# *In silico* Structural Study of Random Amino Acid Sequence Proteins Not Present in Nature

by **Katarzyna Prymula**[a])[b]), **Monika Piwowar**[a]), **Marek Kochanczyk**[a])[c]), **Lukasz Flis**[a])[c]), **Maciej Malawski**[d])[e]), **Tomasz Szepieniec**[e]), **Giovanni Evangelista**[f]), **Giuseppe Minervini**[f]), **Fabio Polticelli**[f]), **Zdzisław Wiśniowski**[a]), **Kinga Sałapa**[a]), **Ewa Matczyńska**[a]), and **Irena Roterman***[a])[1])

[a]) Department of Bioinformatics and Telemedicine, Collegium Medicum – Jagiellonian University, Lazarza 16, PL-31-530 Krakow (phone and fax: +48 12 619 96 93; e-mail: iroterman@cm-uj.krakow.pl)
[b]) Faculty of Chemistry, Jagiellonian University, Ingardena 3, PL-30-060 Krakow
[c]) Faculty of Physics, Astronomy and Applied Informatics, Reymonta 4, PL-30-059 Krakow
[d]) Institute of Computer Science AGH, Mickiewicza 30, PL-30-059 Krakow
[e]) Academic Computer Center CYFRONET, Nawojki 11, PL-30-950 Krakow
[f]) Department of Biology, University Roma Tre, Viale G. Marconi 446, I-00146 Rome

The three-dimensional structures of a set of 'never born proteins' (NBP, random amino acid sequence proteins with no significant homology with known proteins) were predicted using two methods: Rosetta and the one based on the 'fuzzy-oil-drop' (FOD) model. More than 3000 different random amino acid sequences have been generated, filtered against the non redundant protein sequence data base, to remove sequences with significant homology with known proteins, and subjected to three-dimensional structure prediction. Comparison between Rosetta and FOD predictions allowed to select the ten top (highest structural similarity) and the ten bottom (the lowest structural similarity) structures from the ranking list organized according to the RMS-D value. The selected structures were taken for detailed analysis to define the scale of structural accordance and discrepancy between the two methods. The structural similarity measurements revealed discrepancies between structures generated on the basis of the two methods. Their potential biological function appeared to be quite different as well. The ten bottom structures appeared to be 'unfoldable' for the FOD model. Some aspects of the general characteristics of the NBPs are also discussed. The calculations were performed on the EUChinaGRID grid platform to test the performance of this infrastructure for massive protein structure predictions.

**Introduction.** – The search for techniques aimed at the generation of new proteins for pharmacological and biotechnological applications is widely developed nowadays [1–3]. This involves the selection of proteins of desirable activity among those present in Nature, as well as the production of new polypeptide compounds resulting from libraries of peptides with random amino acid sequences [4][5]. The final aim of these

[1]) *G. Evangelista* and *F. Polticelli* are the authors of the software for random sequences generation and selection against the Non-Redundant sequence database, *G. Minervini* is the one who actually set up Rosetta on grid and generated the predictions. *M. Malawski* and *T. Szepieniec* were responsible to set up the FOD model on the grid system. *M. Kochanczyk*, *L. Flis* prepared the program according to FOD model and introduced some modifications (the procedure avoiding the atoms overlapping) and RMS-D calculation allowing structural comparison. *K. Prymula* was responsible for monitoring the progress in calculation. *M. Piwowar* and *E. Matczyńska* are the authors of program checking the amino acid sequences. *Z. Wiśniowski* and *K. Sałapa* performed the statistical calculations. *I. Roterman* is the author of FOD model.

© 2009 Verlag Helvetica Chimica Acta AG, Zürich

studies is the selection of polypeptides with enhanced biological activity, altered catalytic properties, or higher structural stability. It can be assumed that amino acid sequences not observed in real proteins may be an abundant source of unknown biological activities, which – eventually – may be useful for biomedical applications.

An estimate of the number of possible random sequences of, for example, just 70 amino acid residues [4] leads to conclude that the existing sequences which occur in real proteins are a tiny minority of all the possible sequences. Thus it is reasonable to assume that the huge number of potential proteins characterized by sequences not observed in Nature (also known as 'never born proteins' – NBP) can be an ensemble hiding many biological functions not observed in the biochemistry developed so far by living organisms.

The experimental characterization of some NBP produced by phage display techniques and focused on structural aspects was already presented in [4][5]. In that study, the resistance to proteolytic digestion was used as a folding criterion. The high frequency of folded structures in a totally random library observed in the above mentioned study suggests that the globular folding represents a general feature of hetero-polypeptide structures. Thus the selection of corresponding properties such as specific binding and/or catalysis seems to be possible in the set of NBP.

The search for bioactive protein molecules was one of the aims of the EUChinaGRID project (EUChinaGRID – www.euchinagrid.eu). The search for biologically active NBP is performed using two different structure prediction methods: a stochastic one represented by the Rosetta *ab initio* method [6] and a heuristic one, the 'fuzzy-oil-drop' (FOD) model [7], assuming the hydrophobicity irregularity (deficiency) as the criterion for active site identification [8–10]. The proteins selected on the basis of the similarity between the structures predicted according to both methods are planned to be synthesized and analyzed by NMR as an experimental validation of the structures generated *in silico*.

As a preliminary study, before the analysis of the ensemble of NBPs, the performance of the FOD model in the identification of the active site region of a protein molecule was tested on existing proteins of defined polypeptide chain length (70 amino acid residues) [11–13]. Following this approach, the localization of the potential active site (including any ligand binding cavity) is defined on the basis of the specificity of the hydrophobicity deficiency/excess distribution in a particular protein under consideration [14][15]. In the above mentioned preliminary study, the irregularities of the hydrophobicity distribution (the basis of the FOD model) of natural proteins appeared to be highly structure/function specific [9][10]. The same technique was thus applied for the structural and functional comparative analysis of FOD and Rosetta predicted structures.

**Principle of the Methodology.** – The protein structure prediction is usually based on the internal non-bonding interactions which are optimized (global or local energy minima) to represent the structure of the protein. The natural environment is usually mimicked by the presence of water molecules. The model presented in this work (applied for protein structure prediction) assumes the presence of an external force field generating the environment for the folding polypeptide chain. The external force field is expressed by a three-dimensional *Gauss* function. The standard interpretation of this function as probability density distribution is treated as hydrophobicity density distribution.

The three-dimensional *Gauss* function is as follows:

$$Ht_j = \frac{1}{Ht_{\text{sum}}} \exp\left(\frac{-(x_j - \bar{x})^2}{2\sigma_x^2}\right) \exp\left(\frac{-(y_j - \bar{y})^2}{2\sigma_y^2}\right) \exp\left(\frac{-(z_j - \bar{z})^2}{2\sigma_z^2}\right)$$

$Ht_j$ is assumed to represent the hydrophobicity distribution in a particular grid point belonging to the protein body. The hydrophobicity maximum is located in the center of the ellipsoid and decreases in a distance-dependent manner according to the three-dimensional *Gauss* function. The mean value at which the *Gauss* function reaches its maximum is localized at the (0,0,0) point in a coordinate system. The values of standard deviation $\sigma_x$, $\sigma_y$, $\sigma_z$ calculated separately for each dimension (axis) represent the size of the drop which depends on the length of the polypeptide under consideration. The length of the polypeptide determines the size of the protein molecule and thus the size of the ellipsoid expressed by $\sigma_x$, $\sigma_y$, $\sigma_z$. The detailed analysis of the relation between the length of the polypeptide and the size expressed in the ellipsoid parameters is presented elsewhere [16].

Before the external hydrophobic force field can be defined, the protein molecule must be oriented in the space according to the following procedure:

1. The geometric center of the molecule must be localized in the center of the coordinate system.

2. The longest distance between two residues (represented by the effective atom – geometric center of side chain of the amino acid) must overlap one of the axes (say *x*-axis).

3. The molecule must be rotated around the *x*-axis to orient the longest inter-projections (on *y*,*z* plane) distance along the *y*-axis.

4. The linear size (the maximum inter-atomic distance along the *x*, *y*, and *z* axes), increased by 9 Å in each direction (the cutoff distance for hydrophobic interaction), allows the calculation of $\sigma_x$, $\sigma_y$, $\sigma_z$.

This is how the geometric parameters of the protein molecule can be interpreted according to the *Gauss* function.

The empirical (observed) distribution of hydrophobicity can be different than the idealized one. The empirical hydrophobicity distribution can be calculated according to the *Levitt* [17] function:

$$\tilde{H}o_j = \frac{1}{\tilde{H}o_{\text{sum}}} \sum_{i=1}^{N} H_i^r \begin{cases} \left[1 - \frac{1}{2}\left(7\left(\frac{r_{ij}}{c}\right)^2 - 9\left(\frac{r_{ij}}{c}\right)^4 + 5\left(\frac{r_{ij}}{c}\right)^6 - \left(\frac{r_{ij}}{c}\right)^8\right)\right] & \text{for} \quad r_{ij} \le c \\ 0 \quad \text{for} \quad r_{ij} > c \end{cases}$$

where $Ho_j$ represents the empirical hydrophobicity value characteristic for the position of the *j*-th grid point, $H_i^r$ represents the hydrophobicity characteristic of the *i*-th amino acid, $r_{ij}$ is the distance between the *j*-th grid point (its hydrophobicity is equal to 0.0) and the *i*-th effective atom in the amino acid, and *c* expresses the cutoff distance, which has a fixed value of 9.0 Å following the original paper. The grid point collects the hydrophobicity interaction in its close neighborhood (9 Å). $Ho_{\mathrm{sum}}$ represents the sum of all the grid points hydrophobicity. Any hydrophobicity scale available in literature may be applied to calculate the observed hydrophobicity density [18–22].

The grid system mimicking the environment is defined with the constant step size (detailed information can be found in [7]).

Since both values are standardized (the coefficient $1/H_{\mathrm{sum}}$), the differences between theoretical and empirical values expressing hydrophobicity density in a particular point of space can be calculated according to:

$$\Delta\tilde{H}_j = Ht_j - Ho_j$$

The $\Delta\tilde{H}_j$ values measure the discrepancies between expected (theoretical) and observed (empirical) hydrophobicity distribution. The lower the difference the more the hydrophobic residues are buried in the central part of the globule and better is the exposure of hydrophilic residues on the surface of the protein body.

The protein folded in high accordance with the idealized 'fuzzy oil drop' distribution satisfies the condition to be very well soluble in water solution. In consequence such molecule is unable to represent any biological function understood as the tendency to interact with other molecules (ligand, proteins). This is why the discrepancies between these two distributions are observed. Some of them are identified as aim-oriented (or function-oriented).

Since the *Gauss* function is of continuity character, $\Delta\tilde{H}_i$ values when calculated for the positions of effective atoms (averaged position of side chain) reveal the specific characteristics of each residue with respect to the hydrophobicity distribution.

The profile of $\Delta\tilde{H}_i$ (expressing the value of difference for each amino acid) reveals some maxima, which are related to hydrophobicity deficiency. The hydrophobicity deficiency ($\Delta\tilde{H}_i > 0$) seems to represent a potential binding site. The potential ligand may adhere in this area as the complementary element compensating the hydrophobicity deficiency and yielding a regular smoothed hydrophobicity distribution. Negative $\Delta\tilde{H}_i$ values represent areas of higher hydrophobicity than expected. Areas with such characteristics, when localized on the surface of protein, seem to represent potential areas responsible for protein-protein complex creation.

The protocol of the folding simulation according to the presented model is as follows:

1. Orientation of the molecule as described above.

2. Energy (internal non-bonding interactions) minimization procedure is performed.

3. Optimization procedure aimed at obtaining hydrophobicity distribution accordance between observed and expected hydrophobicity distribution is performed.

The three steps are repeated iteratively until the convergence level reaches the expected value. Each iteration is performed for a smaller size of the 'drop' (decrease of $\sigma_x$, $\sigma_y$, $\sigma_z$ values) until the size reaches the volume appropriate for the particular polypeptide chain length (relation between number of amino acids in the chain and the volume of the protein in its native form is presented in [16]).

Step 2 of the presented procedure is a sort of relaxation due to the absence of any external constrains, while step 3 has the character of a squeezing step pushing the molecule to increase its density (the starting structure displays a very low packing density) [23].

A molecule folded according to the presented model produces the molecule with hydrophobic residues buried entirely in the interior of the molecule and hydrophilic residues exposed on the surface. A molecule with these characteristics is perfectly well soluble although deprived of any kind of biological activity (understood as the tendency to interact with ligands, substrates, and proteins).

The $\Delta\tilde{H}_j$ profile (calculated for points representing the effective atoms) reveals the discrepancies which in some proteins appeared to be of aim-oriented form, ensuring a high specificity *versus* the potential interacting (complexing) molecules [9][10]. This is why the form of the final $\Delta\tilde{H}$ profiles for proteins folded *in silico* presented in this article are taken as the criteria for biological function prediction, at least to the extent of identification of potential protein surface areas ready to interact with other molecules.

The described model was applied for protein folding simulation as well as for potential biological function recognition of molecules under consideration. The $\Delta\tilde{H}$profiles obtained for proteins discussed in this work are compared with profiles of other proteins available in the PDB (proteins of around 70 amino acids length) [11–13]. The characteristics of their biological activity with respect to $\Delta\tilde{H}$profiles was taken as the basis for possible biological activity recognition of 'never born proteins' at least limited to the recognition of areas on the protein surface ready to interact with other molecules.

The presented FOD model oriented on the active site identification belongs to the tools available online [24–36]. The detailed comparison of FOD model in respect to Sumo [37][38] and ProFunc [39] was given in [8]. Available methods are of stochastic character while the FOD model is rather of heuristic character and is possible to be applied individually for any protein molecule.

**Results.** – The comparative analysis presented in this paper is focused on the structural and functional comparison of structures generated using Rosetta and the FOD model to discuss the possible bioactivity of the proteins under consideration. The presence of hydrophobicity irregularities in the proteins generated in this work allows the identification of putative active sites also through comparison with the hydrophobicity profiles of real proteins [4][5].

*Structure Comparison.* The RMS-D values were calculated for all proteins (objects) for final structures selected according to the procedure described in the *Exper. Part.* The selected proteins are presented in *Table 1.* The additional optimization procedure (*in vacuo* and in water solvent) was applied for all the discussed structures folded according to FOD model. The final results measuring the structural similarity are shown in *Table 2.*

The RMS-D values are quite large for proteins of this small size (70 amino acids). It must be emphasized that the sequences under consideration are highly peculiar. There is no structural database for this kind of sequences. The highest values of RMS-D (R structures) are due to the fact that these polypeptides appeared to be unfoldable for the FOD model (S structures).

*Secondary Structure Identification.* Secondary structure content of the final predicted structures using DSSP program [40] is given in *Table 3.*

Table 1. *Amino Acid Sequences of the Proteins Selected for Analysis*

| Sequence ID | Sequence |
|---|---|
| 102 | GGNIQNDYIGVETGGVQSMQPHVFAVRPYPGETQAIARNQQGVNRDQTCVCPTTCMNGGDMCPMPTSNYN |
| 372 | HDACGGEDRPDVCLEPTHEHAPMAICRLKFRSTTSDFMKWGYFWLPSPSLLSLTTWRKTIKRFVIYHHSM |
| 386 | SECVEGVKTFFKFCRNARHVGTEQDQPCVHSSPHLIYPDHLEQGTILKDWNVWKYCFMVFDIGAGWSDRY |
| 435 | HKHLFAYNYMSDHTQRFRSAQYTICTSMFVNDNRPLLNDAPFEYLHWYSFLFFMCLHKDCTPLKRYFEQC |
| 438 | IKGLYSTNMKEGVMNLKDAKQHYERDKAESMTRFCEYIQIACVQAPHIWPFNTSYLFCGQKWQTRDGMIL |
| 595 | DGCGCLEPMFDYIHFDRAFDTKFVGITWVADLRQWSGHLCTYAELNRPTCTGEDQVCQCDVNHGRIPCIK |
| 913 | NRNKTEIEWHTCKIAWNCQLHKDDAPGIFMGHTYSSNANGYECPKMQLRCRTAYAQYMHLQGQFCQNPND |
| 1000 | RNALFDGIPTVYCWTLADSQWWACYQYRALCCIKGRCVFERITDSYVRMVTKLIRRFGYANPHPFNCECT |
| 1056 | KMHLDAIESKYWHVPTQTVNDSALFAPTQEMLAPSSSVYYLLINMSRSYHEFLVRVKKPMEDEACNQCVA |
| 1134 | TCSDDRVPSHQTDAFNQHQFITLRLWFDFFWYRKRMHTGVSARSDEDNGRSCSNQWSDDMSGCRWQQDCY |
| 1167 | VPARGFWLGHQPIVWWHDCTYWTPPLLLASWFEWCIGVCRKSLSAWVNTVELYIKEETVPKYWTVASEPH |
| 1281 | GDTPRQFWQWWQDQGNHMEDDDYHPDYCHHGKGLKLSKALPPPIAEEIIWDEAAYSPLSPVRQGGYQKKC |
| 1349 | GVGDSACCHMNASTPNWEMVHKWWHCKKDTPRICNTIFAMLTQLLQWNQLPWQRLQFSQWEIWMHCWNMV |
| 1356 | WMPQCKVHDGYDCSIMFAHKNPLYQYKAYMANEAAVPRRRTEQCCQYGQGYWETMHDPMTMMHKHLGKHA |
| 1570 | CGNVYMFDVCIDHDWDQTDHIMWQLGKYNGCCNPHFHEWSEWYPFFFFLLAVADCRTGVWLNQLDFTRKP |
| 1736 | YIDFRLSCCLGGQCWSFMYQWIQTFCRSASSLWMAWVQCFNVIVVINPWYMYYTQCRYCMCCDVYHCGQS |
| 2265 | WTSEGDFSWLDAFYWCKKMWQFVSDFPHHAQVEVNFQPWEARIRWHSDFKALQGKMPGNHWHGYTRCPMQ |
| 2300 | MDQSSDSLEVNWEDSLQVGTWGDIDLKLRMNFSWWCLKFWMNQTGVNASNSTGSHDGICHIRMRFSCHWW |
| 2748 | TWYIRMTGSLDFLTDDKRFRQTMKQDTMQPDHKILWKQPINYARIIEANLAKEWFIREYNHNMQSWENGT |
| 3208 | SMSWLFTADGFNMNSPENIVWMANLIAGCKWRNQMPVNQPIDCATKQMADQLETFQPPNSFMNLCSIYEC |

Interestingly, the $\beta$-structure is underrepresented in S forms (FOD-calculated structures), although the content of the $\beta$-structure in R form (Rosetta calculation) is the lowest one in comparison to other secondary structural forms. A similar $\alpha$-helical percentage was found in structure 3208, a very high percentage was found in both approaches for structure 1570. A similar percentage for random coil content was found for structure 2748. The percentage of turns appeared to be similar in 2300, 1281, 1167, 386, and 372.

*Active Site Recognition.* The identification of areas of high discrepancy *versus* the idealized hydrophobicity distribution is assumed to express the functional specificity (highly positive $\Delta\tilde{H}_j$ values – area of hydrophobicity deficiency – are assumed to indicate areas ready to interact with a specific ligand with hydrophobicity distribution complementary to the hydrophobic cavity; highly negative $\Delta\tilde{H}_j$ values – higher than expected hydrophobicity – if occurring on the surface of the protein, are assumed to indicate areas for potential protein–protein interactions).

The $\Delta\tilde{H}_j$ profiles for selected structures are given in *Fig. 2*, and 3-D representation of the hydrophobicity irregularity distribution over the molecules under consideration is presented in *Fig. 3*.

The profile of $\Delta\tilde{H}_j$ observed for molecule 1000 presented in *Fig. 2* characterized by the SE parameters represents the example of lowest difference between R and S structure. Additionally the analysis of SE parameters for *in vacuo* and in water simulation (after FOD folding based simulation and additional *in vacuo* and water optimization without any constraints) show that the influence of environment is negligible in this case. The structural changes resulted from the additional optimization

Table 2. *RMS-D Values Calculated for Backbone Atoms.* Simple FOD simulation is compared to Rosetta (*d*). Influence of correction procedure is shown in: *a*) original FOD structure *vs.* additional optimization Amber *in vacuo*, *b*) original FOD structure *vs.* optimization in explicit water, and *c*) relative change between *in vacuo* and Amber with Amber in water environment. Additional energy minimization of FOD structures did not significantly influence the similarity between Rosetta and FOD structures: *d*) simple FOD simulation *vs.* Rosetta, *e*) FOD with Amber *in vacuo* correction *vs.* Rosetta, *f*) FOD with Amber in water correction *vs.* Rosetta.

| ID | *a*) Opt. *in vacuo vs.* FOD | *b*) Opt. in water *vs.* FOD | *c*) Change between *in vacuo* and water | *d*) FOD *vs.* Rosetta | *e*) FOD (*in vacuo*) *vs.* Rosetta | *f*) FOD (in water) *vs.* Rosetta |
|---|---|---|---|---|---|---|
| 102 | 1.642 | 1.418 | 1.283 | 7.562 | 7.773 | 7.723 |
| 372 | 1.193 | 1.194 | 1.468 | 12.139 | 12.373 | 12.326 |
| 386 | 1.277 | 1.226 | 1.254 | 23.139 | 23.163 | 23.437 |
| 435 | 1.685 | 1.344 | 1.619 | 10.140 | 10.381 | 10.479 |
| 438 | 1.432 | 1.558 | 1.313 | 28.222 | 28.600 | 28.240 |
| 595 | 1.632 | 1.378 | 1.448 | 9.687 | 9.865 | 10.209 |
| 913 | 1.364 | 1.281 | 1.458 | 14.866 | 14.788 | 15.159 |
| 1000 | 1.568 | 1.362 | 1.408 | 7.077 | 7.115 | 7.464 |
| 1056 | 1.393 | 1.445 | 1.335 | 7.848 | 7.923 | 7.916 |
| 1134 | 1.691 | 1.483 | 1.552 | 25.990 | 26.240 | 25.628 |
| 1167 | 1.748 | 1.820 | 1.334 | 19.861 | 18.459 | 19.082 |
| 1281 | 1.339 | 1.515 | 1.401 | 8.955 | 9.370 | 9.390 |
| 1349 | 1.174 | 1.392 | 0.989 | 7.493 | 7.895 | 7.865 |
| 1356 | 1.287 | 1.143 | 1.185 | 7.628 | 7.637 | 7.529 |
| 1570 | 1.610 | 1.754 | 1.363 | 15.409 | 15.347 | 15.367 |
| 1736 | 1.438 | 1.298 | 0.998 | 20.011 | 19.897 | 19.787 |
| 2265 | 1.441 | 1.530 | 1.265 | 31.807 | 31.640 | 31.614 |
| 2300 | 1.706 | 1.551 | 1.162 | 6.693 | 6.815 | 6.693 |
| 2748 | 1.456 | 1.287 | 1.189 | 18.899 | 18.363 | 18.654 |
| 3208 | 1.537 | 1.510 | 1.595 | 23.033 | 23.023 | 23.123 |

procedure changed the RMS-D value (the spatial positioning of particular residues) keeping the general distribution of hydrophobicity irregularity conserved.

The similarity/differences between $\Delta\tilde{H}_j$ profiles for R *versus* S form can be measured quantitatively by applying the correlation between $\Delta\tilde{H}_j$ R and $\Delta\tilde{H}_j$ S measurement. The correlation coefficients (calculated according to *Pearson* [41] and *Spearman* [42]) are given in *Table 4*.

The highest similarity of hydrophobicity irregularity in both structures S and R was found for proteins 1281, 2300, 595, and 372. Taking into account the $\Delta\tilde{H}_j$ characteristics of proteins of similar polypeptide chain length deposited in the Protein Data Bank (PDB) [11–13][43], for the structures generated according to the FOD model, the functional characteristics presented in *Table 5* can be hypothesized. The potential binding sites of more than 400 proteins (of 70 amino acid residues in their sequence) present in PDB were previously analyzed [11–13] and treated as a database for the analysis discussed in this article. The analogy may be treated only qualitatively, taking into account the information entropy (SE) values and $\Delta\tilde{H}_j$ profiles similarity and subjective visual interpretation.

Table 3. *Percentage of Secondary Structures in the Proteins under Consideration*

| Sequence ID | Rosetta (R structures) | | | | FOD (S structures) | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | Turn | RC[a]) | $\alpha$ | $\beta$ | Turn | RC[a]) |
| 102 | 52.9 | 5.7 | 21.4 | 20 | 12.9 | 0 | 48.6 | 38.6 |
| 372 | 41.4 | 5.7 | 30 | 22.9 | 30 | 2.9 | 32.9 | 34.3 |
| 386 | 60 | 2.9 | 14.3 | 22.9 | 17.1 | 0 | 14.3 | 68.6 |
| 435 | 42.9 | 8.6 | 25.7 | 22.9 | 27.1 | 0 | 38.6 | 34.3 |
| 438 | 37.1 | 2.9 | 38.6 | 21.4 | 62.9 | 0 | 14.3 | 22.9 |
| 595 | 42.9 | 8.6 | 27.1 | 21.4 | 24.3 | 0 | 37.1 | 38.6 |
| 913 | 52.9 | 5.7 | 22.9 | 18.6 | 21.4 | 2.9 | 42.9 | 32.9 |
| 1000 | 44.3 | 0 | 22.9 | 32.9 | 17.1 | 2.9 | 42.9 | 37.1 |
| 1056 | 38.6 | 11.4 | 31.4 | 18.6 | 25.7 | 0 | 47.1 | 27.1 |
| 1134 | 51.4 | 0 | 30 | 18.6 | 25.7 | 0 | 37.1 | 37.1 |
| 1167 | 51.4 | 5.7 | 21.4 | 21.4 | 28.6 | 0 | 25.7 | 45.7 |
| 1281 | 20 | 8.6 | 31.4 | 40 | 41.4 | 0 | 35.7 | 22.9 |
| 1349 | 60 | 8.6 | 15.7 | 15.7 | 21.4 | 0 | 44.3 | 34.3 |
| 1356 | 25.7 | 8.6 | 44.3 | 21.4 | 42.9 | 0 | 34.3 | 22.9 |
| 1570 | 71.4 | 5.7 | 8.6 | 14.3 | 55.7 | 0 | 20 | 24.3 |
| 1736 | 65.7 | 5.7 | 18.6 | 10 | 18.6 | 0 | 24.3 | 57.1 |
| 2265 | 30 | 20 | 31.4 | 18.6 | 24.3 | 0 | 22.9 | 52.9 |
| 2300 | 32.9 | 5.7 | 37.1 | 24.3 | 15.7 | 0 | 35.7 | 48.6 |
| 2748 | 35.7 | 5.7 | 37.1 | 21.4 | 61.4 | 0 | 17.1 | 21.4 |
| 3208 | 41.4 | 5.7 | 37.1 | 15.7 | 47.1 | 0 | 21.4 | 31.4 |

[a]) RC = random coil.

*Entropy Scale.* The entropy scale (SE, $SE_{max}$, $SE_{rel}$ and I) calculated for maxima (+) and minima (−) distribution along the polypeptide chain under consideration is presented in *Table 6*.

The highest similarity between S and R forms was found for structure 1570, according to all SE based parameters. The lowest similarity was found for the structure 2265 taking SE parameters as criteria for comparison. For each sequence pair-wise alignments with all other sequences, using the LALIGN program from the FASTA package [44] version 35 was performed (scoring matrix BLOSUM50 and gap open/extension penalties equal to $-10/-2$ ). The lowest similarity to other sequences is exhibited by sequence 1056 (lowest maximum *Waterman-Eggert* score equals to 37, lowest number of results obtained – detected similarities). The highest sequence similarity characterizes sequences 1349 and 1736 (*Waterman-Eggert* score equals to 71). Maximum similarity scores and scores for alignment of two identical sequences are given in *Table 7.*

*The High RMS-D Structures.* The high values of RMS-D were obtained due to structures S, which appeared as 'unfoldable' with a high percentage of random coil structural forms and very elongated forms.

The structure of a real protein present in PDB, a dimer of two chains of 70 amino acids each (PDB code 2NWT), represents a structure with a highly unfolded form (*Fig. 4*). The $\Delta\tilde{H}_j$ profile for this protein displays only one fragment of positive $\Delta\tilde{H}_j$ values. The SE parameters appeared to be very similar to those calculated for S structures of high RMS-D values (measuring the discrepancy between R and S forms,

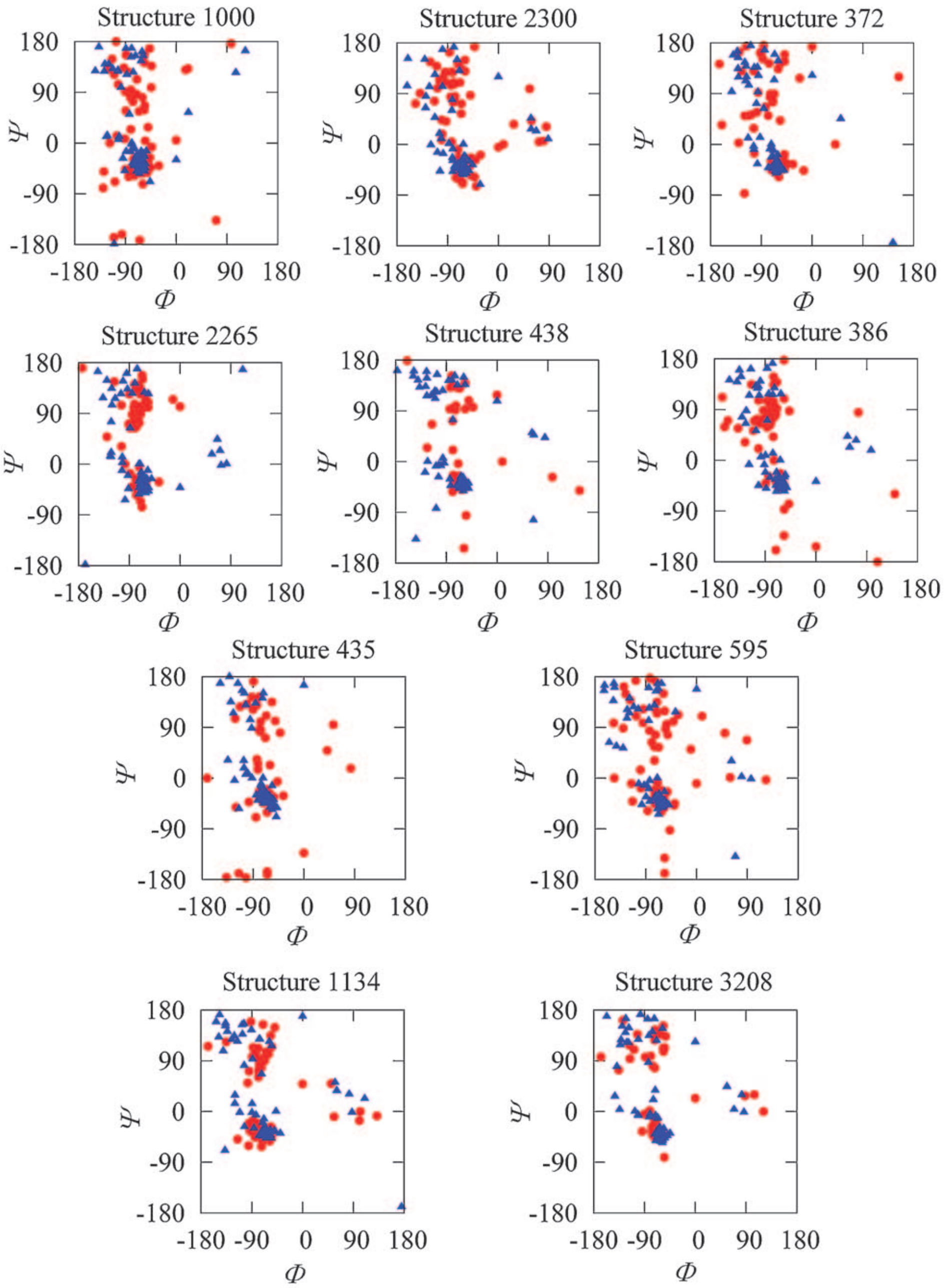

Fig. 1. *Φ,Ψ Angles distribution on the* Ramachandran *map for selected R and S structures.* Red circles: S structures; blue triangles: R structures.

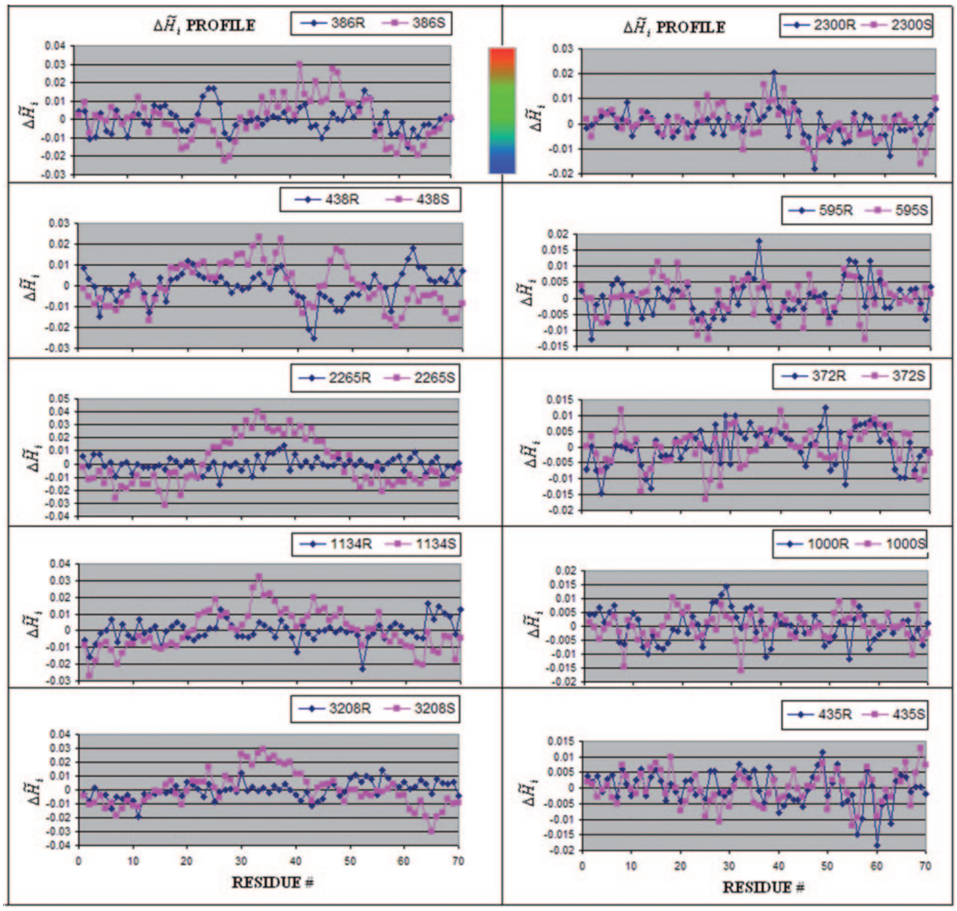

Fig. 2. *ΔH̃ Profiles for selected molecules*

see *Table 8*). According to $\mathrm{SE}_-$ the closest molecule to 2NWT is 2748S, according to $\mathrm{I}_+$ the 438S, according to $\mathrm{SE}_-$ the structures 386S and 438S, and according to $\mathrm{I}_-$ the structure 1736S.

De novo *Designed Proteins.* Another protein of 70 amino acid residues present in the PDB fits well with the subject of this article. This is a synthetic construct (PDB code 2AVP, chain A, *Table 9*) and is a very good example for the discussion focused on the expected binding sites (see *Fig. 5*).

This protein appears to be similar to proteins generated and presented in this work: according to $\mathrm{SE}_+$ of the protein 2300S, according to $\mathrm{I}_+$ of the proteins 435S, 1056R, 1000S, 3208R, 1356R, 1570R, and 372R, according to $\mathrm{SE}_-$ of the proteins 372R, 1134R, and 2300R, and according to $\mathrm{I}_-$ of the proteins 1000R and 435 R.

The 2AVP appeared to be most similar for the following proteins: 372, 1056, 435 taking the SE parameters as criterion for comparison, although only experimental verification could measure the degree of structural similarity.

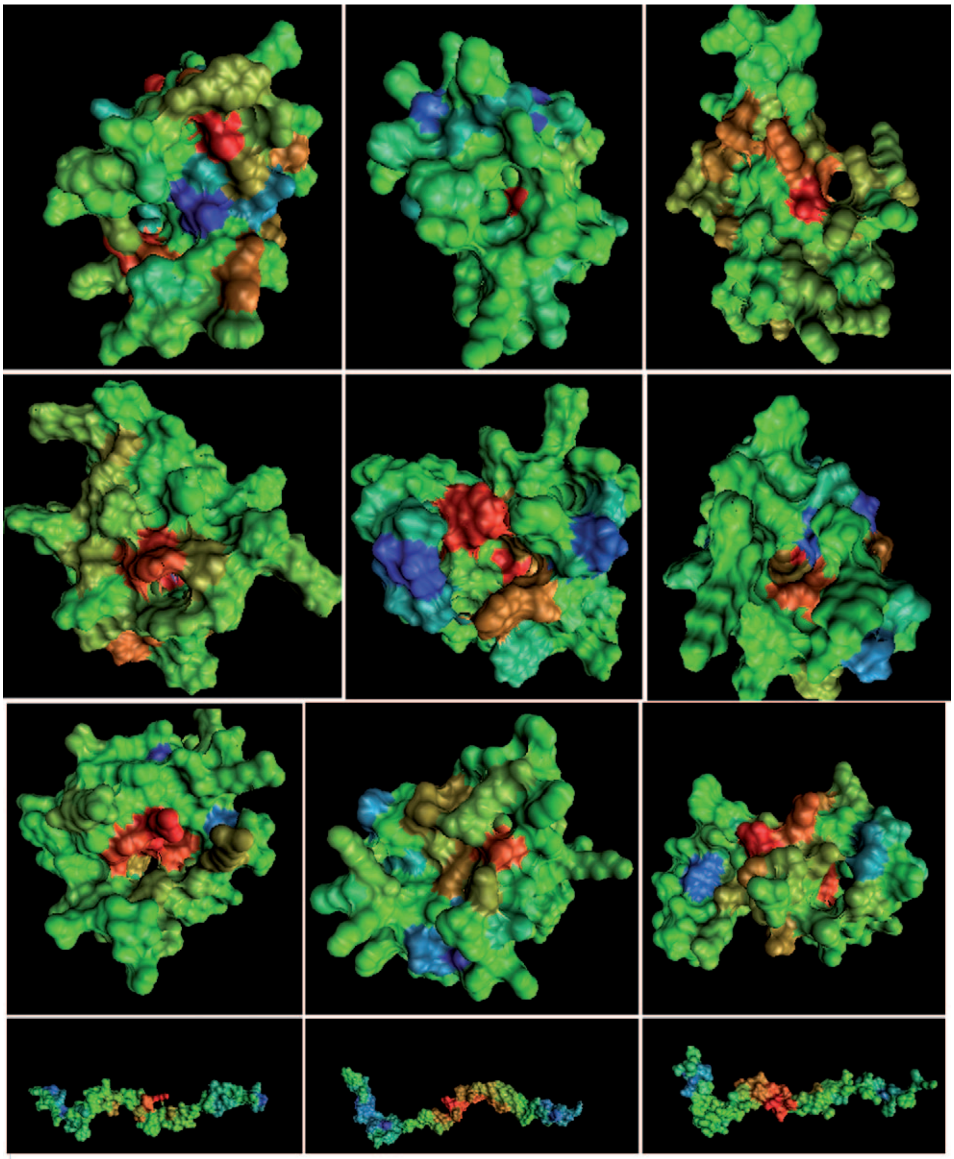

Fig. 3. *3D Presentation of selected structures.* Similar $\Delta\tilde{H}_j$ profiles (from left to right: 1000, 2300, 435): first row, forms R; second row, forms S. Different $\Delta\tilde{H}_j$ profiles (from left to right: 1134, 2265, 3208): third row, forms R; fourth row, forms S.

The proteins with 70 amino acid residues in the polypeptide chain categorized according to their biological function (as it is given in PDB) were taken as control group (P group) for comparative analysis with the R and S forms of NBP. These groups were: ribosomal proteins, RNA/DNA binding proteins, metal binding proteins, hem

Table 4. *Correlation between $\Delta\tilde{H}_j$ Values as Predicted in S and R Structures*. The statistically significant correlations are presented in bold character.

| Sequence ID | | Normal distribution | *Pearson* | *Spearman* |
|---|---|---|---|---|
| 102 | R | yes | **0.28644** | **0.364955** |
| | S | yes | | |
| 372 | R | yes | **0.46118** | **0.509859** |
| | S | yes | | |
| 386 | R | yes | **0.37244** | **0.407821** |
| | S | yes | | |
| 435 | R | no | | **0.487219** |
| | S | yes | | |
| 438 | R | no | | 0.204689 |
| | S | yes | | |
| 595 | R | yes | **0.45341** | **0.453396** |
| | S | yes | | |
| 913 | R | yes | – 0.1479 | – 0.133024 |
| | S | yes | | |
| 1000 | R | yes | | **0.383921** |
| | S | no | | |
| 1056 | R | yes | | **0.343190** |
| | S | no | | |
| 1134 | R | no | | 0.147826 |
| | S | yes | | |
| 1167 | R | yes | **0.323134** | **0.265611** |
| | S | yes | | |
| 1281 | R | yes | **0.51589** | **0.501986** |
| | S | yes | | |
| 1349 | R | yes | **0.41180** | **0.432211** |
| | S | yes | | |
| 1356 | R | yes | **0.28936** | **0.280203** |
| | S | yes | | |
| 1570 | R | yes | 0.16194 | 0.101146 |
| | S | yes | | |
| 1736 | R | yes | | 0.027242 |
| | S | no | | |
| 2265 | R | yes | | 0.221555 |
| | S | no | | |
| 2300 | R | no | | **0.489039** |
| | S | yes | | |
| 2700 | R | yes | | **0.254343** |
| | S | no | | |
| 3208 | R | yes | 0.15917 | 0.158044 |
| | S | yes | | |

binding proteins (cytochromes), antibiotics, toxins, enzymes (EC 2.7.7.6), growth factors, serine protease inhibitors, antifreeze proteins, chaperones, and proteins of unknown function. The two additional groups: R and S of proteins of structure generated according to Rosetta and FOD model were taken for comparative analysis using all SE parameters as criterion for comparison.

Table 5. *The Short Characteristics of the Final 20 Structures Generated According to the R and the S Model.* The short characteristics of possible biological activity are suggested according to the basis of the observation of real proteins of 70 amino acid residues in polypeptide chains.

| Sequence ID | R Form | S Form |
|---|---|---|
| 435 | Few dispersed cavities open too much to bind ligand | Quite good cavity for ligand binding |
| 2300 | Antifreeze-like protein | Quite good ligand binding cavity |
| 595 | Antifreeze-like protein | Quite good cavity to bind ligand |
| 372 | Quite good ligand binding cavity | Quite good ligand binding cavity |
| 1000 | Highly dispersed hydrophobic deficiency/excess | Quite well-defined ligand binding cavity |
| 102 | Highly hydrophobic deficiency area on the surface | Quite well-defined ligand binding cavity |
| 1349 | Quite well-defined ligand binding cavity | Quite well-defined ligand binding cavity |
| 1356 | Antifreeze-like protein | Quite well-defined ligand binding cavity |
| 1056 | Antifreeze-like protein | Antifreeze-like protein |
| 1281 | Few ligand binding cavities | One ligand binding cavity |

*Statistical Analysis of the R and S Forms With Respect to Real Proteins with 70 Amino Acid Residues in Polypeptide Chain.* Only results which appeared to verify the significant differences are shown and discussed here.

All significant differences found according to the statistical analysis point out the R group as significantly different *versus* P and S group. The $SE_+$, $SE_{+max}$, and $L_+$ (number of fragments of positive $\Delta\tilde{H}_j$), $SE_{-max}$, and $L_-$ (number of fragments of negative $\Delta\tilde{H}_j$) parameters appeared significantly different for the R group *versus* the S and P groups according to the non-parametric test (*Wilcoxon* and *Kruskal-Wallis*, *Table 10*).

One way variance analysis (ANOVA) for $I_+$ (the distribution of which appeared to be of normal character) and $I_-$ verified the R group as also significantly different with respect to the S and P groups (*Table 11* and *Fig. 6*). The general characteristic of all three groups is given in *Table 12*.

According to the statistical analysis, the group of proteins R appeared to be significantly different analyzing the SE parameters. Higher values of SE parameters mean significantly higher differentiation in the sense of the distribution of fragments of positive and negative $\Delta\tilde{H}_j$ values. This was seen also in the discussion of possible biological activity. The surface of R proteins was covered by a significantly higher number of areas of opposite hydrophobicity character. The relation of areas of positive and negative $\Delta\tilde{H}_j$ characteristics in natural proteins (group P) and in proteins S (generated according to 'fuzzy oil drop' model) appeared similar.

**Discussion.** – In this article, a large data base of random 70 amino acid residues long protein structures, generated according to two different structure prediction/folding simulation methods, was analyzed with respect to structural similarity and possible functional characteristics, taking the SE parameters as the criterion for comparison. Taking this analysis into account, some characteristics pointing out a potential bioactivity for some members of the data set were presented in *Table 5*.

Table 6. *SE Parameters Describing the Structural/Functional Specificity of the Objects.* The first (upper) values correspond to the R model, the lower values to the S model: *in vacuo* and in water solution. All values are in bits.

| Sequence ID | $\mathrm{SE}_{+}$ | $\mathrm{SE}_{+\mathrm{max}}$ | $\mathrm{SE}_{+\mathrm{rel}}$ | $\mathrm{I}_{+}$ | $\mathrm{SE}_{-}$ | $\mathrm{SE}_{-\mathrm{max}}$ | $\mathrm{SE}_{-\mathrm{rel}}$ | $\mathrm{I}_{-}$ |
|---|---|---|---|---|---|---|---|---|
| 1000 | 3.071 | 3.907 | 0.214 | 63.608 | 3.200 | 3.807 | 0.159 | 55.904 |
| | 3.011 | 3.585 | 0.160 | 31.503 | 3.142 | 3.459 | 0.091 | 41.142 |
| | 3.275 | 3.322 | 0.158 | 39.900 | 3.063 | 3.460 | 0.114 | 44.346 |
| 2300 | 3.231 | 4.087 | 0.209 | 81.805 | 3.443 | 4.087 | 0.157 | 67.001 |
| | 2.072 | 3.459 | 0.401 | 28.050 | 2.662 | 3.321 | 0.199 | 41.082 |
| | 2.324 | 3.000 | 0.225 | 30.351 | 1.382 | 3.000 | 0.539 | 26.363 |
| 372 | 2.625 | 3.700 | 0.290 | 42.854 | 3.354 | 3.807 | 0.120 | 53.900 |
| | 2.999 | 3.585 | 0.163 | 50.782 | 3.114 | 3.459 | 0.010 | 35.855 |
| | 3.046 | 3.585 | 0.150 | 43.761 | 3.165 | 3.460 | 0.080 | 35.801 |
| 435 | 3.907 | 3.907 | 0.076 | 50.372 | 2.969 | 3.907 | 0.240 | 59.562 |
| | 2.804 | 3.170 | 0.115 | 32.144 | 2.509 | 3.170 | 0.208 | 28.818 |
| | 2.913 | 3.585 | 0.187 | 39.183 | 3.105 | 3.585 | 0.134 | 36.668 |
| 595 | 2.935 | 3.700 | 0.207 | 52.263 | 3.004 | 3.585 | 0.162 | 43.924 |
| | 2.287 | 2.807 | 0.185 | 24.772 | 1.976 | 2.585 | 0.236 | 14.669 |
| | 2.481 | 3.000 | 0.173 | 29.061 | 2.037 | 2.807 | 0.274 | 20.633 |
| 102 | 2.652 | 3.700 | 0.283 | 23.169 | 2.634 | 3.585 | 0.265 | 41.950 |
| | 2.806 | 3.585 | 0.217 | 45.020 | 3.116 | 3.585 | 0.131 | 42.656 |
| | 2.863 | 3.585 | 0.131 | 42.656 | 2.863 | 3.459 | 0.172 | 38.706 |
| 1056 | 2.817 | 3.585 | 0.214 | 39.822 | 2.755 | 3.585 | 0.231 | 53.033 |
| | 2.867 | 3.459 | 0.171 | 38.990 | 2.867 | 3.322 | 0.137 | 38.307 |
| | 2.745 | 3.459 | 0.206 | 33.664 | 3.036 | 3.322 | 0.086 | 35.910 |
| 1281 | 2.729 | 3.907 | 0.301 | 54.508 | 3.170 | 4.000 | 0.207 | 55.488 |
| | 2.912 | 3.590 | 0.158 | 38.128 | 2.917 | 3.459 | 0.446 | 38.984 |
| | 2.602 | 3.322 | 0.217 | 27.638 | 2.407 | 3.322 | 0.275 | 35.825 |
| 1349 | 2.828 | 3.170 | 0.108 | 25.611 | 2.883 | 3.322 | 0.132 | 38.408 |
| | 2.750 | 3.000 | 0.083 | 26.050 | 1.648 | 3.170 | 0.480 | 17.993 |
| | 2.627 | 3.000 | 0.124 | 21.021 | 1.892 | 3.170 | 0.403 | 28.625 |
| 1356 | 2.548 | 3.459 | 0.263 | 41.104 | 2.550 | 3.459 | 0.262 | 24.111 |
| | 2.582 | 3.459 | 0.253 | 36.666 | 3.045 | 3.322 | 0.083 | 30.775 |
| | 2.521 | 3.459 | 0.271 | 30.708 | 2.687 | 3.322 | 0.191 | 40.628 |
| 913 | 2.976 | 3.907 | 0.238 | 52.821 | 3.181 | 3.807 | 0.164 | 49.739 |
| | 2.619 | 3.459 | 0.243 | 40.773 | 1.964 | 3.459 | 0.432 | 26.997 |
| | 2.577 | 3.000 | 0.141 | 21.816 | 1.985 | 3.000 | 0.338 | 19.585 |
| 1570 | 3.289 | 3.807 | 0.136 | 42.646 | 2.867 | 3.807 | 0.247 | 38.724 |
| | 2.202 | 3.000 | 0.266 | 36.660 | 3.045 | 3.322 | 0.350 | 30.275 |
| | 2.471 | 3.459 | 0.285 | 49.109 | 2.626 | 3.459 | 0.241 | 39.655 |
| 2748 | 2.255 | 1.067 | 0.321 | 26.486 | 2.684 | 3.322 | 0.192 | 41.059 |
| | 0.654 | 2.585 | 0.747 | 10.424 | 1.814 | 2.585 | 0.298 | 20.955 |
| | 0.766 | 2.585 | 0.703 | 14.877 | 2.104 | 2.585 | 0.186 | 13.087 |
| 1167 | 2.260 | 3.585 | 0.369 | 45.051 | 2.900 | 3.700 | 0.216 | 44.720 |
| | 1.639 | 2.807 | 0.416 | 17.008 | 1.998 | 2.807 | 0.288 | 21.301 |
| | 1.815 | 3.000 | 0.345 | 29.147 | 2.121 | 3.000 | 0.293 | 26.713 |
| 1736 | 2.846 | 3.585 | 0.206 | 45.916 | 2.616 | 3.460 | 0.244 | 41.504 |
| | 1.692 | 2.807 | 0.397 | 16.928 | 2.674 | 3.000 | 0.108 | 21.626 |
| | 1.798 | 2.807 | 0.359 | 29.408 | 2.568 | 3.000 | 0.144 | 22.197 |
| 3208 | 2.815 | 3.700 | 0.239 | 40.823 | 2.530 | 3.807 | 0.335 | 49.295 |
| | 1.349 | 3.000 | 0.550 | 17.679 | 1.611 | 3.170 | 0.492 | 27.314 |
| | 1.354 | 2.807 | 0.922 | 17.185 | 0.990 | 1.585 | 0.375 | 8.554 |

*Table 6* (cont.)

| Sequence ID | $SE_+$ | $SE_{+\max}$ | $SE_{+\mathrm{rel}}$ | $I_+$ | $SE_-$ | $SE_{-\max}$ | $SE_{-\mathrm{rel}}$ | $I_-$ |
|---|---|---|---|---|---|---|---|---|
| 386 | 2.859 | 3.807 | 0.249 | 52.991 | 3.036 | 3.700 | 0.179 | 44.384 |
| | 1.255 | 2.807 | 0.564 | 24.980 | 1.859 | 2.807 | 0.337 | 22.584 |
| | 1.467 | 3.000 | 0.511 | 29.503 | 1.607 | 3.000 | 0.464 | 30.307 |
| 1134 | 3.156 | 4.000 | 0.211 | 62.283 | 3.432 | 4.000 | 0.142 | 60.288 |
| | 1.648 | 2.000 | 0.176 | 10.011 | 1.045 | 2.322 | 0.550 | 6.861 |
| | 1.032 | 2.000 | 0.484 | 7.583 | 1.087 | 2.322 | 0.632 | 19.371 |
| 438 | 2.320 | 3.322 | 0.302 | 30.383 | 1.947 | 3.170 | 0.386 | 27.223 |
| | 0.775 | 1.585 | 0.511 | 2.592 | 1.797 | 2.000 | 0.102 | 8.946 |
| | 0.827 | 1.585 | 0.478 | 9.058 | 1.895 | 2.000 | 0.053 | 8.434 |
| 2265 | 3.630 | 4.523 | 0.197 | 73.147 | 3.992 | 4.460 | 0.104 | 95.248 |
| | 0.090 | 1.000 | 0.909 | 6.461 | 1.018 | 1.585 | 0.357 | 8.050 |
| | 0.078 | 1.000 | 0.922 | 0.014 | 0.990 | 1.585 | 0.357 | 8.554 |

Table 7. *Maximum Similarity Scores for Pairwise Sequence Alignments*

| Sequence ID | Maximum score | Score for 100% identity |
|---|---|---|
| 1000 | 54 | 531 |
| 102 | 38 | 528 |
| 1056 | 37 | 471 |
| 1134 | 53 | 537 |
| 1167 | 55 | 542 |
| 1281 | 42 | 544 |
| 1349 | 71 | 564 |
| 1356 | 47 | 536 |
| 1570 | 67 | 564 |
| 1736 | 71 | 568 |
| 2265 | 56 | 562 |
| 2300 | 56 | 535 |
| 2748 | 44 | 496 |
| 3208 | 54 | 512 |
| 372 | 52 | 516 |
| 386 | 57 | 527 |
| 435 | 67 | 519 |
| 438 | 42 | 496 |
| 595 | 45 | 538 |
| 913 | 49 | 529 |

For proteins shown in *Table 5* presenting a 'Quite good cavity to bind the ligand', the hydrophobicity deficiency is localized in a deep cavity (in the hydrophobic core of the molecule) with gradual decrease of hydrophobicity deficiency in the direction of the surface, reaching zero discrepancy *versus* the idealized hydrophobicity distribution on the surface. According to the characteristics of natural proteins analyzed with the same method [11–13], this is potentially an optimal binding site for a specific ligand. Thus it appears that some of the NBPs could indeed display a ligand binding activity and potentially also a catalytic activity.

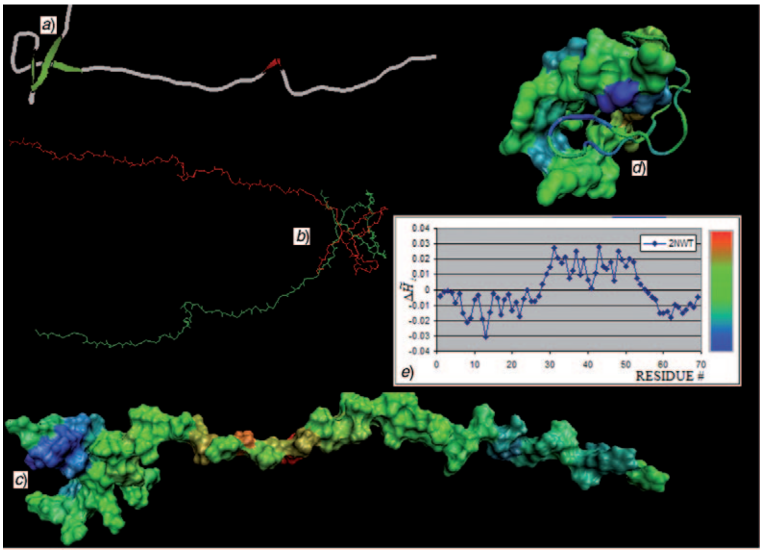

Fig. 4. *The 2NWT protein* (status of unknown function). *a*) Secondary structure of the chain A, *b*) dimeric structure of the protein, *c*) chain A in the color representation proportional to $\Delta\tilde{H}_j$ value (shown in *e*)), *d*) the complexation area showing the dark blue (hydrophobicity higher than expected on the surface of protein), *e*) $\Delta\tilde{H}_j$ profile as calculated for chain A in 2NWT protein.

Table 8. *SE Parameters for the Protein 2NWT – Chain A and B* (unknown function) *in Comparison with S Proteins, which Failed to be Folded*

| Protein or Sequence ID | $SE_+$ | $SE_{+max}$ | $SE_{+rel}$ | $I_+$ | $SE_-$ | $SE_{-max}$ | $SE_{-rel}$ | $I_-$ |
|---|---|---|---|---|---|---|---|---|
| 2NWT-A | 0.00 | 0.00 | NotDef | 0.00 | 0.96 | 1.00 | 0.04 | 2.09 |
| 2NWT-B | 0.75 | 2.58 | 0.71 | 8.91 | 1.84 | 2.81 | 0.34 | 21.72 |
| 3208 | 1.35 | 3.00 | 0.55 | 17.68 | 1.61 | 3.17 | 0.49 | 27.31 |
| | 1.35 | 2.81 | 0.92 | 17.18 | 0.99 | 1.58 | 0.37 | 8.55 |
| 1134 | 1.65 | 2.00 | 0.17 | 10.01 | 1.04 | 2.32 | 0.55 | 6.86 |
| | 1.03 | 2.00 | 0.48 | 7.58 | 1.08 | 2.32 | 0.63 | 19.37 |
| 438 | 0.77 | 1.58 | 0.51 | 2.59 | 1.80 | 2.00 | 0.10 | 8.95 |
| | 0.83 | 1.58 | 0.48 | 9.06 | 1.89 | 2.00 | 0.05 | 8.43 |
| 2265 | 0.09 | 1.00 | 0.91 | 6.46 | 1.02 | 1.58 | 0.38 | 8.05 |
| | 0.08 | 1.00 | 0.92 | 0.01 | 0.99 | 1.58 | 0.38 | 8.55 |

Some of the proteins analyzed display highly dispersed fragments of hydrophobicity deficiency and excess, meaning that there are many small areas suitable for hydrophobic ligand binding. However, such cases were not observed in real proteins present in PDB. It cannot be excluded that in these cases both used methods fail in finding the right solution for the folding of the corresponding sequences or, in alternative, that the structures predicted may be only marginally stable. This is a point that will be clarified only through production and experimental structural characterization of these proteins.

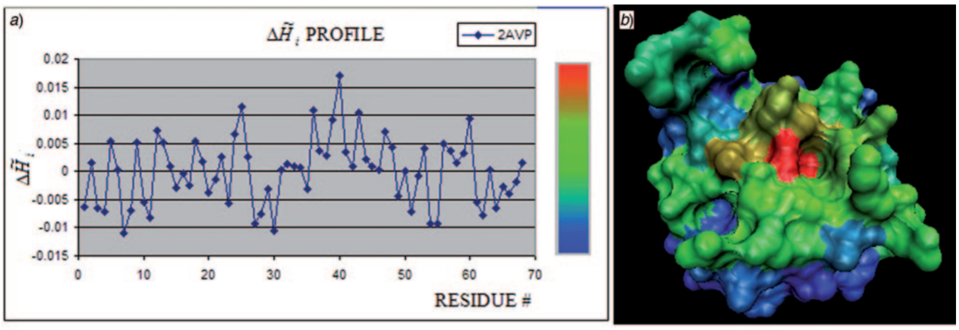

Fig. 5. *The 2AVP protein of the category of* de novo designed *proteins* (according to PDB notation). *a*) $\Delta\tilde{H}_j$ Profile of 2AVP protein, *b*) 3-D presentation of 2AVP according to $\Delta\tilde{H}_j$ value in color scale shown in *a*).

Table 9. *Short Description and SE Parameters for the* de novo *Designed Protein 2AVP-A, Which Is an 8 Repeat Consensus TPR Superhelix* de novo *Synthetic Construct*

| | $SE_{+}$ | $SE_{+max}$ | $SE_{+rel}$ | $I_{+}$ | $SE_{-}$ | $SE_{-max}$ | $SE_{-rel}$ | $I_{-}$ |
|---|---|---|---|---|---|---|---|---|
| 2AVP-A | 2.017 | 3.8071 | 0.312 | 41.74 | 3.537 | 3.807 | 0.071 | 57.32 |

Regarding the proteins classified as 'Antifreeze-like protein', this term should be understood as follows. The FOD model directs the hydrophobic residues toward the center of the protein molecule with simultaneous exposure of the hydrophilic residues on the protein surface. A polypeptide chain folded ideally according to this model should be a molecule exposing all the hydrophilic residues on the surface with well concentrated hydrophobic amino acids in the center of the protein body. Such molecule would be very well soluble with no special activity in the sense of interaction with other molecules. The binding characteristics of such a molecule could well be represented by highly soluble antifreeze molecules, which according to the FOD model seem to be folded in high accordance with this model displaying negligible small (or none) hydrophobicity distribution discrepancy *versus* the theoretical distribution.

A general comment concerns the S structures interpreted as 'unfolded'. Among the structures of real proteins present in the PDB, one molecule was found to be classified as unfolded (PDB code 2NWT). The SE parameters of this structure are highly comparable to those of the structures analyzed in the present work and classified as unfolded. This particular protein is quite peculiar. It is an element of a large protein complex (ribosome). Thus the structure is generated independently and afterwards complexed to the larger multi-molecular complex. Alternatively, its polypeptide chain could fold in a specific cavity of another component of the complex. Calculated parameters for this protein seem to be similar to proteins of S form characterized by high RMS-D values (calculated *versus* the Rosetta structures). This observation is even more interesting taking into account that the protein under consideration has the status 'unknown function' according to the PDB classification.

The structures characterized by high RMS-D values, comparing the R and S forms, could also be interpreted as a failure of the S model. It was indeed observed that some

Table 10. *Results of the Statistical Analysis Verifying Difference Significance between Group R* vs. *Group P and S as Obtained According to* Wilcoxon (Rank Sums) *for Parameters Given in a Head of each Sub-Table*

SE$_+$ *Kruskal-Wallis* test value of statistics = 7.5115, $p = 0.0234$

| Proteins | *N* | Sum of Scores | Expected under *Ho* | Std. dev. under *Ho* | Mean score |
|---|---|---|---|---|---|
| P | 182 | 19760 | 20293 | 367.813 | 108.571 |
| R | 20 | 2969 | 2230 | 274.000 | 148.450 |
| S | 20 | 2024 | 2230 | 274.000 | 101.200 |

SE$_{\max+}$ *Kruskal-Wallis* test value of statistics = 17.526, $p = 0.0002$

| Proteins | *N* | Sum of Scores | Expected under *Ho* | Std. dev. under *Ho* | Mean score |
|---|---|---|---|---|---|
| P | 182 | 19234.5 | 20293 | 365.369 | 105.684 |
| R | 20 | 3369 | 2230 | 272.180 | 168.450 |
| S | 20 | 2149.5 | 2230 | 272.180 | 107.475 |

L$_+$ *Kruskal-Wallis* test value of statistics = 17.526, $p = 0.0002$

| Proteins | *N* | Sum of Scores | Expected under *Ho* | Std. dev. under *Ho* | Mean score |
|---|---|---|---|---|---|
| P | 182 | 19234.5 | 20293 | 365.369 | 105.684 |
| R | 20 | 3369 | 2230 | 272.180 | 168.450 |
| S | 20 | 2149.5 | 2230 | 272.180 | 107.475 |

SE$_-$ *Kruskal-Wallis* test value of statistics = 8.787, $p = 0.0124$

| Proteins | *N* | Sum of Scores | Expected under *Ho* | Std. dev. under *Ho* | Mean score |
|---|---|---|---|---|---|
| P | 182 | 19671 | 20293 | 367.813 | 108.082 |
| R | 20 | 3035.5 | 2230 | 274.000 | 151.775 |
| S | 20 | 2046.5 | 2230 | 274.000 | 102.325 |

SE$_{-\max}$ *Kruskal-Wallis* test value of statistics = 13.642, $p = 0.0011$

| Proteins | *N* | Sum of Scores | Expected under *Ho* | Std. dev. under *Ho* | Mean score |
|---|---|---|---|---|---|
| P | 182 | 19466.5 | 20293 | 365.238 | 106.959 |
| R | 20 | 3232 | 2230 | 272.082 | 161.600 |
| S | 20 | 2054 | 2230 | 272.082 | 102.725 |

L$_-$ *Kruskal-Wallis* test value of statistics = 13.6417, $p = 0.0011$

| Proteins | *N* | Sum of Scores | Expected under *Ho* | Std. dev. under *Ho* | Mean score |
|---|---|---|---|---|---|
| P | 182 | 19466.5 | 20293 | 365.238 | 106.959 |
| R | 20 | 3232 | 2230 | 272.082 | 161.600 |
| S | 20 | 2054.5 | 2230 | 272.082 | 102.725 |

Table 11. *The Results of ANOVA Test to Verify the Differences between R* vs. *P and S Group for I$_+$ and I$_-$* (I$_+$ ANOVA $F = 9.75$, $p < 0.001$)

| Proteins | *N* | Mean I$_+$ | Mean I$_-$ |
|---|---|---|---|
| P | 182 | 33.676 | 36.424 |
| R | 20 | 47.383 | 49.273 |
| S | 20 | 33.833 | 34.333 |

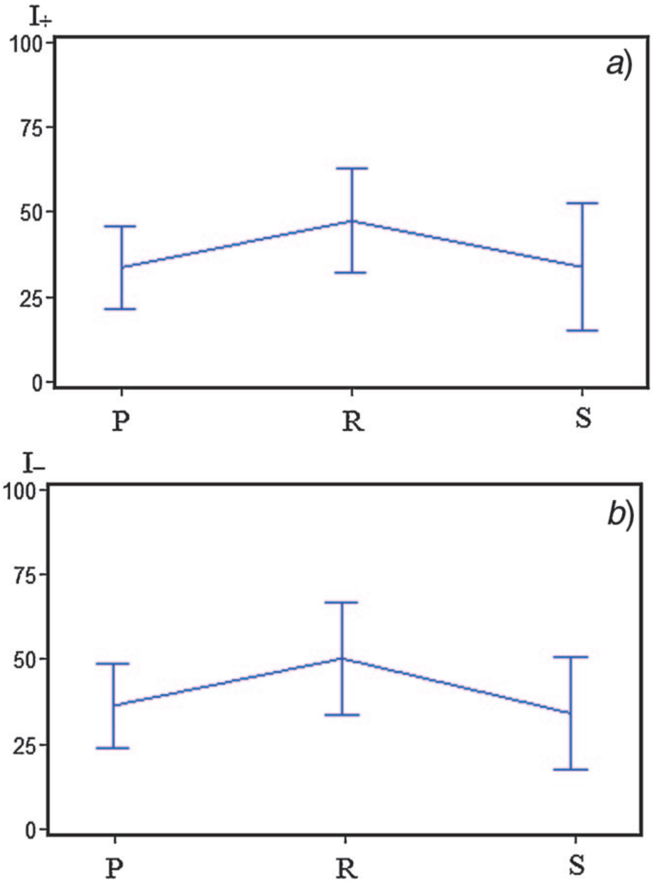

Fig. 6. *The parameters* (mean value and standard deviation) *calculated according to ANOVA test. a*) for $I_+$ and *b*) for $I_-$.

sequences were too difficult to be folded according to the FOD model as compared to the Rosetta model. The hydrophobic deficiency of the surface of the protein molecule, observed for this class of proteins in the S model means that the hydrophobic core (assuming it exists in the protein molecule) gets 'opened' too much and gets exposed on the surface of the molecule. Potentially, such areas could interact with other proteins with a highly hydrophobic area on the surface. However, according to the analysis of about 300 70 amino acid residues long proteins present in the PDB, such a situation does not occur in real proteins.

**Conclusions.** – This work was carried out with two main aims. The first one was to assess the possibility of exploring the enormous sequence space of proteins not present in nature in the search for protein molecules endowed with potentially useful biological functions. To this aim, we tested the performance in terms of efficiency and the agreement between two protein structure/folding prediction software tools, the Rosetta

Table 12. *The Characteristics of All Three Groups* (P: upper row, S: middle row, R: lower row in each cell) *Expressed by Mean Values, Standard Deviation, Minimum Value, Maximum Value, and Number of Amino Acids in Polypeptide Chain*

| Variable | Mean | Std. Dev. | Min | Max | *N* |
|---|---|---|---|---|---|
| *N* Amino acids | 69.417 | 1.721 | 64 | 73.000 | 182 |
| | 70.000 | 0 | 70 | 70.00 | 20 |
| | 70.000 | 0 | 70 | 70.00 | 20 |
| $SE_{+}$ [bit] | 2.473 | 0.605 | 0 | 3.535 | 182 |
| | 2.171 | 1.128 | 0.966 | 3.582 | 20 |
| | 2.872 | 0.388 | 2.254 | 3.630 | 20 |
| $SE_{+\max}$ [bit] | 3.338 | 0.460 | 0.097 | 4.087 | 182 |
| | 3.134 | 0.919 | 1.00 | 4.000 | 20 |
| | 3.734 | 0.304 | 3.170 | 4.524 | 20 |
| $\Delta SE_{+}$ [bit] | 0.865 | 0.354 | 0 | 1.829 | 182 |
| | 0.963 | 0.494 | 0.324 | 2.133 | 20 |
| | 0.861 | 0.252 | 0.299 | 1.325 | 20 |
| $I_{+}$ [bit] | 33.676 | 12.214 | 0 | 78.869 | 182 |
| | 33.833 | 18.989 | 2.722 | 62.601 | 20 |
| | 47.383 | 15.312 | 23.169 | 81.805 | 20 |
| $L_{+}$ [bit] | 10.527 | 2.610 | 1 | 17.000 | 182 |
| | 10.150 | 4.380 | 2.000 | 16.000 | 20 |
| | 13.600 | 3.050 | 9.000 | 23.000 | 20 |
| $SE_{-}$ [bit] | 2.611 | 0.488 | 0.957 | 3.573 | 182 |
| | 2.464 | 0.768 | 1.072 | 3.539 | 20 |
| | 2.957 | 0.768 | 1.072 | 3.539 | 20 |
| $SE_{-\max}$ [bit] | 3.389 | 0.398 | 1 | 4.087 | 182 |
| | 3.217 | 0.724 | 1.585 | 4.000 | 20 |
| | 3.719 | 0.301 | 3.170 | 4.459 | 20 |
| $\Delta SE_{-}$ [bit] | 0.779 | 0.324 | 0.042 | 2.203 | 182 |
| | 0.752 | 0.339 | 0.307 | 1.462 | 20 |
| | 0.761 | 0.235 | 0.439 | 1.277 | 20 |
| $I_{-}$ [bit] | 36.424 | 11.424 | 2.087 | 71.439 | 182 |
| | 34.333 | 15.171 | 7.310 | 58.985 | 20 |
| | 49.273 | 15.178 | 24.111 | 95.248 | 20 |
| $L_{-}$ [bit] | 10.835 | 2.569 | 2 | 17.000 | 182 |
| | 10.250 | 3.878 | 3.00 | 16.000 | 20 |
| | 13.450 | 2.946 | 9.00 | 22.000 | 20 |

*ab initio* protein structure prediction software and the FOD model for protein folding simulations. Results obtained are encouraging in terms of the number of amino acid sequences that can be sampled, though many of the random sequences analyzed seem to be challenging with respect to the reliability of the structure prediction. In fact a low level of agreement has been observed between the predictions carried out with the two complementary methods. The origin of this result has to be analyzed in detail and could well be due to the intrinsic nature of the sequences analyzed and possibly to a low tendency to achieve a stable three dimensional form. Once a larger sample of NBP will be studied it will be possible to analyze statistically significant deviations of NBP from real proteins in terms for example of sequence composition and presence/abundance of particular amino acid residues. Nonetheless, the FOD model seems to be able to capture structural properties of the proteins analyzed which can lead to hypothesize the putative functional properties of a protein molecule. Attempts to produce and experimentally characterize some of the proteins described in the present work are currently in progress and will allow to validate the computational approach undertaken.

The second aim of this work was that of setting up an infrastructure for massive protein structure prediction initiatives exploiting the potential of grid computing techniques within the framework of the EUChinaGRID project. This was an interdisciplinary collaborative effort that involved grid experts, bioinformaticians, and biochemists. From this viewpoint, results presented in this work demonstrate that the infrastructure is able to support the prediction of a huge number of protein structures and to provide tools that allow the access to this enormous computing power also to researchers not trained in grid computing. This type of infrastructures could be exploited not only for the study of non natural proteins, the test case and the topic of the present work, but also for initiatives aimed at predicting the structure of a large number of natural proteins for biomedical purposes.

The problem of 'noisy' sequences seems to be of high importance in relation to biological function in immunoglobulin production [45]. The hyper-variable loops sequences and structures seem to be the excellent database relevant to NBP investigation. The comparative analysis of these sequences and structures with respect to the NBP problem will be the subject of a prospective research project.

The authors wish to tank Prof. *Luisi*, who actually devised the original idea and the definition of NBPs. The authors are very grateful to Prof. *Leszek Konieczny*, Institute of Medical Biochemistry, Jagiellonian University – Collegium Medicum Krakow, Kopernika 7, Poland, for fruitful discussion. Many thanks for administrative work to *Smietanska Anna*, *Kruk Alicja*, *Klepacz-Zielinska Ewa*, *Rak-Gubernat Katarzyna*, and *Starzyk Stanisław*.

The work was financially supported by the European Commission EUChinaGRID grant (contract number: 026634).

## Experimental Part

*Random Amino Acid Sequences Generation.* Random amino acid sequences (70 amino acid residues long) were generated using the utility RandomBlast. The utility has been described in detail elsewhere [46]. Briefly, RandomBlast consists of two main modules: a *pseudo* random sequence generation module and a Blast software interface module. Random numbers generated by the utility are translated into single character amino acid code using a conversion table. Single amino acids are then concatenated to reach the specified sequence length. Each generated sequence is then given as input to the second RandomBlast module, an interface to the blastall program [47], which searches for statistically significant similar sequences in the non-redundant NCBI protein sequence database [48]. Blastall output is then retrieved by RandomBlast, and the *E*-value [49] extracted from it. If the *E*-value is greater than or equals the threshold chosen by the user, the sequence is valid and is added to the output file. Note that in our case we regard as valid only the sequences that do not display significant similarity to any protein sequence present in the database, so that, contrary to the normal Blast usage, valid sequences are those displaying an *E*-value higher than the threshold. For the production of NBP sequences the *E*-value threshold was set to 1.0 in order to be sure to sample the sequence space far away enough from the ensemble of natural known proteins.

*Rosetta Model Description.* Rosetta-abinitio is an *ab initio* protein structure prediction software which is based on the assumption that in a polypeptide chain local interactions bias the conformation of sequence fragments, while global interactions determine the three-dimensional structure with minimal energy which is also compatible with the local biases [6]. To derive the local sequence–structure relationships for a given amino acid sequence (the query sequence) Rosetta-abinitio uses the Protein Data Bank to extract the distribution of conformations adopted by short segments in known structures. The latter is taken as an approximation of the distribution adopted by the query sequence segments during the folding process [6]. In details, Rosetta workflow can be divided into two phases. In the first phase, the query sequence is divided in fragments of 3 and 9 amino acids. The software extracts from the data base of protein structures the distribution of three-dimensional structures adopted by these fragments based on their specific sequence. For each query sequence a fragments data base is derived which contains all the possible local structures adopted by each fragment of the entire sequence. In the second execution phase, using the derived fragments database, for each query sequence the sets of fragments are assembled by Rosetta in a high number of different combinations using a Monte Carlo procedure. The resulting structures are then subjected to an energy minimization procedure using a semi-empirical force field [6], in which the principal non-local interactions considered by the software are hydrophobic interactions, electrostatic interactions, main chain H-bonds, and excluded volume. The structures compatible with both local biases and non-local interactions are ranked according to their total energy resulting from the minimization procedure. In the present work, for each query sequence only the highest ranking predicted structure was considered for further analysis (see below).

*'Fuzzy-Oil-Drop'* (FOD) *Model Based Protein Folding.* The model under consideration is based on two main assumptions: *1*) the simulation of the protein folding process rather than the prediction of the protein structure; *2*) a multi-step nature of the protein folding process is assumed according to the experimental observations indicating the presence of intermediates in this process:

$$\mathrm{U} \rightarrow \mathrm{I}_1 \rightarrow \mathrm{I}_2 \rightarrow \ldots\ldots\ldots\ldots \mathrm{I}_k \rightarrow \ldots\ldots\ldots\ldots \mathrm{N}$$

where U unfolded, N native form, I intermediates, the number of which is assumed to be two in the model applied in this work:

$$\mathrm{U} \rightarrow \mathrm{ES} \rightarrow \mathrm{LS} \rightarrow \mathrm{N}$$

where ES early stage, LS late stage intermediate.

The ES model is based on the following assumption: the conformational subspace is limited to the area of the *Ramachandran* map which appears to be optimal for backbone conformation [50][51].

The structure of LS is generated in the presence of an external force field of hydrophobic character. This external force field is expressed by a three-dimensional *Gauss* function which represents the hydrophobicity (traditionally the value of the *Gauss* function is interpreted as probability) density distribution, in agreement with the commonly accepted model of hydrophobic core in proteins [52]. The method accepts all conformations of the folding polypeptide which display optimal internal (side chain–side chain) interaction and additionally optimal hydrophobic side chain–side chain interaction accordant with the external force field. Minimization of the difference between the idealized (*Gauss* function) distribution and that observed in the folding polypeptide directs the folding process toward hydrophobic core generation with the simultaneous exposure of hydrophilic residues on the protein surface.

A protein structure generated according to this model produces a very well soluble molecule with no specific activity (in terms of exposure of surface areas with hydrophobicity deficiency/excess [10]).

Analysis of hydrophobic density distribution with respect to the idealized one reveals highly specific irregularity most frequently localized in a well-defined area, which very often represents the active site of the protein molecule [9]. This observation allows the use of the above described method for biological activity (ligand binding cavity) identification.

Some attempts to fold a protein in the presence of a specific target molecule (co-enzyme, ligand, or even substrate) have been undertaken to verify the possible participation of these molecules during the folding process [14][15].

The small-size-protein molecules (relatively low number of degrees of freedom) are not always able to eliminate the hydrophilicity buried or hydrophobicity exposed in the final structure. The degree of the disagreement is highly sequence specific. The similar example (56 amino acid residues) analyzed in [7] shows that the hydrophobicity discrepancy present in the *in silico* folded protein molecule may be of high accordance with the one present in the crystal structure, evident from the parallel curves on the $\Delta\tilde{H}_j$ profile representing the *in silico* folded and the native structure of this protein.

The profile representing the $\Delta\tilde{H}_j$ values distribution along the polypeptide chain appeared to be highly specific for the folded protein and thus it may suggest the possible binding site.

The $\Delta\tilde{H}_j$ profile ($\Delta\tilde{H}_j = \tilde{H}_j^t - \tilde{H}_j^o$ where $\tilde{H}_j^t$ and $\tilde{H}_j^o$ are idealized and observed hydrophobicity distribution for each *j*-th amino acid residue, resp.) may be used for structural/functional similarity search. The comparative analysis may be easily performed comparing the entropy (information entropy [53]) of the folded chain comparing the distribution of $\Delta\tilde{H}_j$maxima and $\Delta\tilde{H}_j$ minima observed in a particular protein with the entirely random distribution of fragments (their length and intensity) of positive and negative $\Delta\tilde{H}_j$ values.

The details concerning the early stage (ES) intermediate structure generation were presented in details elsewhere [51][54].

*Calculation Protocol.* The simulation of the folding process according to the FOD model consists of 20 consecutive minimizations of two kinds: the standard internal energy minimization (non-bonding interaction) and the $\Delta\tilde{H}_j$ minimization (as described above), applied alternating one after the other.

The internal energy is calculated according to the ECEPP/3 force field [55–59], and its optimization is performed by a numerical method with analytic gradient and *Hessian* approximation from second update (until convergence is reached).

The hydrophobicity driven optimization procedure is performed according to the non-gradient *Rosenbrock* method (up to 5000 steps – estimated as optimal on test runs). After every pair of simulated hydrophobic collapse (constraint in the form of limited oil-drop size) and consecutive relaxation (absence of any external constraints), a *Gaussian* expressing the theoretical spatial hydrophobicity distribution in $\Delta\tilde{H}_j$ calculation is monotonically shrinking in three directions. In the final step of the $\Delta\tilde{H}_j$ minimization, the size of the 'fuzzy oil-drop' reaches the shape that is in approximation characteristic for water-soluble proteins of given length [54].

According to the model description, amino acid residues in the FOD model (particularly for the hydrophobicity driven step) are represented by their effective atoms (averaged position of atoms in side chains). These steps are followed by the standard nonbonding interaction optimizations. The number of these steps was insufficient to eliminate possible highly energetic mutual approaches of atoms. An increase in the length of final optimization procedure (non-bonding interaction) was necessary due to the unusual character of the random amino acids sequences under consideration.

This is why an additional optimization procedure was applied: energy minimization *in vacuo* and in the presence of water solvent. Amber 9 molecular dynamics suite [60][61], was used with empirical force field ff99 [62]. Both *in vacuo* and *in explicit* solvent calculations, non-bonding pair-wise interactions were cut off at 12 Å. The TIP3P three point rigid water model was used for the solvent [63].

For each structure, the calculation procedure took 50000 optimization steps and led to the decrease of the overall potential energy in the system. Additionally calculations in explicit solvent were preceded by water molecules equilibration (20000 steps).

*Protein Selection.* The three dimensional structures of 3300 NBP amino acid sequences, generated both according to Rosetta (called R in this article) and the FOD methods (called S in this article – as they appeared using the standard procedure implemented on the grid system), were ordered according to RMS-D values obtained by a pair-wise structural comparison using the *Kabsch* method [64]. The ten top (the lowest RMS-D values pairs) and ten bottom (the highest RMS-D values pairs) structure pairs were selected for detailed analysis in this work. The complete set of proteins in forms R and S will be analyzed and presented in further articles.

The amino acid sequences of the selected proteins are reported in *Table 1*.

*Binding Site Characterization.* The entropy of a binding site involving residues close in sequence is low, compared with the entropy of a binding site formed by residues evenly distributed in sequence. Information entropy (SE) calculated for fragments with positive $\Delta\tilde{H}_j$ ($\Delta\tilde{H}_j^p$) measures the amount of uncertainty about the organization of residues forming the binding site (see the examples discussed in [7]).

$$\mathrm{SE}_{+} = -\sum_{j}^{K} p_j \log_2 p_j [bit] \tag{1}$$

where $K$ denotes the number of fragments with positive $\Delta\tilde{H}_j$, and

$$p_j = \sum_{i=1}^{N_{ij}} \frac{\Delta\tilde{H}_i^p}{\Delta\tilde{H}_t^p} \tag{2}$$

where $N_{ij}$ represents the number of positive $\Delta\tilde{H}_i^p$ values belonging to $j$-th fragment and $\Delta\tilde{H}_t^p$ is the sum of all positive $\Delta\tilde{H}_j$ values in the whole polypeptide chain.

The $\mathrm{SE}_{+}$ value characterizes a particular protein and may describe an active site (fragments with positive values). Another quantity used in the study of binding sites is the information (I) necessary to localize residues creating the binding site. The participation of particular residues in the active site creation is understood as a probability expressing conjunction of events (close mutual localization) and can be created according to *Eqn. 3*.

$$\mathrm{I} = -\log_2 \prod_{j=1}^{K} p_j \tag{3}$$

where $K$ has the same meaning as in the *Eqn. 1.*

Both quantities (SE, I) describe similar characteristics of the event of selected residues participating in the organization of a structural element treated as an active site.

*Computational Platform* (Grid System). Pharmacology appears as the one of the important consumer of large scale computing [65]. As a computing facility for conducting the simulations described in this work, we used the computing Grid provided by the EUChinaGRID project [66]. The intercontinental infrastructure comprised twelve clusters, five of them in China and seven in Europe, with a total of 257 machines with 644 CPUs. The machines are managed using gLite middleware [67], which automatically distributes the computing tasks onto worker nodes, managed in turn by local queuing systems, such as Torque or LSF.

The software used for computation had to be prepared for running on the Grid. This required packaging all the programs, libraries, and steering scripts into a self-contained single distribution, which was deployed on the Grid Storage Element and available for download. Then, a script which was submitted for processing on the Grid was responsible for downloading this pre-packaged distribution, installing it, running and storing the output. As the storage for the generated output, we used gLite Storage Element, and LFC catalogue for registering the files.

To facilitate the management of the experiment which comprised the order of 10000 structures, we set up a database and an experiment management system which was integrated within a portal as a user interface. For automating the process of submission of jobs onto the Grid we used LCG-API [68] and GridSphere framework [69]. This allows to monitor the progress of the experiment, statistics on the completed jobs, *etc.* It is also possible to preview the generated structures and perform simple analysis thanks to the tools integrated with the portal.

Using the Grid infrastructure was proven to be very convenient for the large scale computing. Assuming that computing a single protein structure requires approximately one CPU hour, it would be possible to compute only 24 structures per day on a single machine. Using only a part of the whole Grid infrastructure (there were other jobs running concurrently) we were able to achieve a peak throughput of *ca.* 900 structures per day. As a future work we are investigating the possible usage of tools for management of experiments such as Taverna workflow system [70] or ViroLab virtual laboratory [71].

*Tools Available on the Portal.* Preliminary comparative result analysis is available on the portal. The tools available are the following: visualization of secondary structure assignments, $\Phi/\Psi$ maps showing the dihedral angles distribution in a particular protein and contact maps. The superimposed structures can be immediately visualized by launching a JavaWebStart viewer based on Molecular Biology Toolkit [72].

Another JavaWebStart viewer, Reveal, shows molecular surfaces unraveling areas with high observed *vs.* theoretical hydrophobicity discrepancy [73–75].

*Statistical Analysis.* The sets of parameters describing three groups of proteins: S (according to 'fuzzy oil drop' model), R (according to Rosetta method), and P (natural proteins of 70 aa in polypeptide chain) as described in [11–13]. The non-parametric tests (*Wilcoxon* test and *Kruskal-Wallis* test) were applied for variables representing other than normal distribution and ANOVA test for variables representing normal distribution. The SAS program was applied for calculation to verify the hypothesis of significant differences between parameters as they appeared in a particular group.

## REFERENCES

[1] M. Rosenberg, A. Goldblum, *Curr. Pharm. Des.* **2006**, *12*, 3973.
[2] R. Jaenicke, R. Sterner, *Angew. Chem., Int. Ed.* **2003**, *42*, 140.
[3] C. Vendrely, T. Scheibel, *Macromol. Biosci.* **2007**, *7*, 401.
[4] C. Chiarabelli, J. W. Vrijbloed, D. D. Lucrezia, R. M. Thomas, P. Stano, F. Polticelli, T. Ottone, E. Papa, P. L. Luisi, *Chem. Biodiversity* **2006**, *3*, 840.
[5] C. Chiarabelli, J. W. Vrijbloed, R. M. Thomas, P. L. Luisi, *Chem. Biodiversity* **2006**, *3*, 827.
[6] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, D. Baker, *Methods Enzymol.* **2004**, *383*, 66.
[7] L. Konieczny, M. Brylinski, I. Roterman, *In Silico Biol.* **2006**, *6*, 15.
[8] M. Bryliński, K. Prymula, W. Jurkowski, M. Kochańczyk, E. Stawowczyk, L. Konieczny, I. Roterman, *PLoS Comput. Biol.* **2007**, *3*, e94.
[9] M. Brylinski, M. Kochanczyk, E. Broniaowska, I. Roterman, *J. Mol. Model.* **2007**, *13*, 665.
[10] M. Brylinski, L. Konieczny, I. Roterman, *Bioinformation* **2006**, *1*, 127.
[11] K. Prymula, I. Roterman, *J. Biomol. Struct. Dyn.*, **2009**, *26*, 663.
[12] K. Prymula, I. Roterman, *Entropy* **2009**, *11*, 62.
[13] K. Prymula, I. Roterman, *J. Mol. Mod.*, accepted.
[14] M. Brylinski, L. Konieczny, I. Roterman, *Int. J. Bioinform. Res. Appl.* **2007**, *3*, 234.
[15] M. Brylinski, L. Konieczny, I. Roterman, *Comput. Biol. Chem.* **2006**, *30*, 255.
[16] M. Brylinski, L. Konieczny, I. Roterman, *J. Biomol. Struct. Dyn.* **2006**, *23*, 519.
[17] M. Levitt, *J. Mol. Biol*. **1976**, *104*, 59.
[18] J. Kyte, R F. Doolittle, *J. Mol. Biol.* **1982**, *157*, 105.
[19] D. M. Engelman, T. A. Steitz, A. Goldman, *Annu. Rev. Biophys. Biophys. Chem.* **1986**, *15*, 321.
[20] T. P. Hopp, K. R. Woods, *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 3824.
[21] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, M. H. Zehfus, *Science* **1985**, *229*, 834.
[22] R. Wolfenden, L. Andersson, P. M. Cullis, C. C. Southgate, *Biochemistry* **1981**, *20*, 849.
[23] M. Brylinski, L. Konieczny, I. Roterman, *Int. J. Bioinform. Res. Appl.* **2007**, *3*, 234.
[24] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, *J. Mol. Biol.* **2004**, *344*, 1135.
[25] A. T. Laurie, R. M. Jackson, *Bioinformatics* **2005**, *21*, 1908.
[26] V. J. Gillet, G. Myatt, Z. Zsoldos, A. P. Johnson, *Perspect. Drug. Discovery Design* **1995**, *3*, 34.
[27] J. M. S. Law, D. Y. K. Fung, Z. Zsoldos, A. Simon, Z. Szabo, *J. Mol. Struct. (Theochem)* **2003**, *651–657*, 666.
[28] L. Wei, R. B. Altman, *Pac. Symp. Biocomput.* **1998**, 497.
[29] M. P. Liang, D. R. Banatao, T. E. Klein, D. L. Brutlag, R. B. Altman, *Nucleic Acids Res.* **2003**, *31*, 3324.
[30] D. R. Banatao, R. B. Altman, T. E. Klein, *Nucleic Acids Res.* **2003**, 31, 4450.
[31] J. Ko, L. F. Murga, Y. Wei, M. J. Ondrechen, *Bioinformatics* (Supplement 1) **2005**, *21*, 258.
[32] I. A. Shehadi, A. Abyzov, A. Uzun, Y. Wei, L. F. Murga, *J. Bioinform. Comput. Biol.* **2005**, *3*, 127.
[33] J. Ko, L. F. Murga, P. Andre, H. Yang, M. J. Ondrechen, *Proteins* **2005**, *59*, 183.
[34] K. P. Peters, J. Fauck, J. C. Frommel, *J. Mol. Biol.* **1996**, *256*, 201.
[35] J. An, M. Totrov, R. Abagyan, *Genome Inform.* **2004**, *15*, 31.
[36] M. Hendlich, F. Rippmann, G. Barnickel. *J. Mol. Graph. Model.* **1997**, *15*, 359.
[37] M. Jambon, O. Andrieu, C. Combet, G. Deléage, F. Delfaud, *Bioinformatics* **2006**, *21*, 3929.

[38] M. Jambon, A. Imberty, G. Deléage, C. Geourjon, *Proteins* **2003**, *52*, 137.

[39] R. A. Laskowski, J. D. Watson, J. M. Thornton, *Nucleic Acids Res.* **2005**, *33*, W89.

[40] W. Kabsch, C. Sander, *Biopolymers* **1983**, *22*, 2577.

[41] K. Pearson, *Philos. Trans. R. Soc. London, Ser. A* **1896**, *187*, 253.

[42] C. Spearman, *Am. J. Psychol.* **1906**, *6*, 201.

[43] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, C. Zardecki, *Acta Crystallogr., Sect. D* **2002**, *58*, 899.

[44] W. R. Pearson, D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444.

[45] Y. Chen, G. Varani, *FEBS J.* **2005**, *272*, 2088.

[46] G. Evangelista, G. Minervini, P. L. Luisi, F. Polticelli, *Bio-Algorithms Med-Syst.* **2007**, *2*, 27.

[47] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, *215*, 403.

[48] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, J. U. Pontius, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, E. Yaschenko, *Nucleic Acids Res.* **2005**, *33*, D39.

[49] S. Karlin, S. F. Altschul, *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 2264.

[50] W. Jurkowski, M. Brylinski, L. Konieczny, I. Roterman. *J. Biomol. Struct. Dyn.* **2004**, *22*, 149.

[51] W. Jurkowski, M. Brylinski, L. Konieczny, Z. Wiíniowski, I. Roterman, *Proteins* **2004**, *55*, 115.

[52] W. Kauzmann, *Adv. Protein Chem.* **1959**, *14*, 1.

[53] C. Shannon, *Bell Syst. Tech. J.* **1948**, *27*, 379.

[54] M. Brylinski, M. Kochanczyk, L. Konieczny, I. Roterman, *In Silico Biol.* **2006**, *6*, 589.

[55] G. Nemethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, H. A. Scheraga, *J. Phys. Chem.* **1992**, *96*, 6472.

[56] G. Nemethy, M. S. Pottle, H. A. Scheraga. *J. Phys. Chem.* **1983**, *87*, 1883.

[57] M. J. Sippl, G. Nemethy, H. A. Scheraga. *J. Phys. Chem.* **1984**, *88*, 6231.

[58] F. A. Momany, R. F. McGuire, A. W. Burgess, H. A. Scheraga, *J. Phys. Chem.* **1975**, *79*, 2361.

[59] L. G. Dunfield, A. W. Burgess, H. A. Scheraga, *J. Phys. Chem.* **1978**, *82*, 2609.

[60] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham III, S. DeBott, D. Ferguson, G. Seibel, P. Kollman, *Comp. Phys. Commun.* **1995**, *91*, 1.

[61] D. Bashford, D. A. Case, *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.

[62] J. Wang, P. Cieplak, P. A. Kollman, *J. Comput. Chem.* **2000**, *21*, 1049.

[63] J. P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, *J. Comput. Phys.* **1977**, *23*, 327.

[64] W. Kabsch, *Acta Crystallogr., Sect. A* **1976**, *32*, 922.

[65] B. Zagrovic, C. D. Snow, M. R. Shirts, V. S. Pande, *J. Mol. Biol.* **2002**, *323*, 927.

[66] http://www.euchinagrid.org/.

[67] http://glite.web.cern.ch/glite/.

[68] http://www.gridwisetech.com/content/view/91/96/lang.en.

[69] J. Novotny, M. Russell, O. Wehrens, *Concurrency Comput.: Pract. Exp.* **2004**, *16*, 503.

[70] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, P. Li, *Bioinformatics* **2004**, *20*, 3045.

[71] M. Bubak, T. Gubala, M. Kasztelnik, M. Malawski, P. Nowakowski, P. M. Sloot, in 'Expanding the Knowledge Economy: Issues, Applications, Case Studies, eChallenges e-2007 Conference Proceedings', IOS Press, 2007, p. 537.

[72] J. L. Moreland, A. Gramada, O. V. Buzko, Q. Zhang, P. E. Bourne, *BMC Bioinformatics* **2005**, *6*, 21.

[73] M. Malawski, T. Szepieniec, M. Kochanczyk, M. Piwowar, I. Roterman, *Bio-Algorithms Med-Syst.* **2007**, *3*, 45.

[74] G. Minervini, G. L. Rocca, P. L. Luisi, F. Polticelli, *Bio-Algorithms Med-Syst.* **2007**, *3*, 39.

[75] A. Budano, P. Celio, S. Cellini, R. Gargana, R. Gargana, F. Galeazzi, C. Stanescu, F. Ruggieri, Y. Q. Guo, L. Wang, X. M. Zhang, *Bio-Algorithms Med-Syst.* **2007**, *3*, 33.

*Received November 11, 2008*