Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Never born proteins as a test case for *ab initio* protein structures prediction

Giovanni Minervini[1], Giuseppe Evangelista[1] , Fabio Polticelli[1], Monika Piwowar[2], Marek Kochanczyk[2, 3], Lukasz Flis[3], Maciej Malawski[4], Tomasz Szepieniec[5], Zdzisław Wiśniowski[2], Ewa Matczyńska[2, 6] , Katarzyna Prymula[2, 7], Irena Roterman[2, *]

[1]Department of Biology, University Roma Tre, Viale G. Marconi 446, 00146 Rome, Italy; [2]Department of Bioinformatics and Telemedicine, Collegium Medicum - Jagiellonian University, Lazarza 16, 31-530 Krakow, Poland; [3]Faculty of Physics, Astronomy and Applied Informatics, Reymonta 4, 30-059 Krakow, Poland; [4]Institute of Computer Science AGH, Mickiewicza 30, 30-059 Krakow, Poland; [5]cademic Computer Center CYFRONET, Nawojki 11, 30-950 Krakow, Poland;[6]Faculty of Mathematics and Informatics, Jagiellonian University, Lojasiewicza 6, 30-348 Krakow, Poland; [7] Faculty of Chemistry, Jagiellonian University, Ingardena 3, 30-060 Krakow, Poland; Irena Roterman* - Email: myroterm@cyf-kr.edu.pl; * Corresponding author

received Novemeber 20, 2008; accepted November 23, 2008; published December 06, 2008

Abstract:
The number of natural proteins although large is significantly smaller than the theoretical number of proteins that can be obtained combining the 20 natural amino acids, the so-called "never born proteins" (NBPs). The study of the structure and properties of these proteins allows to investigate the sources of the natural proteins being of unique characteristics or special properties. However the structural study of NPBs can also been intended as an ideal test for evaluating the efficiency of software packages for the *ab initio* protein structure prediction. In this research, 10.000 three-dimensional structures of proteins of completely random sequence generated according to ROSETTA and FOD model were compared. The results show the limits of these software packages, but at the same time indicate that in many cases there is a significant agreement between the prediction obtained.

Keywords: *ab initio*; never born proteins; structure prediction; software; grid

Background:
Statistical considerations highlight that the number of theoretical proteins is enormously greater than the number of natural proteins. The simple combination of the 20 natural amino acids in a sequence of only 100 residues results in a number of 10^{130} possible sequences. These unknown proteins are called "never born proteins" **[1]**. The study of the properties of these proteins, especially when compared with natural proteins, may provide important information about the relationship existent between sequences and structures. Today several programs for the *ab initio* proteins structure prediction are available. They are extremely important for studying proteins of unknown crystal structure. CASP international competition **[2]** opens the opportunity to distinguish several models for their reliability of prediction. Using these programs, especially in conjunction with high throughput computing platforms as Grids **[3]**, may be the ideal solution to study problems such as NBPs, very difficult to investigate using traditional methods. On other hand NBPs represent a good test for the reliability of these programs. Many of these are in fact based on algorithms that extract the information needed to predict the three-dimensional structure of a given protein from databases of natural proteins structures. Others instead are based only on the simulation of amino acids physico-chemical properties. NBPs are proteins whose sequence is generated completely randomly and the prediction of their structure represents an extremely challenging utilization of these software packages. In this work we describe the results obtained from comparison of 10.000 predictions obtained with two different models, one for each class described above, starting from amino acid sequences with no significant homology with natural proteins.

Methodology:
Random amino acid sequences (70 amino acids long) were generated using the utility RandomBlast whose implementation has been described in detail elsewhere **[4]**. The three-dimensional structures for NBPs were predicted with the "Rosetta abinitio" software (Rosetta) **[5]** and the "Fuzzy oil drop method" (Twostage) **[6, 7]**. The RMSD was calculated only for the backbone atoms. Couples with RMSD less than 6.80 Å were subjected to energy minimization using the software CHARMM **[8]** in order to reduce the possible clashes between atoms introduced by the prediction software. The RMSD was recalculated for the resulting structure. The structure alignment and its graphic representation were obtained through the use of the CE protein structures alignment software **[9]**.

Results:
Table 1 (under supplementary material) clearly shows that RMS-D values obtained are never lower than 5.86 Å which represents the best value produced by the pair of predictions for the sequence 7033.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 177-179 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

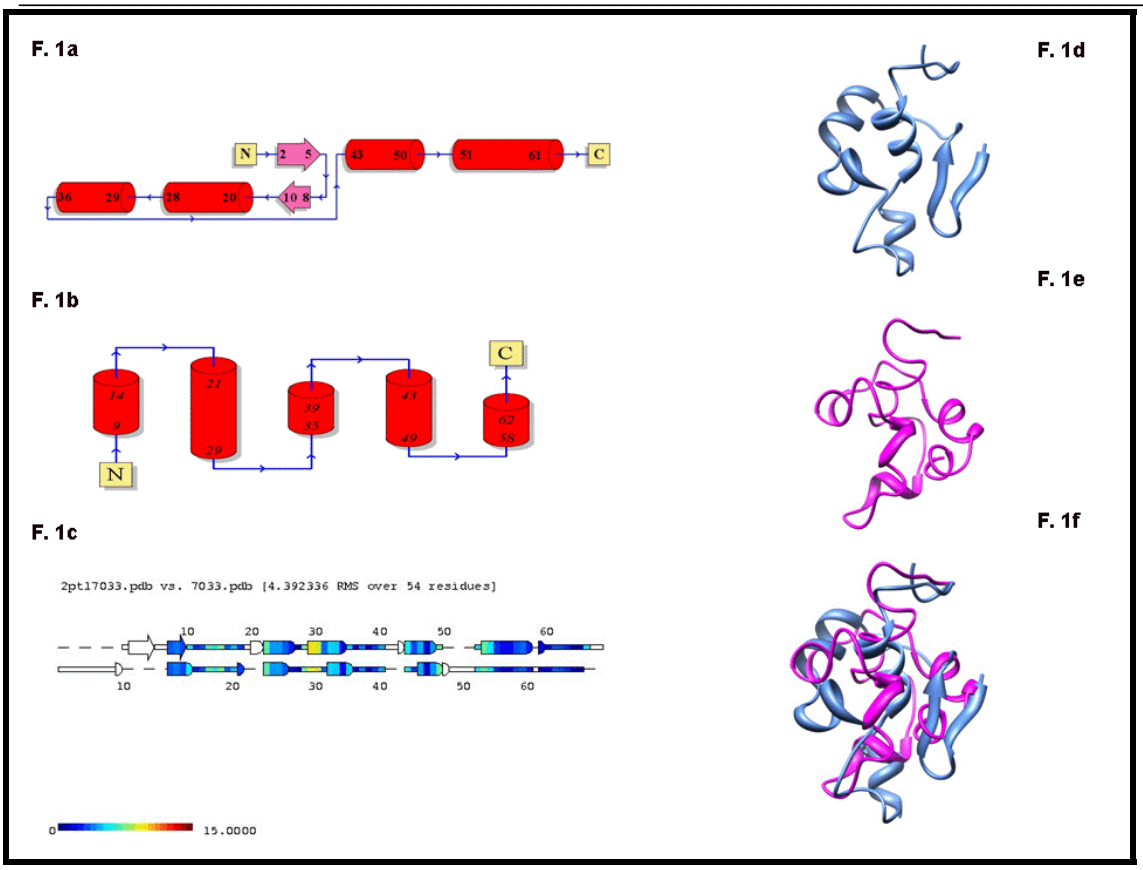

Figure 1: Schematic representation of topology maps, structural alignment and graphics representation of predicted structures for sequence n° 7033. (1a and 1b) Topology maps. (1c) Structural alignment between structures predicted with two different methods. (1d and 1e) Structures obtained with Rosetta and Two stage, respectively. (1f) Superimposition of the two proposed prediction.

Although this is not an exciting result, it is important in that highlights the challenging conditions in which the two software packages were tested. For both methods taken in consideration, the predicted structure is formed largely by alpha helices (Figure 1d and Figure 1e). The related topology maps (Figure 1a and Figure 1b) show that both predictions are characterized by the presence in the C-terminus region of two segments of alpha helix (Rosetta, residues 43-50 and 51-61; Twostage, 43-49 and 62-58), the first of which can be viewed as the central axis of the molecule. Around this central axis are two other alpha-helix regions interspersed by short loops, which seem to embrace the center with the shape of a ring (Rosetta, 20-28 and 29-36; Twostage, 21-29 and 35-39). In total disagreement however, is the prediction for the N-terminus region for which Rosetta predicts the presence of a short beta-sheet while Twostage identifies a further stretch of alpha helix (Rosetta, 5-2 and 8-12; Twostage, 9 -14). The overlap of the predicted structures relative to sequence number 7033 (Figure 1f), illustrates some agreement between the two predictions. The structural alignment (Figure 1c) clearly shows the presence of a good structural similarity and if the region of disagreement at the N-terminus is excluded, the calculated RMSD value decreases to ~4.39 Å, reinforcing the agreement between the two methods.

Conclusion:

This work highlights the difficulty to approach the structural study of proteins like the NBPs. On the other hand it is difficult to think to classic wet lab experiments to approach this problem, especially if one wants to approach the problem in its entirety. The results obtained with this research show that a computational approach shows its weakness as well, especially given the low degree of comparability of the results obtained using two different methods. However this work allows recognizing the limits and the predictive ability of *ab initio* protein structure prediction software packages. Evidence that a low number of predictions are in agreement between the two models can lead to underestimate the effective potential of these methods. At the same time, it must be noted that the programs tested in this study are in agreement in predicting that many of the sequences studied do not have a properly folded structure (data not shown). Finally, evidence that hypothetical proteins such as 7033, can be predicted with sufficient reliability opens up new scenarios of biomedical interest; in other words, from NBPs to new "possible born drugs".

Acknowledgments:

This work has been supported by the European Commission EUChinaGRID project grant (contract number: 026634).

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 177-179 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

References:

[01] C. Chiarabelli *et al.*, *Chem. Biodivers.*, 3: 840 (2006)
[02] http://predictioncenter.org/
[03] http://www.euchinagrid.org/
[04] G. Evangelista *et al.*, *Bio-Algorithms and Med-Systems*, 3: 27 (2007)
[05] C. A Rohl *et al.*, *Methods Enzymol.*, 383: 66 (2004)
[06] M. Brylinski *et al.*, *Comput. Biol. Chem.*, 30: 255 (2006)
[07] Brylinski *et al.*, *J. Biomol. Struct. Dyn.*, 23: 519 (2006)
[08] B. R. Brooks *et al.*, *J. Comp. Chem.*, 4: 187 (1983)
[09] N. Shindyalov and P. E. Bourne, *Protein Eng.*, 9: 739 (1998)

Edited by P. Kangueane

Citation: Minervini *et al.*, Bioinformation 3(4): 177-179 (2008)

Supplementary material

Structure No.	RMSD (Å)
7033	5.76
9214	5.95
435	6.34
6937	6.39
5708	6.48
5829	6.49
2300	6.57
6069	6.57
4899	6.60

Table 1: List of the first nine structures arranged for better RMS-D value.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(4): 177-179 (2008)