Journal of Biomolecular NMR, **20:** 15–21, 2001.
KLUWER/ESCOM

Structure prediction of protein complexes by an NMR-based protein docking algorithm

Oliver Kohlbacher[a,*], Andreas Burchardt[a], Andreas Moll[a], Andreas Hildebrandt[a], Peter Bayer[b] & Hans-Peter Lenhof[a]
[a]*Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany*
[b]*Max-Planck-Institut für molekulare Physiologie, Otto-Hahn-Str. 11, D-44227 Dortmund, Germany*

Received 7 November 2000; Accepted 9 February 2001

Key words: NMR shift prediction, protein docking, structure prediction

Abstract

Protein docking algorithms can be used to study the driving forces and reaction mechanisms of docking processes. They are also able to speed up the lengthy process of experimental structure elucidation of protein complexes by proposing potential structures. In this paper, we are discussing a variant of the protein-protein docking problem, where the input consists of the tertiary structures of proteins A and B plus an unassigned one-dimensional ^{1}H-NMR spectrum of the complex AB. We present a new scoring function for evaluating and ranking potential complex structures produced by a docking algorithm. The scoring function computes a ‘theoretical’ ^{1}H-NMR spectrum for each tentative complex structure and subtracts the calculated spectrum from the experimental one. The absolute areas of the difference spectra are then used to rank the potential complex structures. In contrast to formerly published approaches (e.g. [Morelli et al. (2000) *Biochemistry*, **39**, 2530–2537]) we do not use distance constraints (intermolecular NOE constraints). We have tested the approach with four protein complexes whose three-dimensional structures are stored in the PDB data bank (Bernstein et al., 1977) and whose ^{1}H-NMR shift assignments are available from the BMRB database. The best result was obtained for an example, where all standard scoring functions failed completely. Here, our new scoring function achieved an almost perfect separation between good approximations of the true complex structure and false positives.

Introduction

The goal of protein docking research is the development and implementation of algorithms for predicting the structure and the stability of protein complexes. In the last 15 years, different classes of algorithms have been published for the different classes of protein complexes (protein–ligand, protein–peptide, protein–protein, and protein–DNA). Nowadays, docking algorithms are used in the area of drug design to screen large molecular databases for potential inhibitors of a given enzyme. Docking algorithms that yield good approximations of the protein complexes can also accelerate the lengthy process of experimental structure determination (see below) and can help to build hypotheses about the driving forces of docking reactions.

We are studying the so-called Protein–Protein Docking Problem (PPD problem) that can be formulated as follows: Given the 3D structure of two proteins A and B that form a protein complex AB, compute the 3D structure of the complex AB. In this paper, we are discussing a variant of the PPD problem where the input consists of the tertiary structures of A and B plus an unassigned experimental ^{1}H-NMR spectrum of the protein complex AB.

Early approaches for the PPD problem (e.g. Connolly, 1986; Shoichet and Kuntz, 1991; Katchalski-Katzir et al., 1992; Norel et al., 1994; Fisher et al., 1995; Lenhof, 1997), the so-called Rigid-Body-

[*]To whom correspondence should be addressed at: Informatics Research, Celera Genomics, 45 W. Gude Drive, Rockville, MD 20850, U.S.A. E-mail: oliver.kohlbacher@celera.com

Docking algorithms (RBD), were based on the assumption that the proteins *A* and *B* do not change their structure during the docking process (lock-and-key principle). As a matter of fact, there are many examples where significant structural changes do occur (e.g. induced fit). However, a study by Betts and Sternberg (1999) implies that the lock-and-key principle is a suitable model for most protein docking predictions.

RBD algorithms usually compute the potential complex conformations that show a good matching of large chemically complementary surface regions of *A* and *B* and no (or only small) overlap between the interior of *A* and the interior of *B*. The first step of every RBD algorithm is the generation of a huge number of potential complex conformations that will be evaluated with respect to their geometrical and/or chemical properties in the second step. Each RBD algorithm generates a list of potential complex conformations. The candidates in this list are sorted with respect to a geometric or energetic scoring function such that the first element of the list is the conformation with the best score. Summaries of published RBD techniques can be found in the works by Lengauer and Rarey (1996), Meyer et al. (1996), and Sternberg et al. (1998).

The complexity of the protein docking problem increases tremendously if domains or subdomains of the proteins *A* and *B* significantly change their conformations during the docking process. Sandak et al. (1998) have developed algorithms that are able to handle rotations of rigid protein domains or subdomains (hinge bending).

A key problem in protein docking (and in related problems as well) is the accurate prediction of binding free energies. Over the years, a multitude of different theoretical models have been proposed to estimate binding energies (e.g. Jackson and Sternberg, 1995; Jackson et al., 1998; Totrov and Abagyan, 1994; Trosset and Scheraga, 1999; Wang et al., 1998; Rarey et al., 1997; Hoffmann et al., 1999; Weng et al., 1996).

The integration of experimental data into docking algorithms is a way to improve the quality and reliability of docking results. For ligand docking, it has been shown that the integration of NMR shift information is very useful (Polshakov et al., 1999). Similar results were described recently for protein–protein docking by Morelli et al. (2000). Common to these approaches is the use of distance constraints (intermolecular NOE constraints). To obtain these constraints, it is necessary to assign the majority of the shifts in the spectra used (e.g. heteronuclear ^{1}H-^{15}N-HSQC spectra). Since this assignment is a time-consuming process, the use of unassigned spectra would be desirable. One-dimensional ^{1}H-NMR spectra are the kind of spectra that are easiest and cheapest to obtain. Unfortunately, they contain less structural information than multi-dimensional spectra.

Methods

Preparation of structures

All complex structures were retrieved from the PDB (PDB IDs 1DT7, 1CFF, and 1CKK). From each structure containing several models, we selected the first model in the file. Missing hydrogens were added and all hydrogen positions were optimized in the complex using the AMBER 94 force field (Cornell et al., 1995). The complex structures were then separated into two files, each containing one of the complexed proteins or peptides, which were used in our docking algorithm.

Rigid body docking

For each example, we carried out a rigid body docking using the algorithm described by Lenhof (1995, 1997). The algorithm generates a list of tentative complex structures using geometric and energetic scoring functions. These structures are then evaluated and ranked with respect to our new scoring function.

This function is based on the prediction of the ^{1}H chemical shifts of the proteins. From these shifts, we simulate the ^{1}H-NMR spectrum of the complex and compare it to the experimental spectrum. The deviation of the simulated spectrum from the experimental spectrum is used to rank the structures. The details of these calculations are described in the following sections.

NMR chemical shift calculation

Our shift model decomposes the total chemical shift δ of a proton into four components

$$\delta = \delta_{\mathrm{RC}} + \delta_{\mathrm{A}} + \delta_{\mathrm{JB}} + \delta_{\mathrm{EF}} \qquad (1)$$

where δ_{RC} is the so called random coil shift, δ_{A} is the secondary shift caused by the magnetic anisotropy of the peptide bond, δ_{JB} is the ring current effect as calculated by the Johnson–Bovey theory, and δ_{EF} is the effect of the electric field.

Electric field effect

The electrostatic contribution is generally approximated to be linear in the projection of the electrostatic field E_z on the hydrogen bond:

$$\delta_{\mathrm{EF}} = \varepsilon E_z \tag{2}$$

For the constant ε, we used the parameters proposed by Williamson and Asakura (1993) for both C-H and N-H bonds. The electric field was calculated via Coulomb's law with atomic charges taken from the AMBER 94 force field (Cornell et al., 1995).

Magnetic anisotropy

The magnetic anisotropy of the peptide group is usually modeled by the approach of McConnell (1957). It describes the contribution to the chemical shift of a bond's magnetic anisotropy via the magnetic susceptibility tensor χ:

$$\delta_{\mathrm{A}} = \frac{1}{3N_A R^3} \sum_{i=x,y,z} \chi_{ii}(3\cos^2\theta_i - 1) \tag{3}$$

Here, R is the distance of the proton from the anisotropic bond, N_A is Avogadro's constant, and θ_i is the angle between the i-axis and the distance vector $\vec{R}$. Again, we used the parameters proposed by Williamson and Asakura for the C=O and C-N bond of the peptide group.

Ring current

The circular π-electron system of aromatic rings induces a magnetic field which changes the effective magnetic field at the nucleus and thus leads to a secondary chemical shift. There are two widely used approaches to calculate the ring current shift: the approach by Haigh and Mallion (1972) and the one by Johnson and Bovey (1958).

We used the Johnson-Bovey model, which is slightly more complicated to implement, but gives better accuracy. The secondary shift caused by the ring current effect is calculated as

$$\begin{aligned}\delta_{\mathrm{JB}} &= \frac{ne^0}{6\pi mc^2 a}\frac{a}{\sqrt{(a^2+\rho)^2+z^2}} \cdot \\ &\left(K + \frac{a^2-\rho^2-z^2}{(a-\rho)^2+z^2}E\right)\end{aligned} \tag{4}$$

where n is the number of aromatic π electrons in the ring, e^0 and m are the electron charge and mass, c is the speed of light, a is the ring radius, ρ and z give the position of the nucleus relative to the ring center in cylindrical coordinates, and K and E are complete elliptic integrals of the first and second kind. We used radii of 1.182 Å for the five-membered rings (His, Trp) and 1.39 Å for the six-membered rings (Phe, Tyr, Trp).

Random coil shift

The random coil shifts were obtained initially from the BMRB (Seavey et al., 1991), which provides a set of reference shifts for amino acid protons. To improve the quality of our shift model, we reparameterized the random coil shifts using a set of 21 proteins of known structure (obtained from the PDB (Bernstein et al., 1977)[1]) and shift assignment (from the BMRB). We calculated the chemical shifts for a training set of 14 proteins using our model and corrected the initial random coil shift for each proton by the average deviation observed in the training set. To verify that this correction did not depend on the protein set, we calculated the shifts of the seven remaining proteins using both the initial and the improved model. For the test set, the standard deviation of all ^{1}H chemical shifts dropped from an initial 0.52 ppm to 0.44 ppm.

All further calculations were carried out using this improved shift model, which is implemented in our Molecular Modeling framework BALL (Kohlbacher and Lenhof, 2000). The software and the parameters used for the calculations are available upon request.

Spectrum synthesis and comparison

Spectrum synthesis from experimental data

The assigned chemical shifts were read from the BMRB files (BMRB IDs 4099, 4284, and 4270). The spectrum was then simulated by assuming a Lorentzian line shape of equal width for each proton. Hence, the total spectrum S was written as a linear combination of Lorentzians:

$$S(x) = \sum_i \frac{1}{1 + \frac{(x-p_i)^2}{W}} \tag{5}$$

where p_i are the peak positions and W determines the (uniform) peak width. Since the line widths of the peaks are not known, we chose an average value of $W = 0.0032\,\mathrm{ppm}^2$.

[1] PDB IDs: 1B4M, 1BHI, 1BHU, 1BLQ, 1BLR 1BMX, 1BQV, 1C05, 1C15 1CB9, 1CEJ, 1CI5, 1CKV, 1LFC, 1AFP, 2BTX, 2CPB, 2CPS, 2IF1, 5PTI, 8TFV.

Spectrum synthesis from candidate structures

For each proton of the tentative complex structures, the chemical shift was calculated according to Equation 1. Then, we removed the most exchangable protons (OH of serine, threonine, tyrosine, aspartic acid, glutamic acid and NH of asparagine, glutamine, arginine and lysine NH_3), which are usually not present in the spectrum in a H_2O solution at neutral pH. We thus obtained a list of observable amide, aromatic, and aliphatic protons, which was used to create a spectrum as described above.

Comparison

For comparison, we sampled the 'experimental' spectrum $S_{\rm exp}$ and the spectrum of each tentative complex structure $S_{\rm cpx}$ in the range between -2 and $+12$ ppm with a total of 5000 regularly distributed positions x_i. The absolute difference area of the two spectra was then obtained as the sum of all unsigned differences:

$$\Delta(S_{\rm exp} - S_{\rm cpx}) = \sum_{i \in [-2,12]} |S_{\rm exp}(x_i) - S_{\rm cpx}(x_i)| \tag{6}$$

The resulting difference areas Δ were normalized by subtracting the smallest occurring area from all other areas. These values were then used to rank the structures.

Results

Our approach is based on unassigned one-dimensional ^{1}H-NMR spectra of the unknown protein complex. Instead of deriving distance constraints from the NMR spectra and integrating this geometric information into the docking process, we predict the ^{1}H-NMR spectra of all tentative complex structures proposed by our docking algorithm. The deviation of these predicted spectra from the experimental complex spectrum serves as a scoring function for the docking algorithm.

By combining docking techniques with NMR data, we intend to achieve two goals. First, we mean to speed up the process of structure elucidation of protein complexes by proposing shift assignments based on docking results. Second, the integration of experimental data can be used to improved the reliability of docking algorithms and to countercheck their results. Additionally, the NMR data can serve as a scoring function for those 'hard cases' where usual energetic scoring functions fail (as is the case for many protein–peptide complexes).

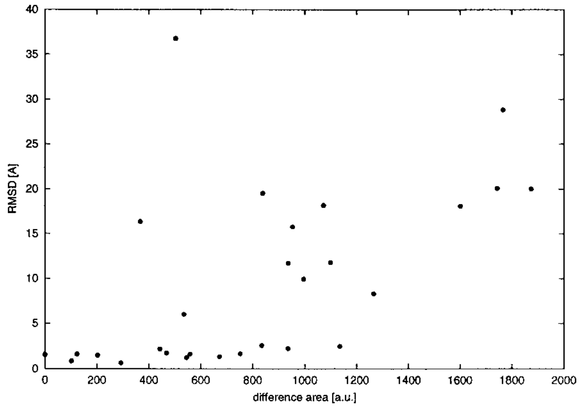

Figure 1. Results of the docking of the S100B(ββ) dimer.

In order to find a suitable test set for our approach we searched the contents of the BMRB (Seavey et al., 1991) for suitable protein complexes of known three-dimensional structure and available ^{1}H-NMR spectra. Unfortunately, the number of candidates is very small. We identified four candidates where the BMRB contained rather complete shift data of the complex and a corresponding structure was deposited in the Protein Data Bank (Bernstein et al., 1977): the complex of calmodulin with the Ca^{2+}-calmodulin-dependent protein kinase kinase (Osawa et al., 1999), the complex of calmodulin with a binding peptide of the Ca^{2+}-pump (Elshorst et al., 1999), the complex of S100B(ββ) with a peptide derived from p53 (Rustandi et al., 1998), and the two identical subunits of the homodimer S100B(ββ).

Since the spectra themselves are not stored in the BMRB, we had to reconstruct approximate experimental spectra from the shift assignments in the BMRB. For each of these complexes, we constructed a set of tentative complex structures via a rigid body docking algorithm (Lenhof, 1997) starting with the two bound structures.

The rigid docking of the four test cases resulted in four sets of tentative complex structures (each set with 24 through 121 different structures). For each of the potential complex structures, we calculated the spectra and determined the difference areas between these spectra and the experimental complex spectrum. In the case of S100B(ββ), the BMRB did not contain the complete shift data of the peptide and the complex, but only the shifts of the two dimer chains A and B.

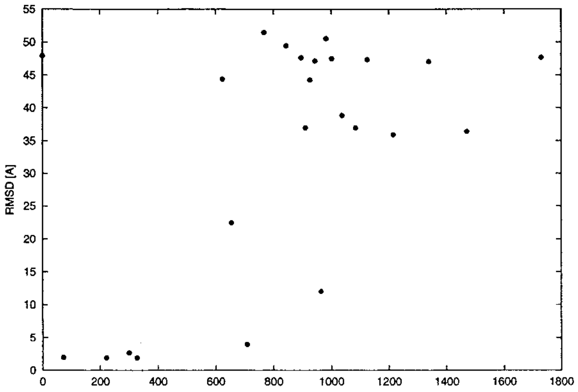

Figure 2. Results of the docking of calmodulin with the binding peptide of the Ca^{2+}-pump.

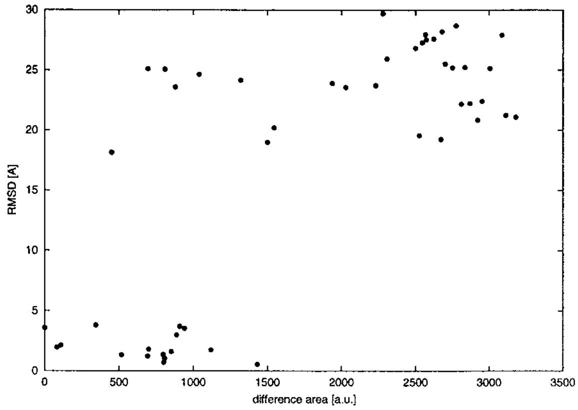

Figure 3. Results of the docking of calmodulin with the Ca^{2+}-calmodulin-dependent protein kinase kinase.

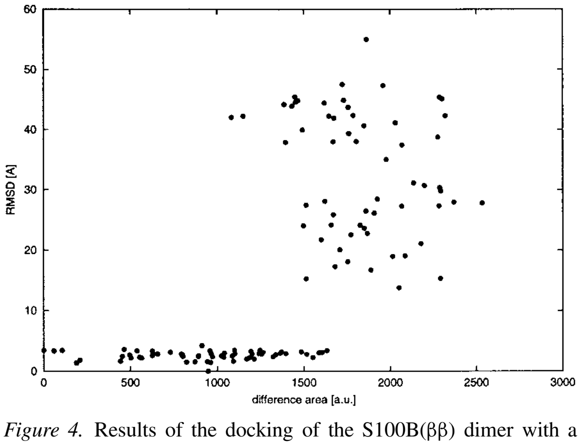

Figure 4. Results of the docking of the S100B(ββ) dimer with a peptide derived from p53.

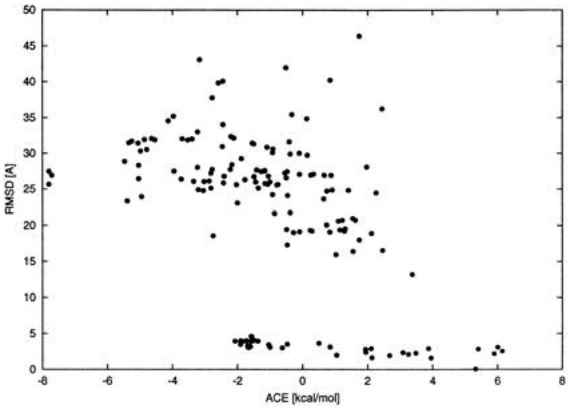

Figure 5. Results of the conventional docking of the S100B(ββ) dimer with a peptide derived from p53. No NMR data was used. Instead, we employed the Atomic Contact Energy (ACE) by Zhang et al. (1997) as a scoring function.

Initial experiments showed that the ranking of structures was significantly improved, if the contributions stemming from the magnetic anisotropy of the peptide group were not included. This seems surprising at first sight, but a detailed analysis of the shifts leads to the conclusion that the effect is a result of overlapping structures. Slightly overlapping structures are a typical result of rigid body docking algorithms. Even small deviations of the true complex structure can bring some groups into a closer spatial vicinity than could be expected from the atoms' van der Waals radii. Since the effect of the magnetic anisotropy grows with the third power of the inverse distance, these collisions lead to enormous changes in the chemical shift. Furthermore, the effect of the magnetic anisotropy is basically a local effect; it depends strongly on the backbone torsion angles (i.e. the secondary structure) and has a much more limited range than ring current and electric field effects. So it should not play a decisive role in the case of protein docking. Hence, we excluded the effect of the magnetic anisotropy between the two docking partners, but included it within each of the partners.

The results of the docking experiments are shown in Figures 1, 2, 3 and 4. In these figures, every point represents a single tentative complex structure. It shows the root mean square deviation (RMSD, y-axis) of the structure from the true complex structure and the normalized difference area of the candidate's spectrum (x-axis). Good approximations of the true complex structure should thus be expected in the lower left corner of the graph.

Except for the complex of calmodulin with the binding peptide of the Ca^{2+}-pump, scoring according to the difference area always identified a good approximation of the true complex structure. The separation between true and false positives was good for the S100B(ββ) dimer and for the complex of calmodulin and kinase, and excellent for the complex of S100B(ββ) and the p53-peptide.

The latter fact is very surprising, since the docking of the small p53-derived peptide (22 amino acids) was impossible using conventional methods. We tested different energy-based scoring functions but we were not able to obtain a reasonable ranking. Figure 5 shows the result of the docking using the Atomic Contact Energy (ACE) developed by Zhang et al. (1997). The first approximation of the true complex structure is ranked as number 48. Other scoring functions, e.g. the use of geometric methods (Katchalski-Katzir et al., 1992) or the inclusion of electrostatics gave very similar results. The problem with this docking example stems basically from the small binding site of the peptide. Most docking algorithms favour structures where the peptide has a larger contact area with the protein. In this case, the use of NMR data was the only possibility to correctly predict the complex structure.

For the complex of calmodulin with the binding peptide of the Ca^{2+}-pump, a false positive structure was ranked number one, followed by the major part of the true positive structures. The reasons for this failure are not yet clear.

Discussion

We have presented a new scoring function for evaluating tentative protein complexes that compares calculated and experimental ^{1}H-NMR spectra and does not use distance constraints (intermolecular NOE constraints). The first docking experiments with bound structures look very promising, but more experiments are necessary to validate the method, i.e. more experiments with bound structures as well as experiments with unbound, native structures. These experiments will also allow us to use true experimental data instead of spectra constructed from shift assignments.

These first experiments indicate that the use of NMR data can improve the reliability and accuracy of docking predictions, but the new method still needs a more extensive validation with experimental data. Unfortunately, the number of protein complexes of known 3D structure with available ^{1}H-NMR spectra or other one-/multi-dimensional spectra is very small.

Since there is more experimental data available for single proteins than for protein complexes, we intend to validate the new method by applying it to tentative protein structures produced by protein structure prediction methods (e.g. threading). We argue that this problem will be even simpler, because the structural differences between the tentative protein structures will be larger than the differences experienced in protein docking.

In close cooperation with NMR spectroscopists we are applying our docking approach to protein complexes with unknown 3D structure. We try to predict the structures of the complexes. The results of our docking algorithm are used to speed up the shift assignment of the complexes. On the other hand, the results of the structure elucidation via NMR will be used to validate our docking algorithms.

Our future research will address the use of ^{13}C- and ^{15}N-NMR spectra as well as the extension of our techniques to multi-dimensional heteronuclear spectra (e.g. ^{1}H-^{15}N-HSQC), which contain more structural information.

Acknowledgements

This work was supported by the DFG research cluster ‘Informatikmethoden zur Analyse und Interpretation großer genomischer Datenmengen’, grants LE952/2-1 and BA1624/3-1.

References

Bernstein, F., Koetzle, T., Williams, G., Meyer Jr., E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535–542

Betts, M.J. and Sternberg, M.J.E. (1999) *Protein Eng.*, **12**, 271–283.

Connolly, M.L. (1986) *Biopolymers*, **25**, 1229–1247.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz Jr., K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. and Kollman, P.A. (1995) *J. Am. Chem. Soc.*, **117**, 5179–5197.

Elshorst, B., Hennig, M., Foersterling, H., Diener, A., Maurer, M., Schwalbe, H., Griesinger, C., Krebs, J., Schmid, H., Vorherr, T. and Carafoli, E. (1999) *Biochemistry*, **38**, 12320–12332.

Fischer, D., Lin, S.L., Wolfson, H.J. and Nussinov, R. (1995) *J. Mol. Biol.*, **248**, 459–477.

Haigh, C.W. and Mallion, R.B. (1972) *Org. Magn. Reson.*, **4**, 203–228.

Hoffmann, D., Kramer, B., Washio, T., Steinmetzer, T., Rarey, M. and Lengauer, T. (1999) *J. Med. Chem.*, **42**, 4422–4433.

Jackson, R.M., Gabb, H.A. and Sternberg, M.J.E. (1998) *J. Mol. Biol.*, **276**, 265–285.

Jackson, R.M. and Sternberg, M.J.E. (1995) *J. Mol. Biol.*, **250**, 258–275.

Johnson, C.E. and Bovey, F.A. (1958) *Chem. Phys.*, **29**, 1012–1030.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Afalo, C. and Vakser, I.A. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 2195–2199.

Kohlbacher, O. and Lenhof, H.-P. (2000) *Bioinformatics*, **16**, 815–824.

Lengauer, T. and Rarey, M. (1996) *Curr. Opin. Struct. Biol.*, **6**, 402–406.

Lenhof, H.-P. (1995) In: *Bioinformatics: From nucleic acids and proteins to cell metabolism* (Eds, Schomburg, D. and Lessel, U.), GBF Monographs Volume 18, pp. 125–139.

Lenhof, H.-P. (1997) In: *Proceedings of the First Annual International Conference on Computational Molecular Biology RECOMB 97*, pp. 182–191.

McConnell, H.M. (1957) *J. Chem. Phys.*, **27**, 227–229.

Meyer, M., Wilson, P. and Schomburg, D. (1996) *J. Mol. Biol.*, **264**, 199–210.

Morelli, X., Dolla, A., Czjzek, M., Palma, N., Blasco, F., Krippahl, L., Moura, J.J.G. and Guerlesquin, F. (2000) *Biochemistry*, **39**, 2530–2537.

Norel, R., Lin, S.L., Wolfson, H.J. and Nussinov, R. (1994) *Biopolymers*, **34**, 933–940.

Osawa, M., Tokumitsu, H., Swindells, M.B., Kurihara, H., Orita, M., Shibanuma, T., Furuya, T. and Ikura, M. (1999) *Nat. Struct. Biol.*, **6**, 819–824.

Polshakov, V.I., Morgan, W.D., Birdsall, B. and Feeney, J. (1999) *J. Biomol. NMR*, **14**, 115–122.

Rarey, M., Kramer, B., Lengauer, T. and Klebe, G. (1997) *J. Mol. Biol.*, **261**, 470–489.

Rustandi, R.R., Drohat, A.C., Baldisseri, D.M., Wilder, P.T. and Weber, D.J. (1998) *Biochemistry*, **37**, 1951–1960.

Sandak, B., Nussinov, R. and Wolfson, H.J. (1998) *J. Comput. Biol.*, **5**, 631–654.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J. (1991) *J. Biomol. NMR*, **1**, 217–236.

Shoichet, B.K. and Kuntz, I.D. (1991) *J. Mol. Biol.*, **221**, 79–102.

Sternberg, M.J.E., Gabb, H.A. and Jackson, R.M. (1998) *Curr. Opin. Struct. Biol.*, **8**, 250–256.

Totrov, M. and Abagyan, R. (1994) *Nat. Struct. Biol.*, **1**, 259–263.

Trosset, J.-Y. and Scheraga, H.A. (1999) *J. Comput. Chem.*, **20**, 412–427.

Wang, J.-M., Xu, X.-J. and Jiang, F. (1998) In: Proceedings of the Fourth Chinese Peptide Symposium (Eds, Xu, X.-J., Ye, Y.-H. and Tam, J.P.), Kluwer, Dordrecht, pp. 106–108.

Weng, Z., Vajda, S. and Delisi, C. (1996) *Protein Sci.*, **5**, 614–626.

Williamson, M.P. and Asakura, T. (1993) *J. Magn. Reson.*, **B101**, 63–71.

Zhang, C., Cornette, J.L. and DeLisi, C. (1997) *Protein Sci.*, **6**, 1059–1064.