J|A|C|S
COMMUNICATIONS

Published on Web 03/20/2004

Distributed Computing and NMR Constraint-Based High-Resolution Structure Determination: Applied for Bioactive Peptide Endothelin-1 To Determine C-Terminal Folding

Hiroyuki Takashima,† Norio Mimura,† Tadayasu Ohkubo,‡ Takuya Yoshida,‡ Haruhiko Tamaoki,‡ and Yuji Kobayashi*,‡

Informatics and Knowledge Management at Novartis Institutes for BioMedical Research, Tsukuba Research Institute, Ohkubo 8, Tsukuba, Ibaraki, 300-2611 Japan, and Graduate School of Pharmaceutical Sciences, Osaka University, 1-6 Yamadaoka, Suita, Osaka, 565-0871 Japan

Received December 11, 2003; E-mail: hiroyuki.takashima@pharma.novartis.com

Three-dimensional structure determination of protein using NMR is often based on the distance geometry method and/or restrained molecular dynamics calculations.[1] Overall convergence of the structures from NMR is based not only on quality of experimental data but also computing power to explore the huge conformational space of the protein. In the structure determination, the conformational space is almost identical with torsion angle space and increases with size of protein. At present, on a single processor, limitation of the computational power restricts the quantities of the calculations, and less than 200 of initial structures or molecular dynamics trajectories have been used to explore the conformational space. Here, we have used a distributed computing implementation to calculate tens of thousands of structures to explore the conformational space comprehensively and to determine high-resolution structures. We applied this method to endothelin-1 (ET-1), a novel cardiovascular bioactive peptide.

ET-1 is a suitable model system of protein folding, composed of secondary structural elements, α-helix and β-turn, which are stabilized by two disulfide bonds in its 21 amino acid sequence. The cystine-stabilized α-helix motif (CSH-motif) was discovered by our group for ET-1[2] and investigated for a series of peptides in a variety of species.[3] The motif is in the N-terminal region of ET-1, resulting in convergence of the region for previously reported NMR studies,[2,4] which are in agreement with X-ray structures.[5] However, there has been no consensus about C-terminal structures,[4c] which were dispersed in the previous NMR studies. The C-terminal residues play an important role in the bioactivities of ET-1.[6] Determination of the C-terminal folding is a key target of the method we present here.

NMR experimental data of ET-1 were collected by standard strategy[7] and conditions.[8] Three hundred nine distance constraints obtained by homonuclear two-dimensional NOESY were subjected to distance geometry and simulated annealing calculations using XPLOR-NIH v2.0.6[9] on the distributed computing of 17 Linux PCs controlled by a SUN GRID engine.

The calculations were started by generating initial structures from random array,[10] and followed by 80-ps simulated annealing (SA) at high initial temperature, 5000 K. The total number of initial structures was increased up to 32 000, and 20 final structures with lowest energies of target function were selected to elucidate root mean square deviations (rmsd). To investigate the effect of sampling scale based on the numbers of initial structures, the rmsd values were compared between smaller and larger subsets of SA structures, from 50 to 32 000. The results, plotted in Figure 1, indicate a clear dependency of the rmsd on the initial structures,

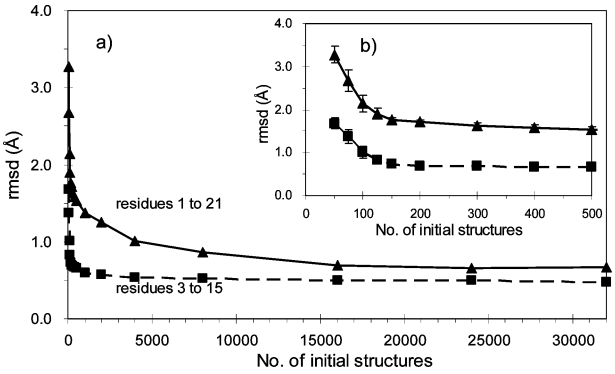

Figure 1. (a) The pairwise averaged rmsd values of all backbone atoms of ET-1 have been elucidated for 20 minimum energy structures out of various numbers of calculated structures with 80-ps simulated annealing at initial temperature 5000 K. (▲) Overlay for all residues from 1 to 21. (■) Overlay for residues 3 to 15. (b) The same plot as (a), expanding the horizontal axis.

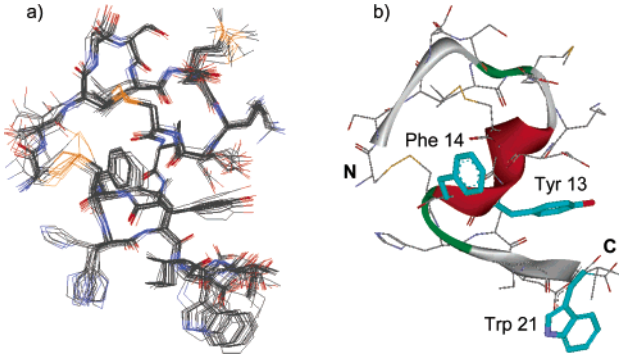

Figure 2. (a) Overlay of 20 minimum energy structures of ET-1, pairwise backbone rmsd 0.58 ± 0.28 Å. (b) The lowest energy structure of (a) with ribbon drawing, schematically representing α-helix and β-turns in red and green, respectively.

even for thousands and tens of thousands of structures. The rmsd decreased exponentially with an increase of initial structures up to 16 000, and after that it was almost constant to 32 000. These numbers are 100 times larger than those used in ordinary structure determination.

The structures determined here have well-defined C-terminal conformation, as shown in Figure 2, and have been deposited in the Protein Data Bank as entry 1v6r. The C-terminal part of the peptide has an extended β-structure and is loosely looped back to the α-helix by a turn in the junction, forming a hydrophobic core around the side chain of Tyr 13. Thus, the side chain of Trp 21, which plays an important role in expression of bioactivities and is

† Tsukuba Research Institute.
‡ Osaka University.

10.1021/ja031637w CCC: $27.50

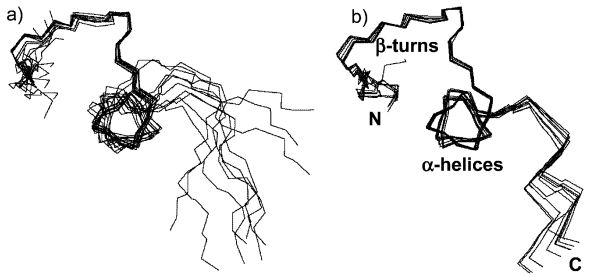

Figure 3. ET-1 structures calculated with (a) 100 initial structures and (b) 32 000 initial structures. Both structures are the overlay of 10 minimum energy structures, fitting residues 3 to 15, to indicate differences in C-terminus convergences.

recognized as a pharmacophore of ET-1 antagonists, is close to the rings of Tyr 13 and Phe 14, with distances of 8.5 ± 0.1 Å and 11.6 ± 0.1 Å measured at C_γ positions, respectively (see Figure 2b). This folding is not consistent with X-ray crystal structure,[5] which has α-helix in the C-terminal region, but is in good agreement with experimental NMR parameters (J-coupling and sequential NOEs).[2,4] These conformational differences between solution and crystal structure had been suggested by a previously reported NMR study.[4c]

On the other hand, the conformation of the N-terminal region, which was already defined in the past studies,[2,4,5] reached baseline much faster. With 200 initial structures, the residues 3 to 15 have an rmsd of 0.7 Å and have well-defined structures. The fast decrease of rmsd is considered to reflect small conformational space restricted by two disulfide bonds, consisting of the CSH-motif that we had investigated. Compared to NOE distance constraints, which imply experimental error margins, the disulfide bond is a strong constraint in the molecular dynamics calculations. Two hundred initial structures is sufficient to investigate such a small conformational space, but not sufficient for a larger one such as a peptide of 21 amino acids.

The rmsd values for all residues, 1 to 21, exhibit a double exponential decay as shown in Figure 1a,b. It is obvious that the faster decay is caused by the convergence of the N-terminal region, followed by a second very slow decay, which is presumably caused by the difference of constraint strength between disulfide bonds and NOE. With the use of a small number of initial structures, i.e., less than 500, the structure determinations will be very likely trapped in the localized folds, and it will bias the overall convergences. Figure 3 represents the convergences of ET-1 with two different numbers of initial structures. When we use a small number of initial structures, i.e., 100, our result of ET-1 (see Figure 3a) shows a dispersion of the C-terminus and rmsd values similar to the previous studies.[2,4] The dispersion is caused from localized folding of the N-terminal trapped by initial structure dependencies, and that is the reason NMR solution structure of ET-1 had missed the convergence in the C-terminal region. It is strongly suggested that the same bias of overall convergences will happen in general structure determinations of proteins with an insufficient number of initial structures or molecular dynamics trajectories.

The most effective and simple way to avoid the initial structure dependency is calculation of the largest possible number of structures to be selected out by their target function energies. The distributed computing is a powerful tool for such a huge number of independent calculations, and its computational power can be easily increased by increasing the number of connected PCs. Our calculations in this study might be a year of calculations for a single computational workstation, but by using 17 PCs, it took only weeks. With 100 PCs, the calculation will take only days. And the cost of 100 PCs is almost the same as one middle range computational workstation.

In general, the conformational space of proteins is much larger than that of ET-1, and the number of initial structures needed to explore them comprehensively is definitely greater than tens of thousands. To use CPU resources of the distributed computing effectively, we propose a two-step strategy of the structure calculation as follows. First, start calculation from about 10 thousand completely random initial structures with tens of picoseconds of relatively short SA at high initial temperature, and second, sort out 10% of the structures by their target function energy to be subjected to hundreds of picoseconds of long SA at relatively low initial temperature. In this strategy, the conformational space will be reduced by the first SA and effectively explored by the second SA. We implemented the strategy using a 113 amino acid protein, neocarzinostatin, obtained strikingly well-defined structures with rmsd 0.28 ± 0.06 Å, and deposited it in the Protein Data Bank as entry lo5p.[11]

Acknowledgment. We thank Dr. Evelyn Stimson and Ms. Susan C. DiClemente for their helpful discussions and SUN Microsystems Japan and Mr. Hideaki Numa for their support.

Supporting Information Available: Details of NMR data and structures (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) (a) Wüthrich, K. *Science* **1989**, *243*, 45−50. (b) Clore, G. M.; Gronenborn, A. M. *Science* **1991**, *252*, 1390−1399.
(2) Tamaoki, H.; Kobayashi, Y.; Nishimura, S.; Ohkubo, T.; Kyogoku, Y.; Nakajima, K.; Kumagaye, S.; Kimura, T.; Sakakibara, S. *Protein Eng.* **1991**, *4*, 509−518.
(3) (a) Kobayashi, Y.; Sato, A.; Takashima, H.; Tamaoki, H.; Nishimura, S.; Kyogoku, Y.; Ikenaka, K.; Kondo, T.; Mikoshiba, K.; Hojo, H.; Aimoto, S.; Moroder, L. *Neurochem. Int.* **1991**, *18*, 525−534. (b) Kobayashi, Y.; Takashima, H.; Tamaoki, H.; Kyogoku, Y.; Lambert, P.; Kuroda, H.; Chino, N.; Watanabe, T. X.; Kimura, T.; Sakakibara, S.; Moroder, L. *Biopolymers* **1991**, *31*, 1213−1220. (c) Tamaoki, H.; Miura, R.; Kusunoki, M.; Kyogoku, Y.; Kobayashi, Y.; Moroder, L. *Protein Eng.* **1998**, *11*, 649−659.
(4) (a) Krystek, S. R., Jr.; Bassolino, D. A.; Novotny, J.; Chen, C.; Marschner, T. M.; Andersen, N. H. *FEBS Lett.* **1991**, *281*, 212−218. (b) Saudek, V.; Hoflack, J.; Pelton, J. T. *Int. J. Pept. Protein Res.* **1991**, *37*, 174−179. (c) Wallace, B. A.; Janes, R. W.; Bassolino, D. A.; Krystek, S. R., Jr. *Protein Sci.* **1995**, *4*, 75−83.
(5) Janes, R. W.; Peapus, D. H.; Wallace, B. A. *Nat. Struct. Biol.* **1994**, *1*, 311−319.
(6) (a) Nakajima, K.; Kubo, S.; Kumagaye, S.; Nishio, H.; Kuroda, H.; Tsunemi, M.; Inui, T.; Kuroda, H.; Chino, N.; Watanabe, T. X.; Kimura, T.; Sakakibara, S. *Biochem. Biophys. Res. Commun.* **1989** *163*, 424−429. (b) Aumelas, A.; Kubo, S.; Chino, N.; Chiche, L.; Forest, E.; Roumestand, C.; Kobayashi, Y. *Biochemistry* **1998**, *37*, 5220−5230.
(7) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; John Wiley & Sons: New York, 1986.
(8) ET-1 was dissolved in H_2O with D_2O (9:1, v/v), at the concentration of 2.5 mM with 5% of deuterated acetic acid, and measured with a Bruker AMX-600 spectrometer at a temperature of 297.7 K. Mixing times were NOESY = 350 ms.
(9) Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, G. M. *J. Magn. Reson.* **2003**, *160*, 66−74.
(10) Cartesian coordinates of all atoms were randomly varied from −10 Å to 10 Å. Randomness of the initial structures was verified by calculation of backbone rmsd value, 12.9 ± 0.5 Å, for all of the sampling scales.
(11) Takashima, H.; Ishino, T.; Yoshida, T.; Hasuda, K.; Ohkubo, T.; Kobayashi, Y., manuscript to be submitted for publication.

JA031637W