ANOLEA: A www Server to Assess Protein Structures

Francisco Melo, Damien Devos, Eric Depiereux and Ernest Feytmans

Facultés Universitaires Notre Dame de la Paix, Department of Biology, Laboratoire de Biologie Moléculaire Structurale.
61, Rue de Bruxelles, B-5000 Namur, Belgium.

Abstract

ANOLEA (**A**tomic **No**n-**L**ocal **E**nvironment **A**ssessment) is a www server that performs energy calculations at the atomic level in protein structures. The calculations involve the non-local interactions between all the heavy atoms of the twenty standard amino acids in the molecule. The input of the server is a PDB file containing one or more protein chains. The output is an energy profile, which gives an energy value for each amino acid of the protein. High energy zones (HEZs) in the profile correlate with errors or with potential interacting zones of proteins. The output of the server also displays the structure in three dimensions, pointing out the high energy amino acids in the protein. This option requires the *CHIME* plug-in, which is freely available on Internet and makes possible, in real time, to rotate, translate and change the point of view and presentation of the molecule in three dimensions. Thus, a fast analysis of a protein structure can be done using a personal computer connected to Internet. The server is available at: **http://www.fundp.ac.be/pub/ANOLEA.html**.

Introduction

The three-dimensional arrangement of the different atoms of a protein contributes to its stability, functionality, specificity and regulation. The stability is the first aspect to be accomplished in order to become a 'possible' protein. The specificity, functionality and regulation of a protein could be seen as additional requirements to become 'useful'. These last two aspects are strongly related because the regulation is always an important part of the protein functionality.

Proteins are not stand-alone systems, but they are dynamic and interacting systems. An isolate protein molecule in water has no meaning, no function, no purpose. Proteins carry out very specific functions interacting with other molecules through molecular recognition. This process is very specific and involves atomic interactions. The understanding of the molecular recognition process is one of the final goals in molecular biology. All molecules in living systems carry their function through the highly specific molecular recognition process. All protein-protein interactions, enzyme-substrate bindings, gene activation and repression by transcription factors, ligand-receptor interactions, etc., are responsible of the rich and complex chemical dynamism that makes life possible. As we know, the same basis makes disease possible. This is one of the reasons why it is central in biology to understand this process. The success of many drugs and chemical compound developments depends on the quality of the protein structure used as template to build them. A protein structure containing errors in important functional regions could avoid a successful design of drugs aimed to cure a disease. In the other hand, a high quality structure can also provide the basis to carry out successfully new *in silica*, *in vitro*, or *in vivo* experiments designed to uncover the function of different residues in the protein. Thus, it is crucial to develop accurate methods to detect errors and to improve protein structure quality.

Several methods are used to detect errors in proteins. We have shown that the method used by our server is more sensitive and accurate in detecting errors than other ones (1). This method refer to a knowledge-based mean force potential (MFP) at the atomic level. A knowledge-based MFP is a potential statistically derived from a non-redundant database of protein structures. Several knowledge-based MFPs have been used in many applications as threading or fold recognition (2, 3), comparative modeling or homology modeling (4), molecular docking or molecular recognition (5), protein structure quality assessment (4), ab initio protein structure prediction (6), and they have proved their usefulness in all these protein structure prediction techniques.

We have derived a MFP at the atomic level using atom type definitions, which makes possible to raise the frequency of observations and to obtain accurate energy functions (7). The MFP at atomic level contains all the terms describing the energy for the non-local pairwise interactions between the heavy atoms of the 20 standard amino acids (1). The environment of each heavy atom in a protein is evaluated by calculating the non-local energy profile (NL-EPs) for it. The profiles are able to point out high energy zones (HEZs) in these proteins, which correlate with punctual errors and, sometimes, with interacting regions of proteins. The MFP is not always able to detect all the errors of protein models (i.e. it gives false negatives), but is very reliable about false positives. Our approach involves only energy terms

without consideration of any other parameter. Here we describe the current implementation of our www server and we also give some examples of its utility in molecular biology.

Methods

We defined a total number of 40 different atom types for all the heavy atoms of the 20 amino acids (7). The atom type definition is based on its connectivity, chemical nature and location level (side-chain or backbone). The MFP was calculated from a set of 147 protein chains obtained from the PDB (8) with complete atomic coordinates, excluding all the proteins with duplicated or missing atoms, structural gaps, or with a number of residues lower than 100 (1). In the case of multimers, when possible, the missing chains were generated using the transformation matrices available in the PDB files. The MFP was calculated on the non-redundant chains, taking into account contacts with the other chains. A distance-dependent MFP (DD-MFP) involving only non-local interactions was developed. The non-local environment of one atom is defined as the set of all the heavy atoms, within an Euclidean distance of 7 Å, that belongs to amino acids that are farther than 11 residues in the chain or that belong to another chain (1). The 7 Å radius was divided into 35 intervals of 0.2 Å each. The MFP was determined symmetrically. The calculation of pairwise pseudo-energy terms has been carried out as described elsewhere (1). The non-local energy profiles (NL-EPs) were calculated as follows: a) The energy value for each atom-atom interaction was taken from the DD-MFP. b) The energy of each residue is the sum of the energies of all its atoms. The NL-EPs are displayed using a window average of 5 residues. Each energy value is divided by kT (0.582 kcal/mol), where k is the Boltzmann's constant and T the absolute temperature. The NL-EPs are then expressed in E/kT units. A threshold of zero units is used to define a HEZ. The non-local contact maps were calculated considering a contact between one atom and any heavy atom, within an Euclidean distance of 7 Å, that belongs to an amino acid that is farther than 11 residues in the chain. Each residue contains the sum of all their atomic contacts. Any pair of residues with ten or more atomic contacts is displayed in the figures as a dot.

Server description

The server is implemented for on line access. It processes the submitted information and immediately returns the results of the calculation via another html page, allowing to perform the analysis interactively. A complete on line help for each parameter is provided.

Input parameters

The server requires four different input parameters: two mandatory are a file in PDB format specifying the coordinates of each heavy atom in the molecule and the protein chain information that specifies on which chain the profile must be calculated and which other chains have to be considered in the calculation. These two parameters must be filled out by the user. The other two parameters are the threshold to define a HEZ and the window average to perform the energy profile. These last two parameters are provided by default and are required for the calculation. Default values can be changed by the user. The sensibility of the HEZ detection is determined by the window average and by the threshold specified. The default values are a window average of 5 residues and a threshold of zero E/kT units. The possible values for the window average parameter are 1, 3, 5, 7 and 9. The threshold can be any real number.

Server output

The server output is divided into five sections: the first one provides the values used to perform the calculations and a five columns output specifying the residue number, the amino acid name (3 letters code), the energy of the residue using the window average specified, a binary label specifying if the amino acid constitutes or not a HEZ with the threshold defined, and the total number of non-local contacts of each residue in the molecule. The second section contains a graphic output showing in three dimensions all the submitted protein chains, pointing out the HEZs in the protein chain on which the calculations were performed. The CHIME plug-in contains many incorporated commands to modify the visualization and appearance of the molecule. However, some options cannot be carried out by the user using only the incorporated commands of CHIME. The last three sections were designed to satisfy the user requirements. An ensemble of buttons that control the visualization of the protein is available. Also, a command line is available, where a RASMOL script can be entered and executed. Although some commands are redundant with the ones available in CHIME, many others are exclusive from the results of the calculation and specific of the submitted protein. The user who has not installed the CHIME plug-in cannot visualize the molecule in three dimensions, but the first section of the output will be always available and it contains all the necessary data to carry out the analysis using another software. However, we

recommend to use all the server capabilities with the aim of saving time to perform the analysis.

Analysis of a protein model

In this section we show an example of the analysis with one particular protein model. This example constitutes a typical result when using our method to test protein models. The NL-EP exhibits oscillation in the energy values through the different amino acids of the structure. In experimentally solved protein structures, all or almost all the residues have an energy below zero. In the case of protein models, some high energy residues are generally found. Figure 1 shows the NL-EP for a particular model of a human fatty acid binding protein and of the X-ray solved protein (2HMB). Practically, the energies of the whole spectrum of the native protein are lower than in the model.

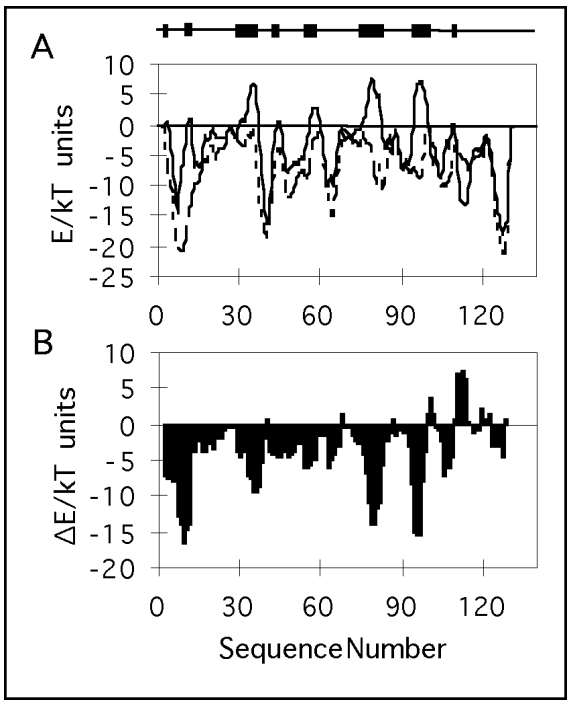

Figure 1: (**A**) Non-local energy profiles of the 2HMB model built on the 1OPA template protein structure (continuous line) and 2HMB X-ray structure (dashed line). On the top of the graph, the HEZs in the model are shown as black filled boxes. (**B**) The energy difference between the model and the X-ray structure.

The native protein has not high energy amino acids. In contrast, eight different regions involving many residues are detected with high energy in the model.

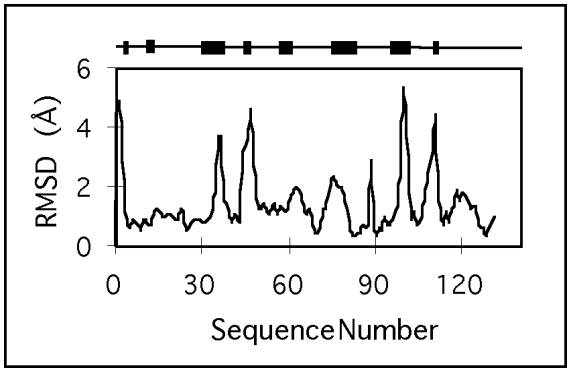

Figure 2: Graph displaying the RMSDs values between 2HMB model and 2HMB X-ray structure after superimposition of all backbone atoms. The RMSD was calculated for each individual residue considering only the backbone atoms.

When all the backbone atoms of both structures are superimposed and the RMSDs between them are calculated, most of the high variation zones are detected by the profile (Figure 2). Seven different regions are found to have a RMSD value higher than 2 Å. Six of these seven regions are accurately pointed out by the profile. In several analyzed cases, the profile is very sensitive in detecting errors and it correlates very well with the main variation zones of the native structure and of the model (1).

When the non-local contact maps of the native protein and of the model are calculated, interesting insights are found (Figure 3). First, many interactions between the different

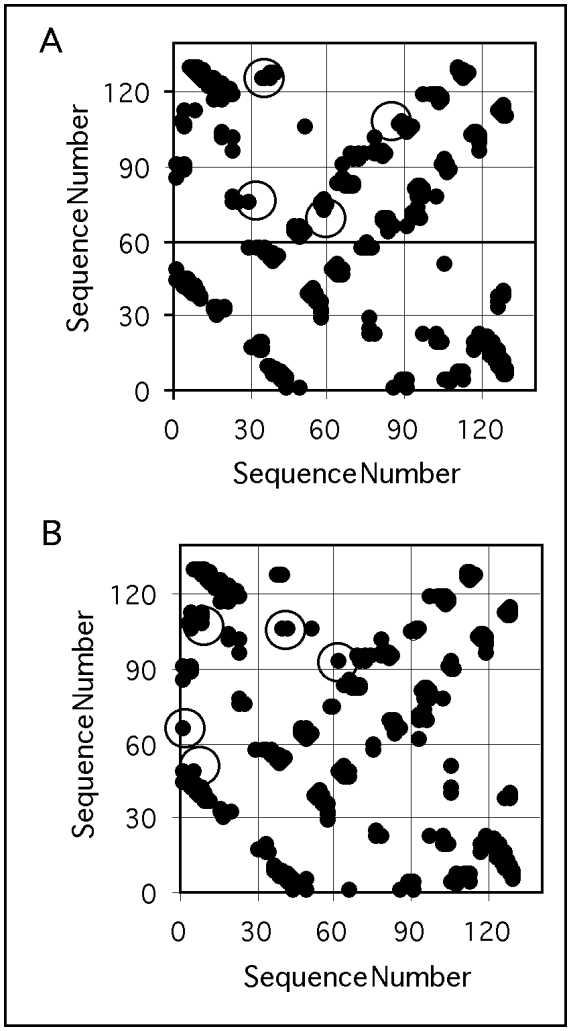

Figure 3: Non-local contact maps of 2HMB model (**A**) and 2HMB X-ray structure (**B**). The contacts in the model that are not observed in the X-ray structure and vice versa are circled.

high RMSD zones in the model are observed in the native protein. Second, some observed contacts in the native protein are not found in the model, and most of them constitute the HEZs of the model that were pointed out by the profile. Third, many contacts are observed in the model and they are not found in the native protein. Again, most of them are within the HEZs and belong to high RMSD regions.

The three-dimensional analysis of the errors in the model shows that most of them belong or are close to the regions connecting secondary structure elements in the structure (Figure 4). It is not trivial that a MFP using only pairwise energy terms can be able to detect the main errors in the model as HEZs. Also, these HEZs are not found in the native structure. The modeling of the loop conformations constitutes a big challenge in protein structure prediction, because they often represent insertions or deletions in homologue proteins. All the current methods often fail to achieve the correct loop conformation (9). Thus, it is very important to have methods able to detect where the errors
are located. This is the first requirement of a method aimed at improving the modeling of these loop regions.

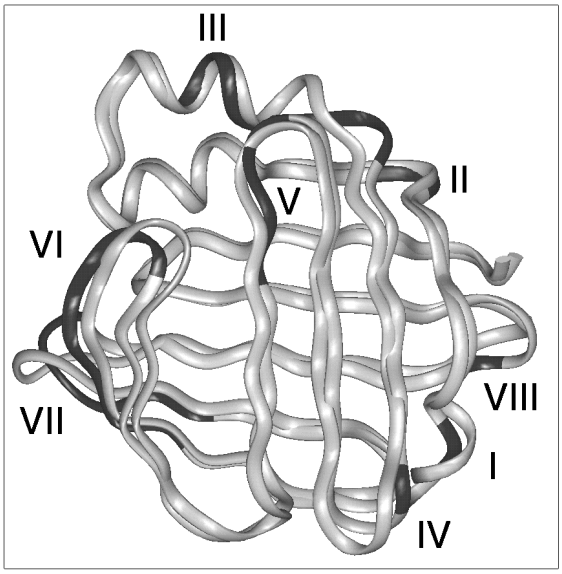

Figure 4: Picture of 2HMB model (blue) and 2HMB x-ray structure (red) after superimposition of all backbone atoms. The HEZs detected by the NL-EP in the model are shown in black. These regions are at residues 3 (I), 11-12 (II), 30-37 (III), 45 (IV), 57-60 (V), 75-82 (VI), 94-99 (VII) and 109 (VIII).

The current server implementation presented here could be very useful to assist the molecular biologist in the correct prediction of a protein structure in the last stages of the model building process. We are currently testing if the atomic level MFP is able to achieve the correct loop conformation through energy minimization simulations.

Acknowledgments

We thank Andrej Sali from the Rockefeller University, NY, USA, for providing coordinates of protein models in order to test our MFP.

References

1.- Melo, F. and Feytmans, E. (1997) Detection of high energy zones in protein structures; correlation with structural errors or with potential protein-protein interacting zones. (*Submitted*)

2.- Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86-89.

3.- Sippl, M.J. (1993) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Computer Aided Mol. Design* **7**, 473-501.

4.- Sippl, M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355-362

5.- Verkhivker, G., Appelt, K., Freer, S.T. and Villafranca, J.E. (1995) Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligan-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Engineering* **8**, 677-691.

6.- Elofsson, A., Le Grand, S.M. and Eisenberg, D. (1995) Local moves: An efficient algorithm for simulation of protein folding. *Proteins: Structure, Function, and Genetics* **23**, 73-82.

7.- Melo, F. and Feytmans, E. (1997) Novel Knowledge-based mean force potential at atomic level. *J. Mol. Biol.* **267**, 207-222.

8.- Bernstein, F.C., Koetzle, T.F., Williams, G.J.B, Meyer, E.F., Brice, M.D. Jr., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

9.- Mosimann, S., Meleshko, R. and James, M.N.G. (1995) A critical assesment of comparative molecular modeling of tertiary structures of proteins. *Proteins* **23**, 301-317.