Open Access

Addressing Two Issues on Conformational Differences and Utility of Random Number Generators in Conformational Sampling

R. S. Rathore*[,1,2]

[1] *Tata Institute of Fundamental Research, Center for Applicable Mathematics, Bangalore 560065, India*

[2] *School of Biotechnology, Devi Ahilya University, Indore 452001, India*

Abstract: Conformation search procedure is carried out to generate an ensemble of conformations that could be used for purposes such as to address the problem of ligand and receptor flexibility during molecular docking, protein structure predictions, loop modeling and to calculate binding properties. Two important aspects have been discussed, the criteria of different conformations for the purpose of effective conformation sampling in systematic conformation searches, and the redundancies of using random number generators in random searches.

INTRODUCTION

Conformation search is a technique, which is employed to examine conformation excursions and build an ensemble in conformational sampling [1-4]. The generated ensemble serves a preliminary data to address flexibility of ligand and enzyme during docking, calculate binding properties, explore dynamics, and to address protein folding and *ab initio* structure prediction problems. The article addresses two important issues in the conformation sampling of a polypeptide. First, to elaborate on the criteria of different conformations for the purpose of effective conformation sampling in deterministic (systematic) conformation searches. Second, it explains the redundancies of using random number generators in stochastic (random) searches.

Distinguishing conformations is key to understand protein functions, mapping their conformational landscape, structure comparisons, clustering and database searches [5-8]. All the computational approaches aiming to perform conformational search procedures confront with a common problem i.e. to define a suitable measure of conformational difference, and thereby map complete or a representative conformational landscape of a molecule. This question is relevant in the computational structural biology and it is the purpose of this article to assort cases where the question may be unambiguously answered. Trivially speaking, the present question matters in the context of molecules possessing a well defined structure, and therefore it is irrelevant for some special classes of proteins such as intrinsically unstructured proteins [9]. In general, the term "*different conformations*" has practically no meaning, where the overall conformation is averaged out.

The systematic or total conformation search is routinely carried out in a grid fashion and an important question is what should be the size of a unit grid (steps of φ and ψ for proteins) as a minimum measure of different conformation, in order to sample the total conformational landscape of a molecule? The systematic conformation search procedure is computationally tractable but still impracticable for macromolecules. A large number of conformers of the order of 2^{100} are possible even for a small protein of about 100-residue length, assuming only two conformations per residue. To generate all these structures, it would require about an age of universe. As a practical solution, a representative set of the total population is generated using stochastic or *Monte Carlo* searches. In this approach, protein structures are generated at random and a variety of random number generators are employed for this purpose.

Taking a reasonable assumption that protein conformations are essentially described by backbone torsion angles, ϕ and ψ, and ignoring the minor flexibility of peptide plane ($\Delta\omega \approx 10^{\circ}$) and molecular geometry, the foregoing question translates into the language of mathematics, how different these values should be for different conformers. Protein structures could be compared in many ways [6], and the root mean square deviation (rmsd) is the most popular measure for its simplicity and robustness. The commonly accepted criterion of similar conformations by model builders and crystallographers is not a universal one [10]. The C^{α}-rmsd value of 0.1-0.6Å for similar conformations (precisely speaking 0.1-0.6Å per residue or atom under consideration) is equivalent to the backbone torsion angle difference of 5-25º (the value of 0.1Å is approximately equivalent to a rotation angle of 4º, assuming bond distance of 1.54Å; supplementary Fig. **1**). Apart from experimental errors, these figures inherently embody the thermal motion of atoms leading to random distribution of their positions to this extent. These figures of similarity are not sacrosanct and depend upon the model quality and the size of macromolecule [11]. For poorly resolved and less accurate structures, more relaxed criterion could be employed. Mathematically answering the question for a general case was elegant but what about for proteins, subjected to systematic changes? We know that the small systematic deviation occurring for each residue within a regular secondary structure would lead to a different type of structure. In

*Address correspondence to this author at the Tata Institute of Fundamental Research, Center for Applicable Mathematics, Bangalore 560065, India; E-mails: ravindranath_rathore@yahoo.com, rathore@math.tifrbng.res.in natural proteins, this type of systematic differences among conjugated residues is unusual. Instead, fluctuating deviations from their regular values are generally observed. As against thermally induced random motion, the systematic motions occur in proteins as a result of protein folding process, conformational transitions and ligand-induced conformational changes. A notable example is a hinge-like motion [12], where even a small conformational change at the pivot would lead to large changes at far away places and the aforementioned criteria of similarity no longer holds true (supplementary Fig. **2**).

Having answered the first question, we are now in a position to appreciate the redundancies of using the random number generators in random sampling. Do random number generators serve any useful purpose in protein structure generations? Random numbers are needed to generate random conformational parameters of polypeptide chain. Typically, random numbers in (0, 1) range are generated which are then converted to corresponding parameters by suitable mapping. For a general peptide sampling, about 10-15 random numbers are needed to populate random conformation angles of backbone and side chains, their preferred values and sometimes geometric parameters. Even considering the unit grid size for a torsion angle down to 1°, there would be only 361 different values to try for. Indeed, the number of trials are even less as only ~50% and ~25% of the total Ramachandran space qualify for the allowed region of Gly and Ala, respectively, and further less for β-branched residues and Pro, whereas only 5-10% remain accessible for non-natural residues such as Aib [13]. Permutations of all these values for an angle with the rests lead to an astronomical combination of conformations. Hence, it is purposeless to employ random number generators to pick one conformation parameter among very few, and their use is an overhead for the program. You would pick these conformational parameters randomly anyway, simply by random shuffling, without using sophisticate random number generators. Although, efficient codes of random number generators are available that can execute the job in just a fraction of seconds. But for large-scale sampling, random numbers are needed in trillions. Profile analysis of a program, employing such calculations would reveal that a single random number generator may consume even up to 1-2% of the total computational time. Given that about dozen such random numbers are expected to be employed, this would be significant time consumption. Certainly, there are even higher computationally expensive jobs during conformation searches such as energy calculations, and they have already been discussed at length elsewhere [1,14]. Contrarily, the numerical counts of randomly generated numbers overwhelmingly exceed the requirement, in other words they are far more accurate (about 5-20 decimal places in (0, 1) range) than the actual demand (only up to couple of decimal places), on an average we are more likely to generate the redundant (same) conformation. It is rather unfortunate that many public and in-house programs, performing conformational search in one way or the other, waste precious time in generating random numbers.

ACKNOWLEDGEMENT

The author acknowledges Science and Engineering Research Council, Department of Science and Technology, Government of India for a Fast-Track scientist award.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] Howard AE, Kollman PA. An analysis of current methodologies for conformational searching of complex molecules. J Med Chem 1988; 31: 1669-75.

[2] Beusen DD, Shands EFB, Karasek SF, Marshall GR, Dammkoehler RA. Systematic search in conformational analysis. J Mol Struct Theochem 1996; 370: 157-71.

[3] Boström J. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. J Comp Mol Des 2001; 15: 1137-52.

[4] Höltje H-D, Sippl W, Rognan D, Folkers G. Molecular modeling: basic principles and applications. Chp. 2, 3rd ed. New Jersey: Wiley-VCH 2008, pp. 32-49.

[5] Smellie A, Kahn SD, Teig SL. Analysis of conformational coverage. 1. Validation and estimation of coverage. J Chem Inf Comput Sci 1995; 35: 285-94.

[6] Koehl P. Protein structure similarities. Curr Opin Struct Biol 2001; 11: 348-53.

[7] Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol 2005; 346: 1173-88.

[8] Aung Z, Tan KL. Rapid retrieval of protein structures from database. Drug Disc Today 2007; 12: 732-9.

[9] Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005; 6: 197-208.

[10] Maiorov VN, Crippen GM. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J Mol Biol 1994; 235: 625-34.

[11] Carugo O. How root-mean-square distance (r.m.s.d.) values depend on the resolution of protein structures that are compared. J Appl Cryst 2003; 36: 125-8.

[12] Flores SC, Lu LJ, Yang J, Carriero N, Gerstein MB. Hinge Atlas: relating protein sequence to sites of structural flexibility. BMC Bioinformatics 2007; 8: 167: 1-20.

[13] Rathore RS. Structural studies of biologically active and conformationally important oligopeptides: implications for *de novo* design. PhD thesis. Department of Physics, Indian Institute of Science: Bangalore 1999.

[14] Vásquez M, Némethy G, Scheraga HA. Conformational energy calculations on polypeptides and proteins. Chem Rev 1994; 94: 2183-239.

Received: June 01, 2009 Revised: July 27, 2009 Accepted: September 14, 2009