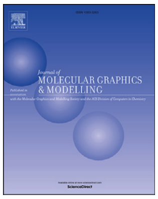
Journal of Molecular Graphics and Modelling 87 (2019) 227–239

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM



The aqueous environment as an active participant in the protein folding process

Małgorzata Gadzała [a, 1], Dawid Dułak [b], Barbara Kalinowska [c, d], Zbigniew Baster [e, f], Michał Byliński [g, h], Leszek Konieczny [i], Mateusz Banach [d], Irena Roterman [d, *]

[a] ACK Cyfronet AGH, Nawojki 11, 30-950, Kraków, Poland
[b] ABB Business Services Sp. z o.o. ul. Żegańska 1, 04-713, Warszawa, Poland
[c] Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, 11 Łojasiewicza Street, Kraków, Poland
[d] Department of Bioinformatics and Telemedicine, Jagiellonian University – Medical College, Łazarza 16, 31-530, Kraków, Poland
[e] Department of Molecular and Interfacial Biophysics, Faculty of Physics, Astronomy, Applied Computer Science Jagiellonian University, 11 Łojasiewicza Street, Kraków, Poland
[f] Markey Cancer Center, University of Kentucky, 789 South Limestone Street, Lexington, KY, USA
[g] Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, 70803, USA
[h] Center for Computation & Technology, Louisiana State University, Baton Rouge, LA, 70803, USA
[i] Chair of Medical Biochemistry, Jagiellonian University – Medical College, Kopernika 7E, 31-034, Kraków, Poland

ARTICLE INFO

Article history:
Received 12 August 2018
Received in revised form
5 December 2018
Accepted 12 December 2018
Available online 14 December 2018

Keywords:
Protein structure prediction
Hydrophobic core
External force field

ABSTRACT

Existing computational models applied in the protein structure prediction process do not sufficiently account for the presence of the aqueous solvent. The solvent is usually represented by a predetermined number of H_2O molecules in the bounding box which contains the target chain. The fuzzy oil drop (FOD) model, presented in this paper, follows an alternative approach, with the solvent assuming the form of a continuous external hydrophobic force field, with a Gaussian distribution. The effect of this force field is to guide hydrophobic residues towards the center of the protein body, while promoting exposure of hydrophilic residues on its surface. This work focuses on the following sample proteins: Engrailed homeodomain (RCSB: 1enh), Chicken villin subdomain hp-35, n68h (RCSB: 1yrf), Chicken villin subdomain hp-35, k65(nle), n68h, k70(nle) (RCSB: 2f4k), Thermostable subdomain from chicken villin headpiece (RCSB: 1vii), de novo designed single chain three-helix bundle (a3d) (RCSB: 2a3d), albumin-binding domain (RCSB: 1prb) and lambda repressor-operator complex (RCSB: 1lmb).

1. Introduction

A characteristic property of proteins – as well as of many other molecules found in living organisms – is their need for interaction with the aqueous environment without which such molecules are unable to perform their biological role, whatever it may be [1–6]. Traditional folding models acknowledge the presence of the solvent by including a certain quantity of water molecules in the bounding box which encapsulates the target polypeptide chain. This approach involves modeling pairwise interactions between residues and water molecules; it can also account for additional environmental stimuli, such as ionic potentials and variable pH.

Although the history of *in silico* protein structure prediction tools goes back several decades, satisfactory models remain elusive [7]. Example of folding software packages include the CHARMM tool – the first one to base the process on molecular dynamic simulations [8–10] and Gromacs [11–15], which introduces an additional viscosity parameter to reflect certain properties of the aqueous environment. Each tool implements its own parameterization and representation of water (e.g. triatomic, biatomic etc.) Interactions between water and the protein are modeled in a pairwise system, involving specific atoms which belong to the residue chain, and individual water molecules (whether tri-, bi- or monoatomic).

In contrast, the approach presented in this work treats the aqueous solvent as a continuous force field. This field is mathematically defined as a 3D Gaussian, which – according to the assumptions which underpin the fuzzy oil drop (FOD) model – represents the idealised (or theoretical) distribution of

[*] Corresponding author.
E-mail addresses: dawid.dulak@gmail.com (D. Dułak), myroterm@cyf-kr.edu.pl (I. Roterman).
[1] Currently: Schibsted Tech Polska Sp. z o. o. ul. Armii Krajowej 28, 30-150, Kraków, Poland.

https://doi.org/10.1016/j.jmgm.2018.12.008
1093-3263/

hydrophobicity in a protein molecule. The fuzzy oil drop model may be regarded as an extension of Kauzmann's discrete oil drop model [16,17], which compares the protein to a drop of oil. When immersed in water, the drop adopts a spherical shape to minimize contact with the polar solvent - what was also discussed in details in famous paper [18] - FOD replaces the discrete representation of hydrophobicity proposed by Kauzmann [16] with a continuous gradient where hydrophobicity increases along with distance from the surface of the protein and peaks at the geometric center of the molecule. It is further assumed that this distribution of hydrophobicity should characterize biologically active proteins (except for their active sites [19–21]) and that the presence of the solvent guides the folding process in such a way as to ensure internalization of hydrophobic residues along with exposure of hydrophilic residues on the surface. The shape and volume of the ellipsoid which encapsulates the protein are determined by the sigma coefficients of the Gaussian (one per axis). Depending on the mutual relations between σ_x, σ_y and σ_z, the protein may appear as a regular sphere (all three coefficients equal) or as an elongated globule [22]. Examples presented in this paper include both properly folded and those which failed, with the aim being to explain the sources of failure.

2. Materials and methods

2.1. Two-stage protein folding model

Building upon experimental research which indicates that protein folding is a multi-stage process [23–25], a model was designed which involves two stages: Early Stage (ES) and Late Stage (LS). The former intermediate is constructed on the basis of the so-called limited conformational subspace where the desired conformational forms are believed to belong to a specific area of the Ramachandran plot. This restriction is justified by relations between the preferred dihedral angles describing the bond between adjacent residues (V_i) and the resulting curvature of successive pentapeptides (R_i). Angles close to 0 produce helical twists; greater values correspond to various loops – all the way to 180°, when the structure becomes a beta sheet. This relation can be traced on the Ramachandran plot by showing that common secondary folds are found on an elliptical path, as described in Refs. [26–29]. Due to the presence of seven distinct density peaks along this path (representing various dihedral angles), a seven-character structural code can be devised [27]. This encoding further enables us to produce a contingency table, expressing the relation between structural codes and preferred secondary folds (for all possible tetrapeptides) [26]. Based on the contents of this table, it is possible to determine the starting structure (i.e. the early stage intermediate) for any given polypeptide chain.

2.2. Late stage

The goal of the Early Stage is to correctly predict the secondary conformational characteristics of a given polypeptide chain. Calculations depend only on the arrangement of residues in the chain and do not acknowledge any inter-atomic interactions or environmental factors [26–29].

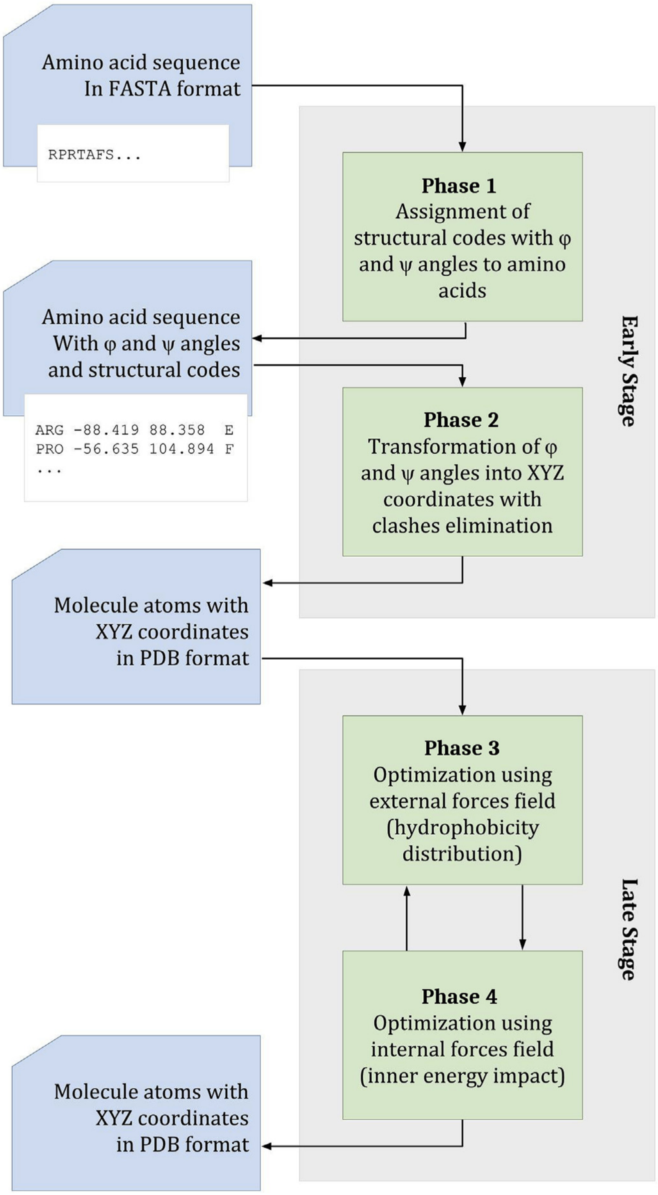
This differs greatly from the process which produces the Late Stage intermediate. At this point we are interested in the protein's tertiary conformation and rely on information concerning mutual interactions between atoms belonging to the protein and those which form the aqueous environment. To achieve this, the Early Stage intermediate is immersed in a solvent whose presence manifests itself as the aforementioned Gaussian distribution of hydrophobicity. This causes hydrophobic residues to congregate at the center of the protein body while hydrophilic residues are exposed on the surface. Optimization of hydrophobic interactions is interleaved with optimization of nonbonding (electrostatic and van der Waals) forces, as well as optimization of covalent bonds inside the model structure. The Gromacs force field and Gromacs program is applied at this point [30].

2.3. Folding workflow

Fig. 1 provides a schematic depiction of the folding process.

In Phase 1, starting with a FASTA sequence, each amino acid is assigned a pair of dihedrals (φ and ψ), based on the contingency table [26,27] and reflected by a structural code (specific fragment of the elliptical path on the Ramachandran plot [27,29]).

Subsequently (Phase 2) each amino acid is mapped to a collection of atoms in 3D space (XYZ coordinates) – φ and ψ may change to avoid steric clashes, but they are restricted to a range of values


Amino acid sequence
In FASTA format
Phase 1
Assignment of structural codes with φ and ψ angles to amino acids
Amino acid sequence
With φ and ψ angles
and structural codes
Phase 2
Transformation of φ and ψ angles into XYZ coordinates with clashes elimination
Early Stage
Molecule atoms with
XYZ coordinates
in PDB format
Phase 3
Optimization using external forces field (hydrophobicity distribution)
Phase 4
Optimization using internal forces field (inner energy impact)
Late Stage
Molecule atoms with
XYZ coordinates
in PDB format

```
RPRTAFS...
```

```
ARG -88.419 88.358  E
PRO -56.635 104.894 F
...
```

Fig. 1. Schematic depiction of the folding process. Inputs and outputs for successive simulations steps are visualized on the left. The right-hand side lists stages of the process, showing the distinction between the Early Stage and the Late Stage.

determined by their assigned structural code.

The following stage – i.e. the Late Stage – involves optimization of external (Phase 3 – hydrophobic forces) and internal forces (Phase 4 – intramolecular force fields and covalent interactions). Here, the number of iterations is determined by the user.

Besides Gromacs, which was used to optimize internal force fields, all other computational modules were developed directly by the authors' research team. Since optimization of external forces relies on so-called effective atoms (averaged-out positions of all atoms comprising the given residue), it may introduce significant steric clashes once the all-atom representation is restored. This optimization step is carried out by a software package called JDrippy and uses Rosenbrock-Palmer's minimization algorithm which would be greatly hampered by recognition of steric clashes [31,32]. This is why steric alignment is carried out only after the optimization step is complete. The same module was used to eliminate steric clashes in the Early Stage intermediate.

An additional component is used to validate the resulting structures by comparing them with structures obtained via experimental means. The comparison criteria are the same as those used in the CASP challenge [33]:

- GDT_TS (Global Distance Test – Total Score) – the main CASP evaluation criterion. An important advantage of GDT_TS is that it is not overly sensitive to the conformation of outlying fragments. It adopts values from the 0–100 range, with higher values indicating better alignment. Values below 20 are regarded as representing entirely different structures.
- TM-score (Template Modeling score) – this criterion is also relatively insensitive to local deviations and provides a good metric for comparison of results listed in Ref. [34]. It adopts values from the 0–1 range. Values greater than 0.5 indicate very good alignment.
- RMS_CA (Root Mean Square Distance for backbone Cα atoms) – one of the original structural comparison metrics. It is the only criterion used in this work which acknowledges the positions of all residues – this may be regarded as a disadvantage (results can be thrown off by poor alignment of “unimportant” structural fragments); however it also shows to what extent the resulting conformation matches the experimentally determined native fold of the protein. Lower values of RMS_CA indicate better alignment.
- QCS (Quality Control Score) – a criterion specifically constructed to mimic human assessment [35]. It focuses on the shape and placement of secondary folds. Higher values indicate better alignment.

In addition to the above, selected structures were singled out for calculation of Rood Mean Square Distance for backbone atoms (bRMSD) coefficients, facilitating comparisons with results reported in Ref. [36]. Lower values of bRMSD indicate better alignment.

Computation of the above values is described in detail in Ref. [37] as applied for structure comparion of models delivered in CASP project. GDT_TS, RMS_CA and TM-score coefficients were calculated using MaxCluster [38]; QCS – using the QCS toolkit [35]; bRMSD – using the ProFit application [39].

Results were visualized using VMD [40], while charts were plotted with Gnuplot [41] and the Highcharts library [42].

The modules are integrated by a dedicated application called DrippyAttack, which enables multiple simulations (with varying parameters) to be launched in parallel on the Zeus cluster at ACC Cyfronet AGH (a supercomputing center in Kraków, Poland). Results are parsed to generate statistics and export data in a format suitable for UI viewers.

2.4. Details of optimization of the external force field (phase 3)

As already mentioned, the JDrippy implementation makes use of the Rosenbrock-Palmer optimization algorithm [31,32]. This is a gradientless method which seeks a minimum in *n* mutually orthogonal dimensions. Initial values of *a* and *b* (parameters as they are defined in the Rosenbrock optimization method) were chosen as 3.0 and 0.9 respectively, with a 10-degree initial step length. Each of the φ, ψ and χ angles (which can be manipulated as they do not participate in rigid secondary structures – we will discuss this in detail further on) is assigned to a coordinate system axis and optimized on the basis of its adherence to the theoretical distribution of hydrophobicity (this is hereinafter referred to as the *FOD discordance function*). Under these criteria, multiples of 360 are assumed to represent identical conformations and the FOD discordance function is computed as follows:

1. A three-dimensional lattice is introduced, with points separated by $d = 5\,Å$ (*NP* represents the total number of lattice points)
2. For each point a theoretical value of hydrophobicity Ht_j is computed as follows:

$$\tilde{H}t_j = \frac{1}{\tilde{H}t_{sum}} \exp\left(\frac{-\left(x_j - \bar{x}\right)^2}{2\sigma_x^2}\right) \exp\left(\frac{-\left(y_j - \bar{y}\right)^2}{2\sigma_y^2}\right) \exp\left(\frac{-\left(z_j - \bar{z}\right)^2}{2\sigma_z^2}\right) \tag{1}$$

where:

Ht_{sum}– sum of $Ht$for all lattice points under consideration.
x, y, z – coordinates of the given point.
$\bar{x}$, $\bar{y}$, $\bar{z}$– coordinates of the point of highest hydrophobicity (center of the lattice).
σ – standard deviation.

$\bar{x}, \bar{y}, \bar{z}$ reflect the placement of the center of the ellipsoid (all three are equal to 0 at the origin of the coordinate system). σ coefficients are calculated as 1/3 of the greatest distance between any effective atom belonging the molecule and the origin of the system, once the molecule has been oriented in such a way that its greatest spatial extension coincides with a specific axis (for each axis separately).

The $1/Ht_{sum}$ coefficient ensures normalization of both distributions (empirical and theoretical) and therefore enables comparative analysis. While theoretical hydrophobicity is defined at any point within the encapsulating ellipsoid, in practice we are only interested in positions which correspond to effective atoms (averaged-out positions of all atoms comprising a given residue). Consequently, the sum has N components where N is the number of residues in the chain. Each component is the theoretical value of hydrophobicity at the position of the given “effective atom” representing the residue under consideration.

3. In addition, the observed value of hydrophobicity Ho_j is determined for each point using the following formula introduced by M. Levitt [43]:

$$H_{O_j} = \frac{1}{H_{O_{SUM}}} \sum_{i=1}^{NP} H_i^r \left[1 - \frac{1}{2}\left(7\left(\frac{r_{ij}}{c}\right)^2 - 9\left(\frac{r_{ij}}{c}\right)^4 + 5\left(\frac{r_{ij}}{c}\right)^6 - \left(\frac{r_{ij}}{c}\right)^8\right)\right] \text{for } r_{ij} \le c \tag{2}$$

where:

c – cutoff distance for hydrophobic interactions.

r_{ij} – distance between the jth lattice point and the center of the ith residue.

In both equations (eq. (1) and eq. (2)) j denotes the position of the *j-th* grid point. Ho_j is an aggregate value describing interactions with neighboring residues (indexed i) at a distance not greater than 9 Å (this distance – c – is treated as the cutoff value for hydrophobic interactions, following the original model [43]). Applying a cutoff value implies that hydrophobic interactions are considered local and depend on the position of each residue. This function is empirically determined and, according to Ref. [43], expresses the force of hydrophobic interactions. H_i^r represents intrinsic hydrophobicity (constant for each residue) according to a predetermined scale, which can be arbitrary (in our study the relevant scale is derived from Refs. [44,45]). r_{ij} is the distance between the i-th residue ("effective atom" position) and the j-th grind point while *NP* is the total number of grid points in the lattice.

The best graphic presentation of the basis of the method is given in Ref. [46] (see Fig. 1 in the cited publication).

Since both distributions are normalized (division by the summary of appropriate H values) the comparison of these two distributions is possible.

4. Fitness score is then given as:

$$\Delta H = \sum_{i=1}^{N} (Ht_i - Ho_i)^2 \quad (3)$$

Higher values of this function correspond to greater deviations from the FOD model.

The volume of the bounding box is determined prior to optimization; details can be found in the section titled *Determination of expected Phase 3 structure size*.

During the optimization process, successive dihedral angles were selected at random. This means that, having calculated the discordance parameter following a change in φ for the second amino acid in the chain, the algorithm may subsequently choose any one of the yet-untested dihedrals. This stochasticity may produce different results for the same set of input data. Consequently, each input set was processed 10 times, following which we selected the three best matches (i.e. lowest discordance) for further tests. One shall mention that the all atom representation is kept along complete procedure. It is used explicitly for internal energy calculation (Gromacs bazed steps). All atom representation is used for "effective atoms" position in the steps oriented on calculation of hydrophobic interaction in FOD based steps.

Minimization of discordance may severely disrupt secondary folds (helices and beta sheets) because it does not check whether the altered residues belong to such structures, and furthermore it focuses only on the positions of effective atoms. This is why we decided to introduce an option to treat secondary folds as rigid entities during the optimization process (but not during optimization of internal forces – as it turns out, the Early Stage intermediate typically does not include well differentiated secondary folds and furthermore its helical twists are not fully normalized; consequently, we wanted to enable Gromacs to introduce slight changes in helices and beta sheets).

The number of optimization steps was set to 100, 600, 1100, 1600 or 2100 in order to determine whether increasing the number of iterations produced better results.

2.5. Resolution of steric clashes (end of phase 3)

Since attempts to avoid steric clashes (i.e. situations where unbound atoms are placed in close proximity to one another) during the minimization process greatly reduced the efficiency of the algorithm, we instead opted to resolve such clashes only after minimization is complete. For each of the three structures produced in the preceding step 10 iterations of the clash resolution procedure were carried out. We then selected the structure with the fewest clashes as long as its volume was not significantly increased by the resolution process (this additional condition was introduced to avoid selecting structures which may be free of clashes but whose packing is deemed insufficient).

The clash resolution process is not expected to remove all clashes but to eliminate as many of them as possible given the limited processing time.

2.6. Determination of the expected phase 3 structure size

The **expected final size** (R_{final}) of the structure produced by the simulation workflow was computed in two ways:

a) Based on the formula published in Ref. [47], as the average value of dimensions of the bounding box containing the protein (for the sake of simplicity)
b) As

$$R_{final} = \sqrt[3]{10^{3.1+0.7725*\log(N)}} \quad (4)$$

where N is the number of amino acids in the protein. The originally published coefficient (3.5671) was lowered to 3.1, which yielded significantly improved predictions of the size of selected structures.

c) As the average value of dimensions of the bounding box containing the native form of the protein as supplied by Gromacs.

In most cases all three methods produced very similar results.

The **expected Phase 3 structure size** (or, more accurately, the average value of dimensions of the bounding box) was calculated as follows:

For N = 1: $R_i = R_{final}$ (5)

$$\text{For N} > 1 \; R_i = R_{start} + (i-1)\frac{R_{start} - R_{final}}{N-1} \quad (6)$$

$$R_i = R_{start} + (i-1) \cdot \frac{R_{start} - R_{final}}{N-1}$$

Where N is the number of iterations (i.e. repetitions of phases 3 and 4), i is the current iteration, R_{final} is the expected final size and R_{start} is the initial size of the Early Stage intermediate.

R_{start} may be computed in two ways:

a) MAX – maximum value of the dimensions of the ES bounding box
b) AVER – average value of the dimensions of the ES bounding box

In some cases R_{start} proved greater than R_{final}, suggesting that the optimization process increased the volume of the protein. This effect contradicts the natural properties of the folding process, however the results provide a useful baseline for comparative analyses.

We also attempted twofold repetition of i for each set of input parameters (R_i).

2.7. Details of optimization of the internal force field (phase 4)

Optimization of internal force fields was carried out using Gromacs v4.5.3, at single precision, using the Conjugate Gradients method with the following coefficients: emtol = 10.0 kJ/(mol * nm) (minimization is achieved if the maximum force present inside the structure is lower than the given threshold); emstep = 0.05 nm (initial step length).

From among the three structures produced in Phase 3 we selected the one for which atom-atom forces (as given by Gromacs) were lowest. This structure was subjected to further optimization of internal force fields. The bounding box was a dodecahedron configured to exceed the actual size of the protein by 1 nm. We applied the Gromacs amber99sb-ildn force field, which aggregates multiple internal force fields [47].

The number of optimization steps was set to 100, 600, 1100, 1600 or 2100 in order to determine whether increasing the number of iterations produced better results.

2.8. Iterations of phase 3 and phase 4

The number of successive iterations of the optimization process (Phase 3 and Phase 4) combined was set to 1, 2*, 5, 10, 10*, 15 and 30* (asterisks indicate that the process was repeated twice for each expected Phase 3 structure size).

2.9. Data

Our analysis involved proteins: Engrailed homeodomain [48] (PDB id: 1enh; Resolution 2.1 Å; R-value Free - unavailable; and R-value Work: 0.197), Chicken villin subdomain hp-35 [49], n68h (RCSB: 1yrf) (Resolution 1.07 Å; R-value work - unavailabe, R-value-free 1.161); Chicken villin subdomain hp-35 [50], k65(nle), n68h, k70(nle) (RCSB: 2f4k) (Resolution 1.05 Å, R-vaue work – unavailable, R-value free – 0.166); Thermostable subdomain from chicken villin headpiece [51] (RCSB: 1vii) (solution NMR), de novo designed single chain three-helix bundle (a3d) [52] (RCSB: 2a3d) (solution NMR), albumin-binding domain [53] (RCSB: 1prb) (solution NMR) and lambda repressor-operator complex [54] (RCSB: 1lmb; resolution – 1.8 Å, R-value work 0.189, R-value free – unavailable).

were downloaded from RCSB [55] selected due to its high conformance with the FOD model, leading us to assume that its folding process proceeds in accordance with FOD criteria [56]. Moreover, this protein is frequently used as a case study by other research teams [34,35]. The remaining proteins were selected to investigate the behavior of the proposed method for chains which vary in terms of their length and FOD conformance. Table 1 provides a summary of the basic properties of each protein, along with their origin and likely function (Table 1.).

The sizes of the proteins' native forms differed somewhat from the corresponding values of R_{final}, however the discrepancies were deemed sufficiently small to enable meaningful analysis (Table 2). For the purposes of the simulation algorithm, values were rounded to the nearest whole number (except for 1ENH where 28.82 was rounded down to 28). The examination of the size of molecule is important due to the decreasing size of 3D Gauss function during the optimization procedure. Each step of the minimization procedure is performed for smaller size of ellipsoid. The relation between ES and native size of molecules was discussed in details in Ref. [57].

Table 2
Comparison of the actual size of the native structure with a size determined solely on the basis of the chain length (size is understood as the average value of the dimensions of the protein's bounding box).

Protein	R_{final} calculated for the native form – 1.6.1.b [Å]	R_{final} produced by eq. 1.6.1.a [Å]
1ENH	28.82	30.16
1YRF	22.96	26.97
2F4K	23.73	26.97
1VII	24.32	27.17
2A3D	32.79	32.59
1PRB	31.41	30.02
1LMB	34.85	34.10

2.10. Assessment of the final structure under the FOD model

As discussed in Section 1.3, assessment of similarities between the target structure and model predictions is based only on geometric factors. The FOD model introduces a formal way to determine the fitness between the observed and theoretical distributions of hydrophobicity in a protein – this fitness is expressed by the so-called relative distance (RD) coefficient, which shows whether the target structure more closely approximates the theoretical Gaussian distribution of hydrophobicity or a reference distribution in which no concentration of hydrophobicity is observed at any point in the protein body. This method of expressing fitness score is independent of chain length and has been described in numerous publications (the detialed description in Ref. [46]). Below, we apply it to assess the similarities between our models and their respective targets.

The description provided below follows CASP naming standards, with “targets” referring to structures deposited in PDB while simulated structures are referred to as “models”.

Additionally the comparison of the size of molecule is important due to the squeezing procedure applied during optimization procesidue. The size of ellipsoid encapsulating the molecule shal be under controll since it may be changed during the calculation. The relation between size of ES intermediate and the native form is discussed in details in Ref. [57].

2.11. Comparative analysis of final structures

The parameters based on the FOD model are included in comparative analysis. The notion applied is as follows: capital O, T symbols represent the observed and theoretical distributions respectively. The indexes “m” and “t” distinguish model and target respectively. The O distribution in any case is the distribution

Table 1
The listed native structures were downloaded from www.rcsb.org in PDB format and subjected to gentle relaxation using Gromacs (100 iterations of the conjugate gradients method). Resulting forms were evaluated for their potential energy and FOD fitness.

PDB ID	FOD discordance	Potential energy	Chain length [aa]	Type	Organism
1ENH	0.00247	−8.82E+03	54	DNA binding	*Drosophila melanogaster*
1YRF	0.00383	−2.52E+03	35	Structural	*Gallus gallus*
2F4K	0.00397	−2.36E+03	35	Structural	*Gallus gallus*
1VII	0.00416	−1.52E+03	36	Actin binding	*Gallus gallus*
2A3D	0.00239	−5.26E+03	73	Synthetic	Synthetic construct
1PRB	0.00448	−1.43E+03	53	Albumin binding	*Escherichia coli*
1LMB	0.00260	−7.00E+03	87	Transcription/DNA	*Escherichia* virus Lambda

calculated according to eq. (2) while T distribution is calculated according to eq. (1). The T distribution expresses the idealised (expected) distribution ahilw O expressed the really observed distribution being the result of the residues positions. Each of them is represented by its intrinsic hydrophobicity H_i^r.

The model and target structures encapsulated in the 3D Gauss ellipsoid can be essesed In respect to their T and O distribution. The similarity of these two distributions can be measured. The Kullback-Leibler entropy D_{KL} is applied to this analysis [58].Where p_i – probability - in our case – observed hydrophobicity distribution, p_i^0– probability – in our case the observed hydrophobicity treated as reference distribution for point *i-th*, N – number of points in the profile.

The D_{KL} is the quantity of entropy character. This is why its value has no immediate interpretation. To solve this problem the second reference distribution is introduced – R. According to this distribution each residues represents equal hydrophobicity. It means there is no concentration of hydrophobicity in any point in the protein. To measure the Relative Distance the RD coefficient was introduced defined as follows:

$$RD = D_{KL}\ (O|T)\ /\ [D_{KL}(O|T) + D_{KL}(O|R)] \tag{9}$$

The *pi* is represented by O and p_i^0 by T and R distribution treated as reference distributions respectively. The interpretation of RD value expresses the similarity of O distribution versus T and R ones. The lower value of RD the closer is the O distribution versus T distribution. It means that the centric hydrophobic core is present in protein under consideration. The values of RD below 0.5 expressed such situation.

3. Results

3.1. Folding simulation

For each protein 800 folding simulations were carried out (Table 3), differing in terms of the number of steps in Phase 3 and Phase 4, iterations of the optimization process (Phase 3 + Phase 4) as well as values of R_{start} and R_{final}. Since for 2A3D two different algorithms produced identical values of R_{final}, all simulations were repeated twice to obtain the same number of results as for other proteins (800 in total).

Note that not all simulations could be carried out to completion as some of them required excessive computing time.

The percentage of results with GDT above 40, i.e. suspected to resemble the native structure – varied (depending on the protein) between 36.6% and 0%. The best results were obtained for 2F4K, while the lowest scorers were 1PRB and 1LMB. 1PRB in particular performed much worse than 2A3D. In addition, very few reasonable results were obtained for 1VII – one of the smallest proteins in the set.

Early Stage structures contained folded helices; however their placement did not always correspond to the location of helices in the native structures (Fig. 2). Moreover, twist angles differed from those observed in native proteins. Taking 1PRB as an example, we can see that the Early Stage intermediate contains two helices (instead of the expected three) and that these helices are incorrectly positioned. Similar inaccurate results are observed in 1VII, 2A3D and 1LMB Early Stage intermediates.

Results with the highest values of GDT_TS, TM-score, RMS_CA and QCS (Table 4.) were singled out for further analysis. The only exception was 1ENH (simulation no 624) which, despite poor similarity coefficients, appears more consistent than other proteins upon visual inspection. Two proteins (1YRF and 2A3D) obtained high scores in all categories, while two more (1ENH, 2F4K) were highly scored in most categories. In terms of GDT and TM-score, 1YRF and 2F4K were very strong performers while 1ENH and 2A3D received satisfactory grades. In contrast, simulations failed to meet expectations for 1PRB and 1LMB.

In addition to the summary presented in Table 4 we also computed RMSD values for 1ENH structures derived from this table, obtaining 4.374 Å, 3.370 Å and 3.514 Å for simulations no. 290, 440 and 624 respectively.

Computation of correlations between the coefficients listed in Table 4 revealed no significant regularities. Only slight correlation (0.60) was noted between RMS_CA and FOD discordance for the final structure (Fig. 3.). Considering QCS, the following was observed: while lower FOD discordance corresponded to lower RMD_CA, high values of QCS were generally retained.

Referring to Table 4:

- No simulations with fewer than 5 iterations are listed.
- Only one simulation involved 2100 Phase 4 steps, and it did not produce the best results in terms of RMS_CA/TM-score/GDT_TS/QCS.
- From among proteins with generally correct folds, only 2A3D increased in volume during Phase 3 ($R_{start} > R_{final}$); all others shrank.
- Most simulated folds exhibited better consistency with FOD (lower RD) than their corresponding native forms.
- Most correctly folded proteins had low internal energy.

It seems intuitive that Phase 3 should reduce the size of the protein; however, when analyzing GDT and Phase 3 size changes in the 800-iteration set no such correlation appears evident (correlation coefficients between −0.02 and 0.06).

Similarly, final FOD discordance did not correlate with GDT values (coefficients between −0.14 and 0.0, except for 1PRB where the value was 0.47). The same is true for internal energy (coefficients between −0.04 and 0.04) and number of steps per simulation (−0.5 to −0.04).

Fig. 4. Visual inspection of overlays of the native backbone (gray) with backbones produced by folding simulations listed in Table 1 confirms the conclusions drawn from analysis of GDT and TM:

Table 3
Folding simulation statistics for each protein in the set.

Protein	No. of simulations				
	planned	completed	completed with TM > 0.4	completed with GDT >40	completed with RMS_CA < 3.2
1ENH	800	98.6%	11.4%	23.6%	0.0%
1YRF	800	99.2%	1.1%	13.9%	4%
2F4K	800	99.5%	1.3%	36.6%	3.6%
1VII	800	99%	0%	3.8%	0%
2A3D	800	98.3%	7.1%	4.1%	0%
1PRB	800	99.5%	0%	0%	0%
1LMB	800	98.6%	0%	0%	0%

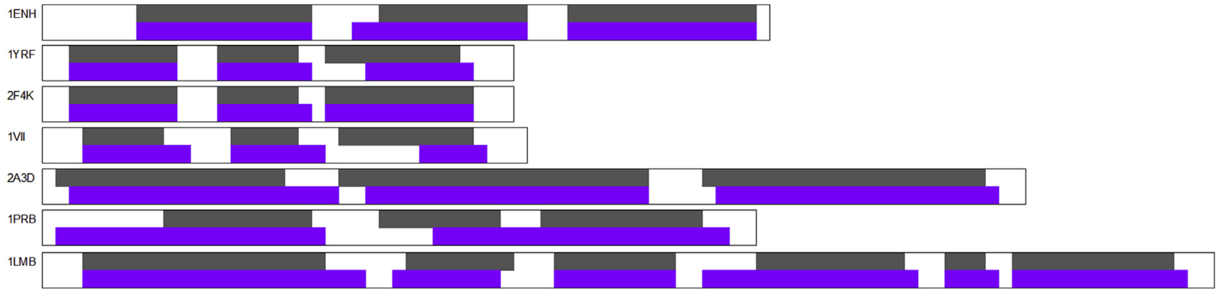


Fig. 2. Comparison of secondary folds – helices present in the native structure (gray) and the Early Stage intermediate (purple) for each protein. Each square corresponds to a single residue. Residues tagged in white do not belong to helices.

Table 4
Summary of the best results obtained for each protein in terms of GDT, TM, RMS_CA and QCS. The exception is 1ENH (simulation no. 624) which did not rank highest in any category. ***M*** next to R_{start} indicates that MAX size was applied (cf. Section 1.6.2.a), while **A** corresponds to the AVER size (Section 1.6.2.b). Lowercase labels („a" and „b") next to R_{final} values correspond to distinct computation algorithms presented in Sections 1.6.1.a and 1.6.1.b respectively.

Protein	Id	GDT	TM	RMS_CA	QCS	Internal energy	FOD discordance	Iterations	R_{start}	R_{final}	Phase 3 steps	Phase 4 steps
1ENH	290	57.87	0.514	4.542	74.04	−7736.8	0.001944	30*	(M) 61.6	(b) 28	600	100
	440	52.32	0.478	3.514	64.82	−7870.8	0.001675	15	(M) 61.6	(b) 28	1100	100
	624	51.39	0.462	3.734	61.13	323 876E2	0.001664	15	(M) 61.6	(a) 30	1600	2100
1YRF	450	70.72	0.498	2.310	70.71	−1362.4	0.002180	30*	(M) 31.6	(b) 23	1100	100
2F4K	703	64.29	0.48	2.342	57.96	−2489.6	0.003319	5	(M) 34.2	(a) 27	2100	1600
	680	33.57	0.35	2.886	61.20	−1854.0	0.003316	5	(M) 34.2	(b) 24	2100	100
1VII	207	53.47	0.37	3.518	39.21	−1945.2	0.003371	5	(A) 20.7	(b) 24	600	1110
	228	39.58	0.306	3.460	39.94	−1788.8	0.003578	5	(A) 20.7	(a) 27	600	1600
	523	50.69	0.319	4.233	54.12	−2423.52	0.003079	5	(M) 34.91	(b) 24	1600	1600
2A3D	055	48.63	0.505	3.968	65.38	−5740.0	0.002529	20*	(A) 29.6	33	100	100
1PRB	237	33.02	0.304	6.494	40.67	683 241E5	0.002609	10*	(A) 21.4	(a) 30	600	1100
	122	33.02	0.316	10.05	26.40	−3241.4	0.004732	15	(M) 37.27	(b) 31	100	1100
	423	27.83	0.260	8.061	43.02	−2588.5	0.003023	10	(M) 37.27	(a) 30	1100	1600
1LMB	625	24.71	0.291	13.13	37.65	−4264.86	0.001368	15	(A) 30.47	(a) 34	1600	100
	796	16.38	0.223	9.884	30.05	−4362.28	0.001050	30*	(A) 30.47	(a) 34	2100	600
	300	20.98	0.267	12.98	41.24	−3034.76	0.001665	15	(M) 60.15	(a) 34	600	100

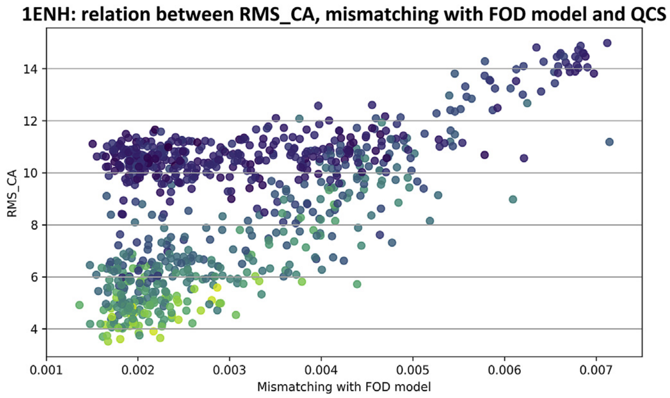


Fig. 3. Simulated folds of 1ENH: relationship between RMS_CA (vertical axis), FOD discordance (horizontal axis) and QCS (yellow – low; turquoise – moderate; navy blue – high).

accurate results were obtained for 1YRF and 2F4K; acceptable results – for 1ENH and 2A3D; poor results – for 1LMB and 1PRB.

Fig. 5 Shows the distribution of amino acid residues in native folds (gray; left) and in selected simulated structures (purple; right). Hydrophilic residues are tagged in blue while hydrophobic residues are tagged in red. Simulation results are never perfectly accurate, but appear consistent with native folds for 1YRF and 2F4K.

The charts shown in Fig. 6 and Fig. 7 reveal changes which occur during the optimization process for 1YRF (no. 450; best simulation results). Fig. 6 shows that discordance progressively decreases while internal energy, having fallen to a certain level, only reverts to that level when the structure is "spoiled" in Phase 3. Each round of Phase 3 disrupts the distribution of internal energy and *vice versa* – Phase 4 reduces the molecule's accordance with the fuzzy oil drop model.

Fig. 7 Shows the evolution of GDT and RMS_CA during the folding process. Clearly, improvements are not inexorable, and some rounds of optimization erase earlier gains even though the final result is a clear improvement over the initial structure. Notably, the best structure (in terms of the abovementioned coefficients) is produced in the 19th iteration (labeled 9.5 on the chart).

3.2. Relationship between the distribution of hydrophobicity in the model and the corresponding distribution in each protein's crystal structure

Table 5 presents the status of the analyzed structures from the point of view of their hydrophobic cores. The second column lists values of RD which indicate the presence of a prominent hydrophobic core in all proteins (RD < 0.5), even though the degree of ordering varies. Regarding simulation results (3rd column in Table 5), it seems that one protein (2A3D) did not generate a hydrophobic core consistent with the FOD model (RD > 0.5).

The values listed in Table 5 characterize the status of the hydrophobic core in the PDB target (which exhibits a prominent core). The second column shows that only 2A3D does not contain a well-formed core in the model structure.

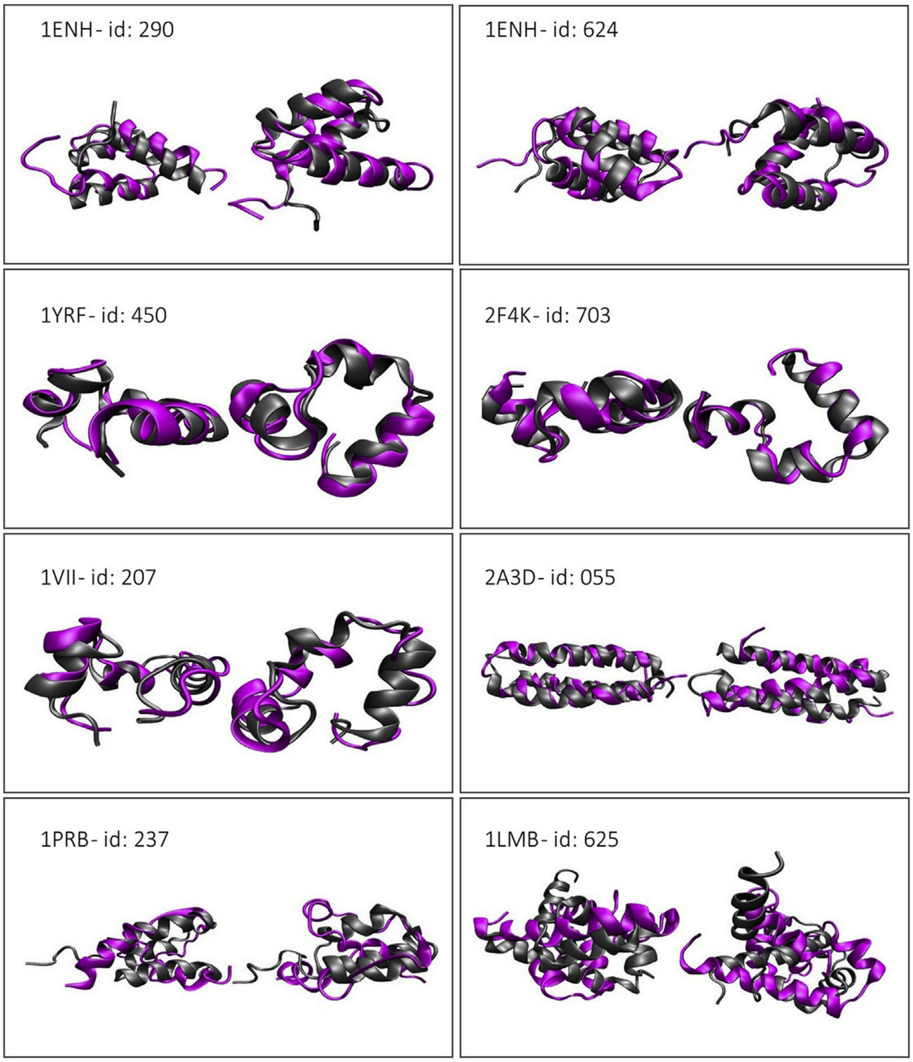


Fig. 4. Overlays of the native backbone (gray) with backbones produced by folding simulations (purple) listed in Table 1.

The accuracy of models with regard to the structure of the hydrophobic cores of target molecules may be visualized by plotting hydrophobicity distributions for the analyzed proteins. The comparison of Ot (observed distribution in target) with Om (observed distribution in model) is possible as shown in Fig. 8 (for proteins where the model is a good match for the target RD < 0.5) and Fig. 9. (for proteins where RD > 0.5, indicating low fitness between the model and the target).

The FOD results plotted in Fig. 8 may be characterized as promising since the model distributions clearly resemble the distribution found in the target. Differences are small and limited to specific residues, while the overall shape of the curve (with its maxima and minima) remains consistent with the target. Of note is the accurate result obtained for 1LMB – the longest chain in the set.

2F4K was modeled correctly despite gaps in its sequence. The same holds for 2ENH, which – in addition to tightly packed secondary folds in the central part of the molecule – includes loose fragments. Thus, obtaining a correct model for this protein can be regarded as a promising result. Similarly, 1PRB – a helical protein – produced results which remain consistent with its target (from the point of view of hydrophobicity distribution).

Fig. 9 reveals fragment which deviate from distributions of hydrophobicity in their respective targets. In the case of 2A3D these deviations are strongly localized while in 1VII and 1YRF they encompass the central part of the chain (although the distribution of local minima and maxima is largely replicated in the model). Visual analysis of the presented charts suggests that the folding process may have been terminated prematurely. Further compression of local maxima would likely have resulted in an increased concentration of hydrophobicity in the core. While RD is above 0.5, large parts of the chain remain consistent with the target.

Other form of comparative analysis is the energy calculation which was performed using the dFire and dFire2 protocol [59] using servers [60,61]. The energy of proteins under considerations are given in Table 6..

The values given in Table 6 reveal higher energy for models in respect to crystal (native) structures, however some of them are quite close. The structure of 1YRF shows the lower energy for model structure.

Taking into account the parameters and conditions for FOD optimization it suggests that the final size was probably to large to stop the optimization procedure. The stronger squeezing of the molecule (decrease of ellipsoid size) is taken under consideration for the further experiments *in silico*.

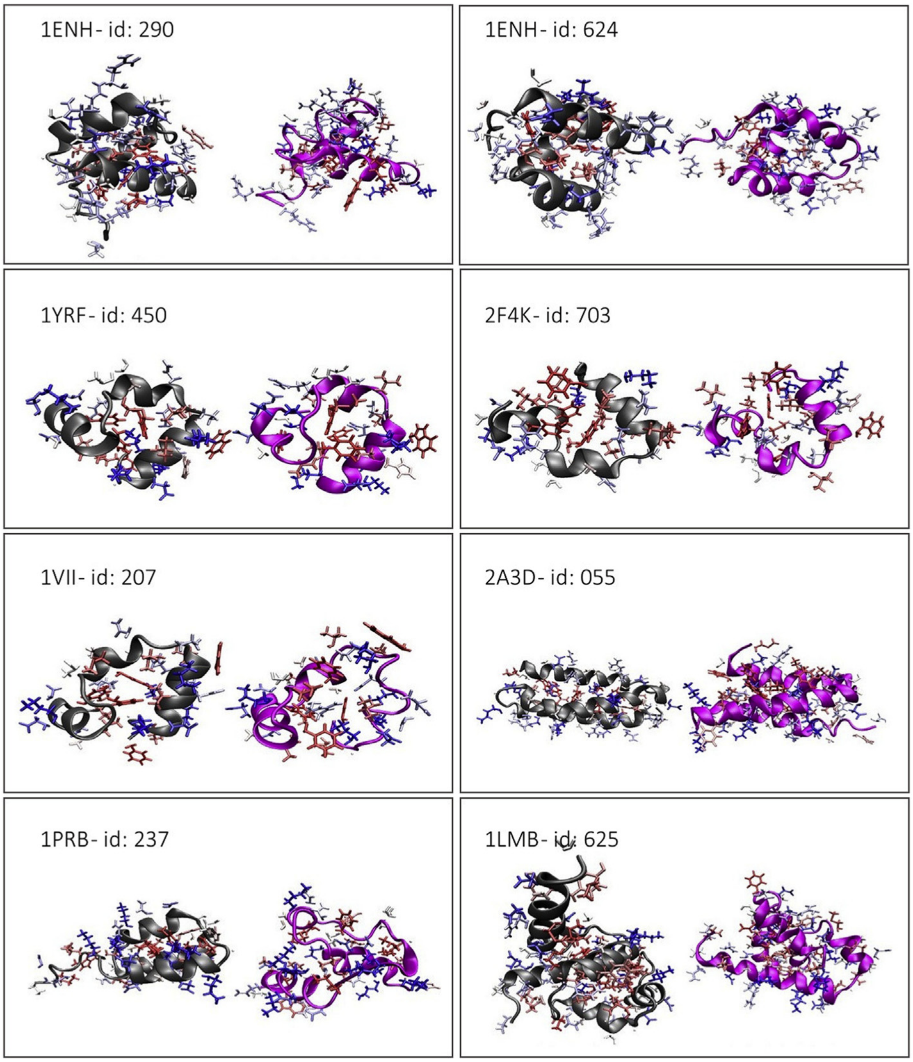


Fig. 5. Comparison of native backbone (gray) with simulated backbone (purple). Two projections are visualized.

4. Discussion

When comparing simulation results for various proteins we relied mostly on the GDT coefficient since it is of central importance in the CASP challenge. Structures characterized by optimal GDT were not always optimal in terms of QCS, TM and RMS_CA; however, it is always recommended to check whether RMS_CA is acceptably low. This is due to the fact that structures in which key fragments deviate strongly from their respective native forms may still rank high on the GDT scale. Regarding RMS_CA, these values are dependent on the length of the protein; as such, it only makes sense to compare them for chains which contain a similar number of residues (e.g. 1YRF, 2F4K and 1VII).

The best results were obtained for the smallest structures (2F4K and 1YRF), however an even more important factor was whether the Early Stage intermediate comprised correctly positioned secondary folds. For example, 1PRB, although smaller than 1ENH or 2A3D, produced significantly worse results in terms of GDT, TM, RMS_CA and QCS (its Early Stage structure was the least accurate from among all proteins in the set). Late Stage folding simulations are quite capable of handling small adjustments in the placement and shape of helical twists – see for example 1YRF.

The above study did not cover proteins whose native forms are greatly divergent from the FOD model. It seems that this discordance – as long as it does not exceed a certain threshold – does not significantly affect folding simulations. The assessment expressed by RMS_D seems to be in contradiction with the level of accordance of hydrophobic core formation. The accordance however depends on the criteria selected for comparison.

On the other hand, taking into account the immanent significance of water environment, the hydrophobic core formation shall be taken into account as the criteria for final structure assessment.

We were not able to identify a meaningful relationship between the adopted parameters (number of iterations; number of steps in Phase 3 and Phase 4 etc.) and the quality of results. Similarly, calculating FOD discordance or total internal energy does not tell us whether the model is sufficiently accurate. Thus, automatic verification of model quality with no prior knowledge about the native fold remains an open issue.

Results obtained for 1ENH (RMS_CA = 3.514 and bRMSD = 3.370 for simulation no. 440; TM-score = 0.514 for simulation no. 290) are slightly better than those reported in Ref. [35] (bRMSD = 3.40) and significantly better than those reported in Ref. [62] (RMS_CA = 4.52) but also somewhat worse (in terms of TM-score)

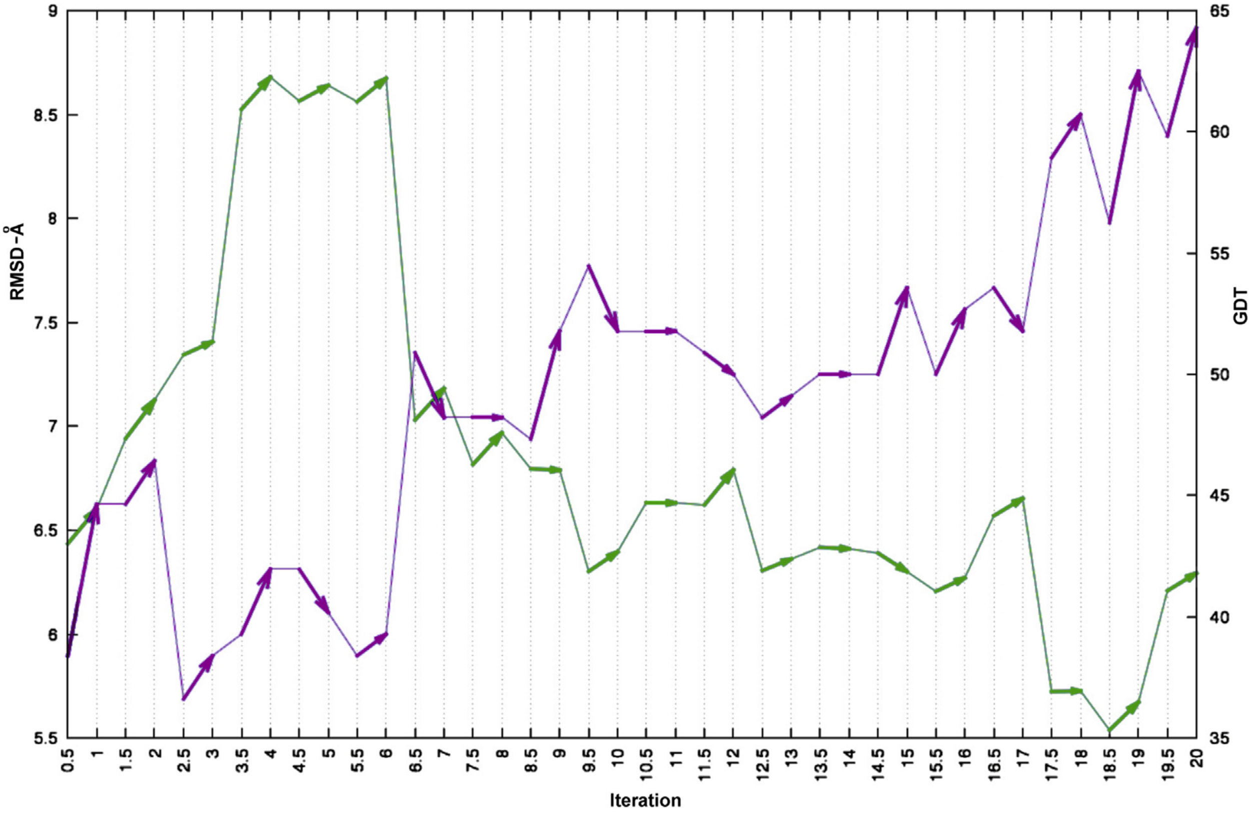


Fig. 6. Tome flow of RMS_CA [Å] (purple) and GDT for 1YRF (green).

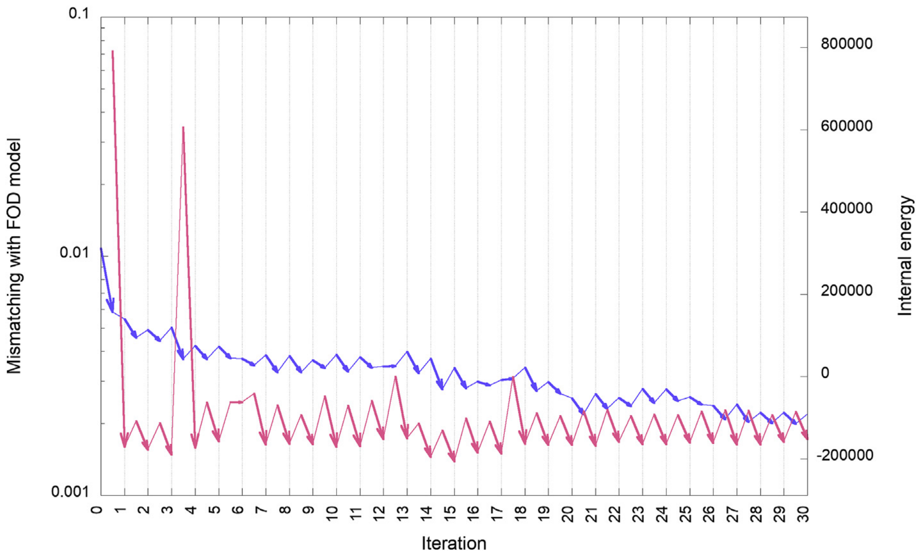


Fig. 7. Changes in internal energy (purple) and FOD accordance (blue) for 1YRF (simulation no. 450). Blue arrows mark point to the effects of Phase 3 optimization (JDrippy) while purple arrows reveal Phase 4 changes (GROMACS). Each iteration shown on the chart corresponds to one change in final size (Phase 3). Since this parameter changes only every other step, "fractional" iterations were introduced.

than those reported in Ref. [33] (TM-score = 0.67).

Folding simulations conducted in the presence of water indicates that the solvent plays an important role in this process. It is, however, important to ensure that the Early Stage structure includes the correct secondary folds since the Late Stage does not redefine such structures – it merely optimizes their shape and location by eliminating potential steric clashes.

It should also be noted that the gradient minimization method employed in Phase 4 (Late Stage) does not entirely eliminate structures which are ruled out by laws of physics and chemistry and does not alter them to produce correct structures. A solution to this problem may involve molecular dynamics simulations, which are

Table 5
Values of RD for model structures (produced by folding simulations) and their PDB counterparts (targets).

PROTEIN (PDB ID)	TARGET	MODEL
	T_t-O_t-R_t	T_m-O_m-R_m
1LMB	0.462	0.413
2F4K	0.287	0.384
1ENH	0.376	0.363
1PRB	0.338	0.437
2A3D	0.419	0.519
1VII	0.282	0.429
1YRF	0.246	0.330

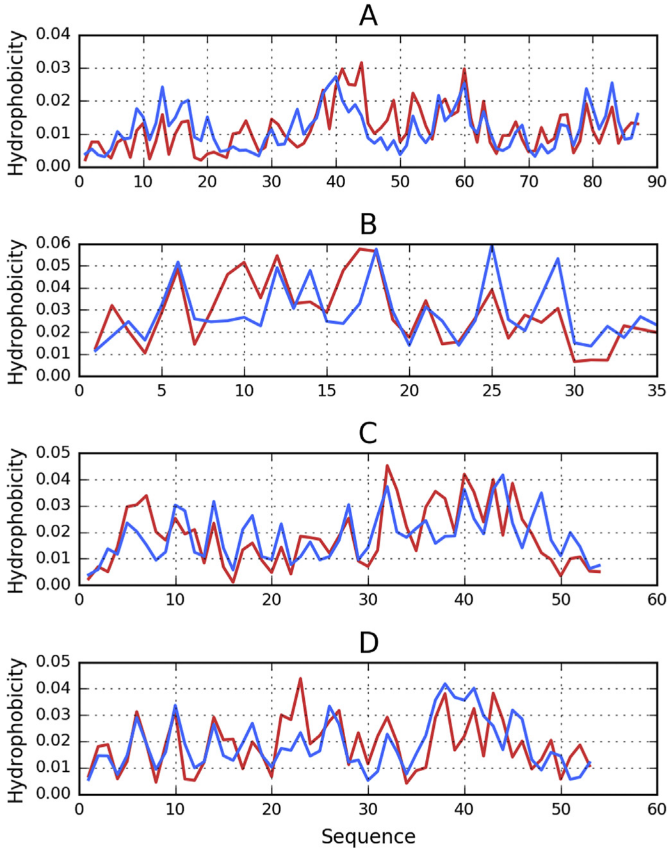


Fig. 8. Observed distributions of models which exhibit deviations from their respective targets - O_t/O_m – proteins with precision expressed by RD < 0.5. Red line – O_m (observed in model), blue line – O_t (observed in target). A – 1LMB, B – 2F4K, C – 1ENH, D – 1PRB

supported by Gromacs, either in Phase 4 or at the very end of the simulation workflow – this should further improve the accuracy and stability of models. Molecular dynamics algorithms may also be applied to automatically prune weak results.

The promising results of Gromacs molecular dynamics simulations with a solvent density model consistent with the 3D Gaussian [46] should be acknowledged. This model mimics progressive restriction of conformational freedom corresponding to increases in density and correlates with the model used in Late Stage Phase 3. Its principal advantage is that it greatly reduces computational complexity. The model may be applied in order to further improve the folding simulation workflow.

Comparison of results with those obtained by other teams [62] is not in our favor; however, given the incomparably greater experience possessed by these teams, the presented method may still be regarded with cautious optimism.

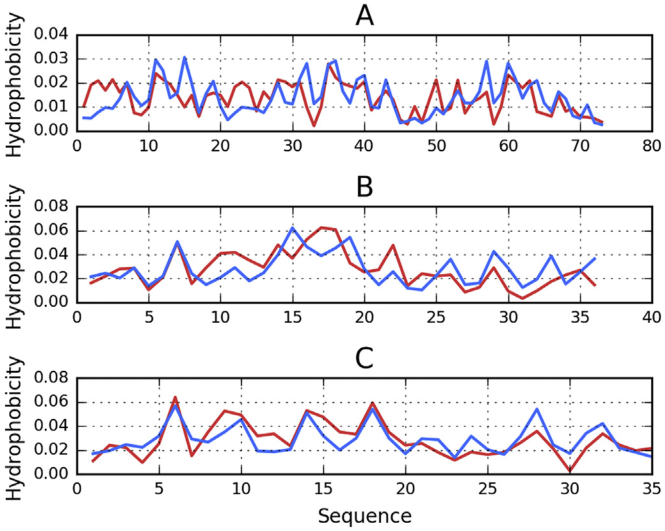


Fig. 9. Observed distributions of models which exhibit deviations from their respective targets - O_t/O_m (RD > 0.5). Frames mark highly discordant fragments. Red line –O_m (observed in model), blue line – O_t (observed in target). A – 2A3D, B–1VII, C – 1YRF

Table 6
The energy values are received using dFire and dFire2 protocols for native structure (as available in PDB) and for models.

		dFire	dFire2	N residues	N atoms
1ENH	Native	−137.5	−83.60	54	466
	Model	−88.97	−61.95	54	466
1LMB	Native	−438.70	−289.04	179	1378
	Model	−134.50	−99.81	87	669
1PRB	Native	−85.93	−55.60	53	418
	Model	−64.85	−46.76	53	408
1VII	Native	−51.86	−35.82	36	294
	Model	−46.53	−34.14	36	294
1YRF	Native	−47.23	−44.48	35	288
	Model	−49.85	−34.85	34	278
2A3D	Native	−136.17	−93.86	73	571
	Model	−113.43	−79.80	72	561
2F4K	Native	−64.11	−40.57	33	270
	Model	−46.19	−34.48	34	276

The presented work relies on assessment of model accuracy on the basis of CASP metrics (GDT_TS, TM-score, QCS, RMS_CA and bRMSD). Nevertheless, it must be emphasized that no single unambiguous accuracy criterion currently exists. Methods employed in CASP (with the exception of FlexE) focus on the geometry and topography of the model structure. However, geometric similarity is not the sole criterion which should be taken into account, and the method proposed in Ref. [36], which makes use of the FOD model, may provide a useful auxiliary criterion. This work presents results obtained using this approach both with regard to the native fold and the outcomes of various simulations.

The need to acknowledge the aqueous environment stems from a set of observations which suggest that this environment is not accurately represented in current algorithms. In particular, the links between tertiary conformational stability and the presence of a hydrophobic core and disulfide bonds are discussed in Ref. [63], showing that both factors may either compound or counteract each other.

The search for new folding simulation models is spurred by the perceived lack of sufficient accuracy among existing tools, despite over 50 years of development [7]. Emphasizing the role of the aqueous solvent is in line with recent studies regarding the influence of water upon the activity and conformational preferences of protein chains [64,65]. The WeFold team [62] suggests that it may be useful to turn to models which have not heretofore been successful, and attempt to combine their strengths with the strengths of existing approaches in hopes of obtaining more accurate results than those produced by any individual platform or tool. The detailed discussion of the folding process with the participation of so called burial potential [64]. The concentration of hydrophobic residues in the center of protein body is the result of external influence of water environment. The model presented in Ref. [65] treats the concentration of hydrophobic residues in the central part of proteins as the results of inter-residual interaction of hydrophobic character.

The presented here approach to folding simulations based on the FOD model appears encouraging – as evidenced by Table 6, which shows strong similarities between the distributions of hydrophobicity in the models and their respective targets. We intend to continue developing this method (as already postulated in Ref. [66]) by introducing multicriteria optimization where both force fields – internal (nonbonding interactions) and external (FOD model, representing the influence of the aqueous solvent) remain in balance, to ensure the production of a compromise structure.

The lowest accordance as received for 1VII is probably due to its length which is the shortest one in the set. According to Ref. [18] polypeptide chain below certain length is not able to generate the hydrophobic core due to the lower number of degrees of freedom.

Results discussed in this work can be found under the following URLs:

For 1ENH
http://protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/1ENH-4.5.3-
s-dm.
http://protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/1ENH-4.5.3-%0Ds-dm.

For 1YRF:
http://protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/1YRF-
gromacs-dm.
http://protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/1YRF-%0Dgromacs-dm.

For 2F4K:
http://1protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/2F4K-4.5.3-
s-dm.
http://1protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/2F4K-4.5.3-%0Ds-dm.

For 1VII:
http://protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/1VII-4.5.3-s-
dm.

For 2A3D:
http://protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/2A3D-4.5.3-
s-dm.
http://protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/2A3D-4.5.3-%0Ds-dm.

For 1PRB:
http://protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/1PRB-4.5.3-
s-dm.
http://protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/1PRB-4.5.3-%0Ds-dm.

For 1LMB:
http://1protein-folding.plgrid.pl/simulationsSummary.php?dir=plgtomanek/1LMB-4.5.3-s-dm.

5. Conclusions

The discussed FOD model is an attempt to present the influence of water environment, which can be compared with others like implicit solvent model [67–70]. The implicit solvent model transforms the exposure of hydrophilic/hydrophobic residues into thermodynamic parameters, while FOD model treats the environment as the active partner in folding process directing this process toward hydrophobic residues concentration in the center of molecule with hydrophilic residues exposure on the surface. This way the FOD model represents an attempt to define the mechanism of folding process. The results show n in this paper look promising taking into account short history of its application. Particularly, the discussion of misfolding proteins (amyloids) makes the method promising [71,72]. Planed application of multicriteria optimization procedure is expected to upgrade the FOD method [73].

Acknowledgements

The authors wish to thank Mariusz Sterzel, Tomasz Szepieniec and Klemens Noga for useful advice and discussions, Maciej Czuchry for help with deploying applications on the PLGrid infrastructure, and Piotr Nowakowski for translation services. We are also indebted to Anna Śmietańska for technical support. Many thanks also to Prof Cezary Czaplewski for help in preparation of the final form of the paper.

This research was supported in part by the PLGrid Infrastructure and the PLGrid Plus project. Financial support was provided under the Jagiellonian University Medical College grant no. K.ZDS/006363.

References

[1] A. Ben-Naim, One-dimensional model for water and aqueous solutions. IV. A study of "hydrophobic interactions", J. Chem. Phys. 129 (10) (2008) 104506.

[2] A. Ben-Naim, Myths and verities in protein folding theories: from Frank and Evans iceberg conjecture to explanation of the hydrophobic effect, J. Chem. Phys. 139 (16) (2013) 165105.

[3] A. Ben-Naim, Theoretical aspects of self-assembly of proteins: a Kirkwood-Buff-theory approach, J. Chem. Phys. 138 (22) (2013) 224906.

[4] A. Ben-Naim, On the so-called gibbs paradox, and on the real paradox, Entropy 9 (3) (2007) 132–136.

[5] A. Ben-Naim, Theoretical aspects of pressure and solute denaturation of proteins: a Kirkwood-buff-theory approach, J. Chem. Phys. 137 (23) (2012) 235102.

[6] A. Ben-Naim, Statistical Thermodynamics for Chemists and Biochemists, Plenum Press, New York, 1992, pp. 459–559.

[7] K.A. Dill, J.L. MacCallum, The protein-folding problem, 50 years on, Science 338 (6110) (2012) 1042–1046.

[8] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J.E. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, et al., CHARMM General Force Field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields, J. Comput. Chem. 31 (4) (2010) 671–690.

[9] D.A. Case, T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, R. Woods, The Amber biomolecular simulation programs, J. Comput. Chem. 26 (16) (2005) 1668–1688.

[10] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, J. Am. Chem. Soc. 117 (19) (1995) 5179–5197.

[11] W.F. van Gunsteren, H.J.C. Berendsen, Groningen Molecular Simulation (GROMOS) Library Manual, BIOMOS, Groningen, 1987, pp. 1–221.

[12] W.F. van Gunsteren, H.J.C. Berendsen, Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry, Angew Chem. Int. Ed. Engl. 29 (9) (1990) 992–1023.

[13] W.F. van Gunsteren, S.R. Billeter, A.A. Eising, P.H. Hunenberger, P. Kruger, A.E. Mark, W.R. Scott, I.G. Tironi, Biomolecular Simulation: the GROMOS96 Manual and User Guide, Verlag der Fachvereine, Zurich, 1996.

[14] W.F. van GunsterenF, The GROMOS Software for Biomolecular Simulation, 2012, Aug. http://www.gromos.net.

[15] W.F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D.P. Geerke, A. Glattli, P.H. Hunenberger, et al., Biomolecular modelling: goals, problems, perspectives, Angew. Chem. Int. Ed. 45 (25)

(2006) 4064–4092.

[16] W. Kauzmann, Some factors in the interpretation of protein denaturation, Adv. Protein Chem. 14 (C) (1959) 1–63.

[17] L. Konieczny, M. Brylinski, I. Roterman, Gauss function based model of hydrophobicity density in proteins, Silico Biol. 6 (1–2) (2006) 15–22.

[18] K.A. Dill, Dominant forces in protein folding, Biochemistry 29 (31) (1990) 7133–7155.

[19] K. Prymula, T. Jadczyk, I. Roterman, Catalytic residues in hydrolases: analysis of methods designed for ligand-binding site prediction, J. Comput. Aided Mol. Des. 25 (2) (2011) 117–133.

[20] M. Banach, L. Konieczny, I. Roterman, Ligand-binding site recognition, in: Irena Roterman-Konieczna (Ed.), Protein Folding Woodehead Publishing (Currently Elsevier), 2012, pp. 80–94.

[21] B. Kalinowska, M. Banach, Z. Wiśniowski, L. Konieczny, I. Roterman, Is the hydrophobic core a universal structural element in proteins? J. Mol. Model. 23 (7) (2017) 205.

[22] I. Roterman, M. Banach, L. Konieczny, Application of the fuzzy oil drop model describes amyloid as a ribbonlike micelle, Entropy 19 (4) (2017) 167.

[23] T.E. Creighton, Protein folding, Biochem. J. 270 (1) (1990) 1–16.

[24] T.L. Religa, J.S. Markson, U. Mayor, S.M. Freund, A.R. Fersht, Solution structure of a protein denatured state and folding intermediate, Nature 437 (2005) 1053–1056.

[25] C. Bystroff, Y. Shao, Modeling protein folding pathways, in: J.M. Bujnicki (Ed.), Practical Bioinformatics, Springer, Heidelberg, 2004, pp. 97–122.

[26] B. Kalinowska, A. Krzykalski, I. Roterman, Contingency Table Browser - prediction of early stage protein structure, Bioinformation 11 (10) (2015) 486–488.

[27] W. Jurkowski, M. Brylinski, L. Konieczny, Z. Wisniowski, I. Roterman, Conformational subspace in simulation of early-stage protein folding, Proteins 55 (1) (2004) 115–127.

[28] B. Kalinowska, P. Fabian, K. Stąpor, I. Roterman, Statistical dictionaries for hypothetical in silico model of the early-stage intermediate in protein folding, J. Comput. Aided Mol. Des. 29 (7) (2015) 609–618.

[29] W. Jurkowski, Z. Baster, D. Dułak, I. Roterman, The early-stage intermediate, in: Irena Roterman-Konieczna (Ed.), Protein Folding, Woodhead Publishing (currently Elsevier, 2012, pp. 1–20.

[30] H.J.C. Berendsen, D. van der Spoel, R. van Drunen, R. Gromacs, A message-passing parallel molecular dynamics implementation, Comput. Phys. Commun. 91 (1–3) (1995) 43–56, https://doi.org/10.1016/0010-4655(95)00042-e.

[31] H.H. Rosenbrock, An automatic method for finding the greatest or least value of a function, Comput. J. 3 (3) (1960) 175–184.

[32] J.R. Palmer, An improved procedure for orthogonalising the search vectors in rosenbrock's and swann's direct search optimisation methods, Comput. J. 12 (1) (1969) 69–71.

[33] CASP Official Website, 2018, May. http://predictioncenter.org.

[34] D. Bhattacharya, UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling, Bioinformatics 32 (18) (2016) 2791–2799.

[35] Q. Cong, L.N. Kinch, J. Pei, S. Shi, V.N. Grishin, W. Li, N.V. Grishin, An automatic method for CASP9 free modeling structure prediction assessment, Bioinformatics 27 (24) (2011) 3371–3378.

[36] A. Verma, W. Wenzel, A free-energy approach for all-atom protein simulation, Biophys. J. 96 (9) (2009) 3483–3494.

[37] M. Gadzala, B. Kalinowska, M. Banach, L. Konieczny, I. Roterman, Determining protein similarity by comparing hydrophobic core structure, Heliyon [serial online] 3 (2) (2017), e00235.

[38] A. Herbert, MaxCluster - a Tool for Protein Structure Comparison and Clustering, 2018, May. www.sbg.bio.ic.ac.uk/~maxcluster/index.html.

[39] A.C.R. Martin, C.T. Porter, ProFit, 2018, May. http://www.bioinf.org.uk/software/profit/.

[40] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, J. Mol. Graph. 14 (1) (1996) 33–38.

[41] http://www.gnuplot.info (2018, May).

[42] Interactive JavaScript charts for your webpage | Highcharts. https://www.highcharts.com, 2018, May.

[43] M. Levitt, A simplified representation of protein conformations for rapid simulation of protein folding, J. Mol. Biol. 104 (1) (1976) 59–107.

[44] M. Brylinski, M. Kochanczyk, L. Konieczny, I. Roterman, Sequence-structure-function relation characterized in silico, Silico Biol. 6 (6) (2006) 589–600.

[45] M. Brylinski, L. Konieczny, I. Roterman, SPI-structure predictability index for protein sequences, Silico Biol. 5 (3) (2005) 227–237.

[46] B. Kalinowska, M. Banach, L. Konieczny, I. Roterman, Application of divergence entropy to characterize the structure of the hydrophobic core in DNA interacting proteins, Entropy 17 (3) (2015) 1477–1507.

[47] The Force Field Jungle, 2018, May. http://www.cgmartini.nl/index.php/blog/265-comparingforcefields.

[48] N.D. Clarke, C.R. Kissinger, J. Desjarlais, G.L. Gilliland, C.O. Pabo, Structural studies of the engrailed homeodomain, Protein Sci. 3 (10) (1994) 1779–1787, 1ENH.

[49] T.K. Chiu, J. Kubelka, R. Herbst-Irmer, W.A. Eaton, J. Hofrichter, D.R. Davies, High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein, Proc. Natl. Acad. Sci. U. S. A. 102 (21) (2005) 7517–7522, 1YRF.

[50] J. Kubelka, T.K. Chiu, D.R. Davies, W.A. Eaton, J. Hofrichter, Sub-microsecond protein folding, J. Mol. Biol. 359 (3) (2006) 546–553, 2F4K.

[51] C.J. McKnight, P.T. Matsudaira, P.S. Kim, NMR structure of the 35-residue villin headpiece subdomain, Nat. Struct. Biol. 4 (3) (1997) 180–184, 1VII.

[52] S.T. Walsh, H. Cheng, J.W. Bryson, H. Roder, W.F. DeGrado, Solution structure and dynamics of a de novo designed three-helix bundle protein, Proc. Natl. Acad. Sci. U. S. A. 96 (10) (1999) 5486–5491, 2A3D.

[53] M.U. Johansson, M. de Château, M. Wikström, S. Forsén, T. Drakenberg, L. Björck, Solution structure of the albumin-binding GA module: a versatile bacterial protein domain, J. Mol. Biol. 266 (5) (1997) 859–865, 1PRB.

[54] L.J. Beamer, C.O. Pabo, Refined 1.8 A crystal structure of the lambda repressor-operator complex, J. Mol. Biol. 227 (1) (1992) 177–196, 1LMB.

[55] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliiand, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242. http://www.rcsb.org/.

[56] I. Roterman, L. Konieczny, M. Banach, W. Jurkowski, Intermediates in the protein folding process: a computational model, Int. J. Mol. Sci. 12 (8) (2011) 4850–4860.

[57] M. Bryliński, L. Konieczny, I. Roterman, Is the protein folding an aim-oriented process? Human haemoglobin as example? Int. J. Bioinf. Res. Appl. 3 (2) (2007) 234–260.

[58] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86, https://doi.org/10.1214/aoms/1177729694. MR 0039968.

[59] H. Zhao, J. Wang, Y. Zhou, Y. Yang, Predicting DNA-binding proteins and binding residues by complex structure prediction and application to human proteome, PLoS One 9 (2014), e96694.

[60] http://sparks-lab.org/tools-dfire.html (Oct 14, 2018).

[61] http://sparks-lab.org/yueyang/download/index.php (Oct 14, 2018).

[62] G.A. Khoury, A. Liwo, F. Khatib, H. Zhou, G. Chopra, J. Bacardit, L.O. Bortot, R.A. Faccioli, X. Deng, Y. He, et al., WeFold: a coopetition for protein structure prediction, Proteins 82 (9) (2014) 1850–1868.

[63] M. Banach, B. Kalinowska, L. Konieczny, I. Roterman, Role of disulfide bonds in stabilizing the conformation of selected enzymes—an approach based on divergence entropy applied to the structure of hydrophobic core in proteins, Entropy 18 (3) (2016) 67.

[64] P. Das, D. Kapoor, K.T. Halloran, R. Zhou, C.R. Matthews, Interplay between drying and stability of a TIM barrel protein: a combined simulation-experimental study, J. Am. Chem. Soc. 135 (5) (2013) 1882–1890.

[65] O.V. Galzitskaya, D.N. Ivankov, A.V. Finkelstein, Folding nuclei in proteins, FEBS Lett. 489 (2–3) (2001) 113–118.

[66] I. Roterman, L. Konieczny, M. Banach, D. Marchewka, B. Kalinowska, Z. Baster, M. Tomanek, M. Piwowar, Simulation of the protein folding process, in: A. Liwo (Ed.), Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes: from Bioinformatics to Molecular Quantum Mechanics, vol. 1, Springer, Berlin, 2014, pp. 599–638.

[67] D.C. Ferreira, M.G. van der Linden, L.C. de Oliveira, J.N. Onuchic, A.F. de Araújo, Information and redundancy in the burial folding code of globular proteins within a wide range of shapes and sizes, Proteins 84 (4) (2016) 515–531, https://doi.org/10.1002/prot.24998.

[68] F.M. Richards, Areas, volumes, packing and protein structure, Annu. Rev. Biophys. Bioeng. 6 (1977) 151–176, https://doi.org/10.1146/annurev.bb.06.060177.001055.

[69] B. Roux, T. Simonson, Implicit solvent models, Biophys. Chem. 78 (1–2) (1999) 1–20, https://doi.org/10.1016/S0301-4622(98)00226-9.

[70] C.G. Ricci, B. Li, L.T. Cheng, J. Dzubiella, J.A. McCammon, Tailoring the variational implicit solvent method for new challenges: biomolecular recognition and assembly, Front. Mol. Biosci. 5 (2018) 13, https://doi.org/10.3389/fmolb.2018.00013, doi: 10.3389/fmolb.2018.00013. eCollection 2018.

[71] D. Dułak, M. Gadzała, M. Banach, M. Ptak, Z. Wiśniowski, L. Konieczny, I. Roterman, Filamentous aggregates of tau proteins fulfil standard amyloid criteria provided by the fuzzy oil drop (FOD) model, Int. J. Mol. Sci. 19 (10) (2018), https://doi.org/10.3390/ijms19102910 pii: E2910.

[72] D. Dułak, M. Banach, M. Gadzała, L. Konieczny, I. Roterman I, Structural analysis of the Aβ(15-40) amyloid fibril based on hydrophobicity distribution, Acta Biochim. Pol. (2018), https://doi.org/10.18388/abp.2018_2647.

[73] L. Konieczny, I. Roterman, Conclusion, in: Irena Roterman-Konieczna (Ed.), Protein Folding, Woodhead Publishing (Currently Elsevier), 2012, pp. 191–201.