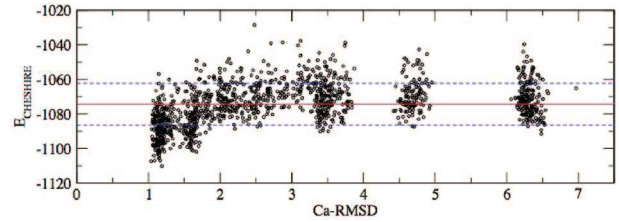

Figure 7. Energy landscape of the Cheshire score $E_{Cheshire}$ (eq 9), for the E9−Im9 complex for the candidate structures generated by the CamDock procedure after their chemical-shift-based refinement.

Table 1. Summary of the Structural Determination Procedure of the E9−Im9 Complex (PDB Code 2K5X)

		E9	Im9
no. of amino acids		134	86
no. of chemical shifts	1Hα	119	85
	15N	122	81
	13Cα	131	86
	13Cβ	95	80
chemical shift correlations	1Hα	0.96	0.83
	15N	0.99	0.99
	13Cα	0.98	0.99
	13Cβ	0.99	0.84

Table 2. Comparison of the Residues at the Interface of the E9−Im9 Complex in the X-ray Structure (PDB Code 1EMV) and in the Structure, Denoted as NMR, Determined Here from Chemical Shifts (PDB code 2K5X)

E9		Im9	
1EMV	2K5X	1EMV	2K5X
	LEU 23	ILE 22	ILE 22
ARG 54		CYS 23	CYS 23
ASN 70	ASN 70	ASN 24	ASN 24
LEU 71	LEU 71	ALA 25	ALA 25
ASN 72	ASN 72	THR 27	THR 27
PRO 73	PRO 73	SER 28	SER 28
SER 74	SER 74	SER 29	SER 29
ASN 75	ASN 75	GLU 30	GLU 30
SER 77	SER 77		GLU 31
SER 78	SER 78	LEU 33	LEU 33
LYS 81	LYS 81	VAL 34	VAL 34
TYR 83	TYR 83	VAL 37	VAL 37
SER 84	SER 84	THR 38	THR 38
PHE 86	PHE 86	GLU 41	GLU 41
THR 87	THR 87	HIS 46	
PRO 88	PRO 88	PRO 47	PRO 47
LYS 89	LYS 89	SER 48	SER 48
ASN 90		GLY 49	
GLN 92	GLN 92	SER 50	SER 50
	GLY 94	ASP 51	ASP 51
GLY 95	GLY 95	ILE 53	ILE 53
LYS 97	LYS 97	TYR 54	TYR 54
VAL 98	VAL 98	TYR 55	TYR 55
TYR 99	TYR 99	PRO 56	PRO 56
		ASP 62	ASP62
		ILE 67	ILE 67

In order to assess the sensitivity of the results on the number of chemical shift restraints used in the calculations, we repeated the calculations by randomly removing about 25% of the chemical shifts, i.e. 200 of the 799 that were available (see Table 1). In this case we found the structure of minimal overall score at 3.47 Å Cα rmsd. Further, in order to assess the relative importance of the Cα chemical shifts, we carried out a separate calculation in which we did not use them as a source of information, thus removing 217 restraints (see Table 1). Since the structure of minimal score was found in this case at 6.25 Å Cα rmsd from the X-ray structure, these results indicate that Cα chemical shifts are a particularly important source of information. In addition, to assess the importance of electrostatic interactions in determining the structure of the complex, we carried out a calculation in which we removed the electrostatic term in the SHIFTX predictions of the chemical shifts. In this case, we found the structure of minimal score at 6.31 Å Cα rmsd from the X-ray structure, a result that suggests that electrostatic interactions are important to stabilize the bound state.

These results indicate that current methods for the determination of protein structures, including those of protein−protein complexes, from NMR chemical shifts are limited by the accuracy to which the chemical shifts themselves can be calculated for a given structure. Further advances in this direction[47,55,35,48] can be expected to lead to an increased accuracy in the resulting structures. In particular, it would be extremely useful to include information about side chain chemical shifts, although at the moment the still rather limited accuracy in their prediction restricts their effective use for structure determination.

Conclusions

We have presented the determination of the structure of the complex formed by an endonuclease colicin (E9) and an immunity protein (Im9) by introducing the CamDock method, which incorporates NMR chemical shifts as structural restraints in a flexible docking approach. This method is not specific to the E9−Im9 complex discussed here but is applicable to other protein−protein complexes for which chemical shift measurements are available. We have shown that this method is particularly suitable in cases in which there are considerable structural modifications upon binding. Thus, the results that we have presented contribute in extending the range of applicability of NMR chemical shifts to the determination of the structures of protein−protein complexes, after the initial reports of their use for the determination of the structures of the native states of globular proteins in solution.[32−34]

Acknowledgment. This work was supported by the Leverhulme Trust (R.W.M., X.S., and M.V.), the European Union (A.C., X.S., and M.V.), the European Molecular Biology Organisation (M.V.), and the Royal Society (M.V.).

JA805258Z

(55) Moon, S.; Case, D. A. *J. Biomol. NMR* **2007**, *38* (2), 139–150.