J Mol Model (2013) 19:4259–4269
DOI 10.1007/s00894-013-1909-6

ORIGINAL PAPER

Hypothetical in silico model of the early-stage intermediate in protein folding

Barbara Kalinowska · Paweł Alejster · Kinga Sałapa · Zbigniew Baster · Irena Roterman

Received: 7 January 2013 / Accepted: 3 June 2013 / Published online: 28 June 2013

Abstract This paper presents a method for determining the structure of the early stage (ES) intermediate in the multistage protein folding process. ES structure is modeled on the basis of a limited conformational subspace of the Ramachandran plot. The model distinguishes seven structural motifs corresponding to seven local probability maxima within the limited conformational subspace. Three of these are assigned to well-defined secondary structures, while the remaining four are found to represent various types of random coils. The presented heuristic approach also provides insight into the reasons behind incorrect predictions occurring when the folding process depends on external factors (e.g., ligands, ions or other proteins) rather than on the characteristics of the backbone itself. The accuracy of the presented method is estimated at around 48 %.

Keywords Early stage · Intermediates · Protein folding · Secondary structure

Introduction

Although the multi-stage nature of protein folding has been confirmed experimentally [1], experimental research into the structure of early stage intermediates remains scant. In silico models are thus needed to supply adequate starting structures for folding simulations. Several authors have recently presented experimentally determined structures which they claim to correspond to early stage intermediates [2].

Many traditional in silico structure prediction methods depend on a set of starting structures, subjected to energy minimization algorithms in hope of arriving at the native form of the analyzed protein. Preparation of such starting structures depends on the model in question, with important differences separating Darwinian and Boltzmann-based approaches [3]. The former model relies on preserving structural similarities as the chain undergoes evolutionary, sequential changes, while the latter treats folding as a spontaneous process triggered by the chain's natural propensity to seek out its global free energy minimum. Determining starting structures is much easier in the Darwinian approach due to the existence of protein homologues whose structure is well known. Pasting together structural motifs from sequentially similar fragments yields useful structures which can then be subjected to energy minimization. The Boltzmann approach is far more challenging and typically relies on Monte Carlo simulations to stochastically select Φ and Ψ angles for a given chain. Some Boltzmann algorithms depends on databases which contain data on short peptide fragments (tripeptides or 9-peptides—such as in the Rosetta package [4]).

The model presented in this paper attempts to simulate early stage intermediates by referring to a limited conformational subspace established within the bounds of the Ramachandran plot [5]. The subspace assumes the form of an elliptical path whose shape and placement are the result of geometric analysis of the polypeptide chain (Fig. 1). The path traverses all fragments of the map which correspond to specific structural motifs. Assuming that the theoretical model is correct, generation of the early stage intermediate may take on two forms. In the *step back* procedure the crystalline form undergoes changes intended to reverse the folding process, replacing the values of Φ and Ψ with corresponding pairs of early-stage angles (Φ_e and

B. Kalinowska · P. Alejster · K. Sałapa · I. Roterman (✉)
Department of Bioinformatics and Telemedicine, Jagiellonian University—Medical College, Lazarza 16, 31-530 Krakow, Poland
e-mail: myroterm@cyf-kr.edu.pl

B. Kalinowska · P. Alejster
Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, Reymonta 4, 30-059 Krakow, Poland

Z. Baster
Faculty of Physics, University of Science and Technology (AGH), Al. Mickiewicza 30, 30-059 Krakow, Poland

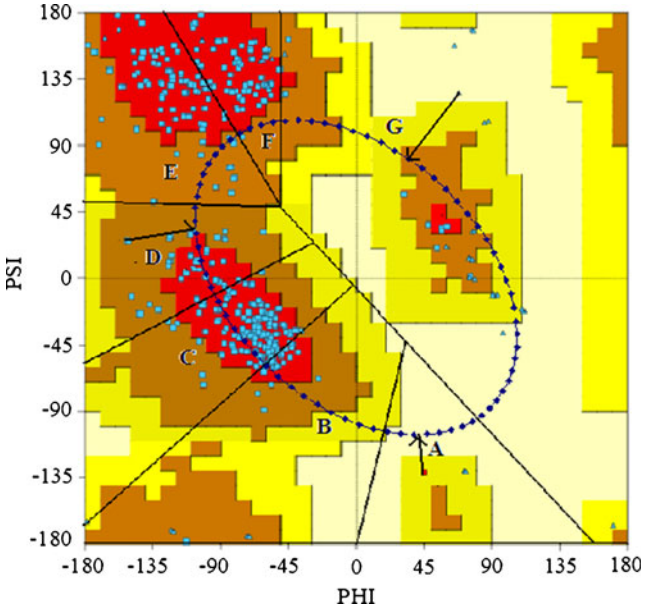

Fig. 1 The *step-back* procedure assigns to each pair of angles $\{\Phi, \Psi\}$ a corresponding pair $\{\Phi_e, \Psi_e\}$ which lies on the elliptical path. Letter codes (*A*, *B*, *C*, *D*, *E*, *F* and *G*) denote local probability peaks (Fig. 2). The derivation of the elliptical path which represents the limited conformational subspace is further explained in [5, 11]. The figure shows three sample pairs of angles and their subspace counterparts

Ψ_e) belonging to the limited conformational subspace. From an algorithmic point of view, each pair of dihedral angles $\{\Phi, \Psi\}$ is matched to a corresponding early-stage pair $\{\Phi_e, \Psi_e\}$ which lies on the elliptical path and is closest to the original pair (see [5–9] for a more detailed description of this process and refer to Fig. 1). Given this distribution of dihedral angles, the limited conformational subspace is partitioned into seven sections, each centered upon a distinct local probability maximum, labeled *A* through *G*. This process established the structural alphabet for the early stage intermediate (Fig. 2) [10].

Applying the above algorithm to a nonredundant set of proteins produces a contingency table which lists the relations between structural motifs and peptide sequences (in our case, the base sequence is a tetrapeptide fragment). Given the number of possible code variations and structural motifs, the size of the contingency table is 160,000×2401. Each cell lists the probability of encountering a specific structural motif for a given sequence of peptides (see Table 1) [10]. The other possible approach, called the *step forward* model, applies the contingency table directly to assign structural motifs to tetrapeptide fragments within the limited conformational subspace. This paper compares both approaches in order to highlight the most common misconceptions associated with the *step forward* procedure. The *step back* algorithm is treated as a baseline when determining the scope of simulation errors and inaccuracies [11]. It should be noted that the limited conformational subspace is—in itself—not a novel concept as theoretical and experimental considerations have led others to suggest similar approaches [12]. The presented derivation based on the geometric model of the polypeptide chain [11] is merely one of many attempts to establish such a subspace.

The effectiveness of the presented early stage intermediate generation method is verified on the basis of a set of protein chains extracted from the Protein Data Bank (PDB). The presented study complements the outcomes of these simulations as a crucial step in the protein folding process [13]. The goal of the early stage analysis step, presented here, is to assess the effectiveness of the proposed model when applied to raw amino acid sequences. This paper shows where the model succeeds and where it fails; it also explains the reasons behind simulation failures (on either stage of the folding process). The main aim of early stage model is to deliver the structural forms for further energy minimization procedures (computational interpretation) and to deliver structural forms which mimic the initial steps of folding process (biological interpretation).

Materials and methods

Data

The testing subset of 250 protein chains has been chosen randomly from nonredundant set of protein structures from PDB. The teaching set of nonredundant set of protein structures has been selected from PDB on a basis of data obtained in December 2011 by means of the BLASTClust tool for protein sequences characterized by sequence identity not higher than 95 %. The testing subset of protein chains is 1 % of the whole nonredundant data basis of proteins. The teaching set did not contain the proteins belonging to test set.

Early stage model

As highlighted above, the in silico folding model applied in our work can be divided into two stages: the early stage (ES) and the late stage (LS) (see [13] for a thorough description of the model). The early stage is simulated by adopting a limited conformational subspace, corresponding to an elliptical path on the Ramachandran plot. For more information on how this subspace is derived refer to [13]. An important property of the presented elliptical path is that it traverses all areas of the plot which correspond to well defined secondary structural motifs (Fig. 1).

The *step back* procedure

The *step back* procedure relies on translating each pair of dihedral angles $\{\Phi, \Psi\}$ into its corresponding "image" which

lies on the elliptical path (limited conformational subspace), using the least-distances rule (Fig. 1). The angles comprising this image will hereafter be denoted as $\{\Phi_e, \Psi_e\}$. Performing these computations for a nonredundant set of proteins yields probability profiles which indicate the likelihood of encountering specific pairs of angles along the elliptical path. A sample distribution (for a randomly chosen amino acid) is visualized in Fig. 2. For each amino acid seven distinct probability peaks can be distinguished; these peaks are assigned letter codes (*A* through *G*). It should be noted that code *C* corresponds to an α-helical structure, code *E* represents a β-sheet while code *G* stands for a left-handed helix. Codes *A*, *B* and *D* all represent poorly ordered structures traditionally referred to as random coils (RC).

The values of $\{\Phi, \Psi\}$ (as they occur in the actual protein) are replaced with $\{\Phi_e, \Psi_e\}$ pairs belonging to the limited conformational subspace. Subsequently each point in the subspace is matched to a local probability peak and assigned a letter code, as shown in Fig. 2 [10].

Contingency table

The procedure described above, when applied to a large number of proteins (nonredundant database), yields a contingency table whose rows (160,000 in all) correspond to tetrapeptide sequences while columns (2401) represent various combinations of structural codes. The tetrapeptide was taken as the basic unit of structure as it is the shortest fragment to which a specific (secondary) structural motif can be unambiguously assigned. The contingency table expresses the correspondence between tetrapeptide sequences and structural motifs occurring in the early stage intermediate. It can be directly exploited in structural simulations of known peptide chains (Table 1). In this study, the contingency table has been created only for the subset of the nonredundant protein structures data base with the protein chains belonging to the testing subset excluded from teaching database. The aim of the modification was to avoid positively biased prediction of protein structure and to separate the elements of teaching set in respect to testing set.

Assignment of structural codes is performed in an overlapping fashion, as shown in Fig. 3. For each input sequence a consensus structure can be defined by taking the most frequently occurring code at a given position in each of the four overlapping structural chains. If no code fulfills this criterion, consensus is based upon the highest probability values in the contingency table.

Statistical analysis

Convergence assessment of both algorithms (*step back* and *step forward*) has been performed by using Chi-square testing as well as analysis of RR (relative risk) [14], OR (odds ratio) [14] and D (distance) [14] parameters. Discrepancies between various structural models can be explained by analyzing the dependencies between correct (or incorrect) simulation results and the involvement of individual residues in interaction with external molecules (e.g., ligands or other proteins). A sample table which expresses these dependencies is shown below (Table 2). External interaction is defined

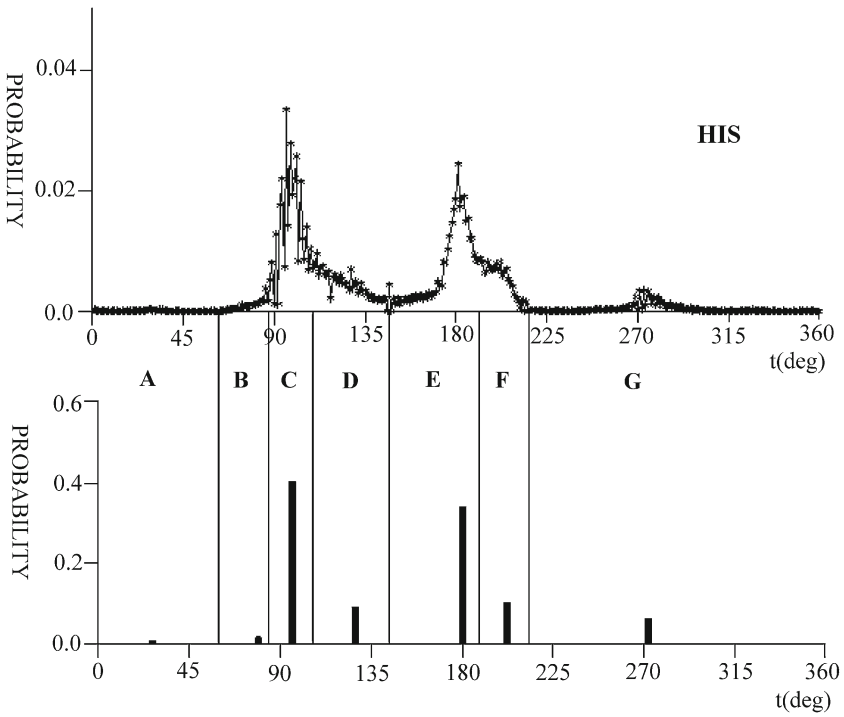

Fig. 2 Sample distribution of probability for a randomly chosen amino acid (histidine). **a** continuous distribution obtained using the *step back* algorithm. **b** discrete distribution obtained using the *step forward* method based on the contingency table. *t*(deg) is the offset (in degrees) along the elliptical path, from an arbitrary starting point in the middle of the lower right-hand corner of the Ramachandran plot

Table 1 Representative fragment of the contingency table generated by the *step back* algorithm

	ABCD	*ADEE*	*CCEG*	…
ADGC	$p_{1,1}$	$p_{1,2}$	$p_{1,3}$	…
ACTW	$p_{2,1}$	$p_{2,2}$	$p_{2,3}$	…
AAMD	$p_{3,1}$	$p_{3,2}$	$p_{3,3}$	…
…	…	…	…	…

Tetrapeptide sequences are listed in boldface while structural codes are italicized

Table 2 A sample grouping data with respect to involvement of residues in external interactions (with ligands and/or other proteins) and the validity of structural code predictions generated using the presented algorithms (e.g., N_{NY} is the number of residues for which structural codes have been incorrectly predicted and which form bonds with external molecules)

		Correct prediction	
		Yes	No
Involvement in external interaction	No	N_{YN}	N_{NN}
	Yes	N_{YY}	N_{NY}

as an engagement of particular residue in ligand (protein, ion, nucleic acid) complexation. This identification is based on *PDBSum* standards (the distance criterion—distance below 4 Å) [15]. A Chi-square test has been applied to assess the dependencies listed above. Values of the Chi-square statistics indicate dependencies ($p<0.05$), which are treated as effects of external interactions upon the conformation of a given amino acid. All relevant calculations were performed using the Statistica package [16].

Results and discussion

Table 3 presents a summarized assessment of the accuracy of *step forward* structural predictions, compared with *step back* simulations. Of note are the large values along the diagonal, which indicate a high ratio of correct predictions (except the position B). Figure 4 contains an equivalent graphical representation of this data. Secondary structural motifs (code C—α-helix; code E—β-structure) are correctly modeled around 55 % of the time. Note the high ratio of correct predictions for codes A, B and D despite their relative scarcity in actual proteins. Code F, traditionally associated with β-like motifs, is also modeled with adequate accuracy (approximately 48 %).

Figure 4a indicates overestimation of code C. The very good result concerns the codes E and F. They represent the β-structural forms quite difficult to be predicted. Very promising is also the code G although classified erroneously as code A. Figure 4b reveals the erroneous recognition of A and G which appear highly entropic (information entropy) sharing a similar likelihood. The positive characteristics concerns the code D, which seems to play an important role as the zone linking the α-helical and β-structural zones on Ramachandran map.

Table 3 reveals overestimation of C-type (α-helical) structures compared to all other structures. Of note is the relative abundance of non-secondary structural motifs. Codes F and G are modeled with high accuracy (24.82 % and 39.87 % respectively). Given the relatively low frequency of A- and D-type motifs, even the obtained prediction values of 4.61 % and 5.97 % (respectively) should be considered satisfactory. The accuracy of prediction for individual amino acids is listed in Tables 3, 4 and Fig. 4. Both diagrams confirm that α-helical motifs are excessively favored with respect to other types of structures.

Characteristics of individual amino acids

To search for the possible specificity of particular amino acid the failure cases were analyzed in respect to each residue individually. High prediction accuracy is noted for ASN in zone *E* and ASP in zone *G*. PHE exhibits affinity for zone *F*, which is an important observation, as, according to research presented in [17], this zone may be associated with amyloidogenesis. One should also note the peculiar properties of CYS, which result from the relatively broad structural variations in this amino acid. While GLY does not appear in zone *G* as frequently as might be expected, HIS seems correctly related to zone *A*. Finally, the results for zone *D* (for all amino acids) appear particularly promising. The importance of this zone for structural modeling has been noted in [18].

```
Sequence:     R P R T A F S S E Q L A R L K R E F N E N R Y L T E R R R Q Q L S S E L G L N E A Q I K I W F Q N K R A K I

Structure 1:  E F E E E F F C C C C C C C C C C C C C C E C E F F C C C C C C C C C C C C G E F C C C C C C C C C C C C - -
Structure 2:  - E F E E F F C C C C C C C C C C C C C C E C C C C C C C C C C C C C C G C C C C E E E E C C C C C C C C C -
Structure 3:  - - C C C C F C C C C C C C C C C C C C C E C E F F C C C C C C C C C C C C G E F C C C C C C C C C C C C C C
Structure 4:  - - - C C C C C C C C C C C C C C C C C C E C E C C C C C C C C C C C C C C G E F C C C C C C C C C C C C - - -

Resulted:     E F C C C F F C C C C C C C C C C C C C C E C E C C C C C C C C C C C C C C G E F C C C C C C C C C C C C C C
```

Fig. 3 Assignment of structural codes to an input sequence of amino acids

Table 3 Frequency of structural code predictions for algorithms based on the contingency table. The table lists the similarities and differences in results obtained using the *step forward* (treated as golden standard) and *step forward* approaches, for the entire protein dataset, grouped by individual structural codes

		Predicted—*step forward*						
		A	B	C	D	E	F	G
Observed—*step back*	A	40.35	61.25	10.79	9.97	6.86	8.52	35.83
	B	3.76	8.44	19.61	4.66	9.18	8.95	2.95
	C	1.35	4.01	19.75	7.53	9.24	6.12	3.52
	D	5.54	15.54	15.61	32.31	13.74	10.37	9.58
	E	1.93	5.63	12.74	9.01	36.22	10.23	3.33
	F	5.04	5.50	11.92	6.83	16.09	49.05	6.86
	G	42.02	0.00	9.57	29.68	8.65	6.76	37.92
Versus *step forward*		100	100	100	100	100	100	100

As can be seen in Fig. 5 and Table 5 that almost all amino acids demonstrate the best predictability for *C* (α-helix) structural code. The second best predicted structural code is *E* (β-structure). However ASP and PHE are the exceptions. The first one represents high predictability for *G* code (left-handed α-helix) and the second one for *F* code (traditionally treated as β-structure although the distinguishing between *E* and *F* structural forms seems to be important).

Individual case studies

In order to ascertain the reasons behind erroneous predictions the authors have performed accuracy analyses for specific proteins. Table 6 lists the best- and worst-case scenarios identified in the course of this study. The distinguishing between helical and differentiated secondary structure is made according to different level of predictability of helical fragments in relation to all other secondary structural form.

The set of well predicted proteins of mainly helical structure represent quite differentiated length of polypeptide chain that suggests the accuracy of prediction not dependent on the size of particular molecule. The proteins of differentiated secondary structure (rather large size with β-structural motifs) appeared surprisingly as predicted correctly quantitatively even higher than helical structures. This observation seems to make the model quite promising.

The lowest predictability was achieved for proteins of low size although of high participation of random coil structures (almost entirely unstructuralized proteins like 3C05 or 2RQW) and proteins of large size of entirely β-structural form.

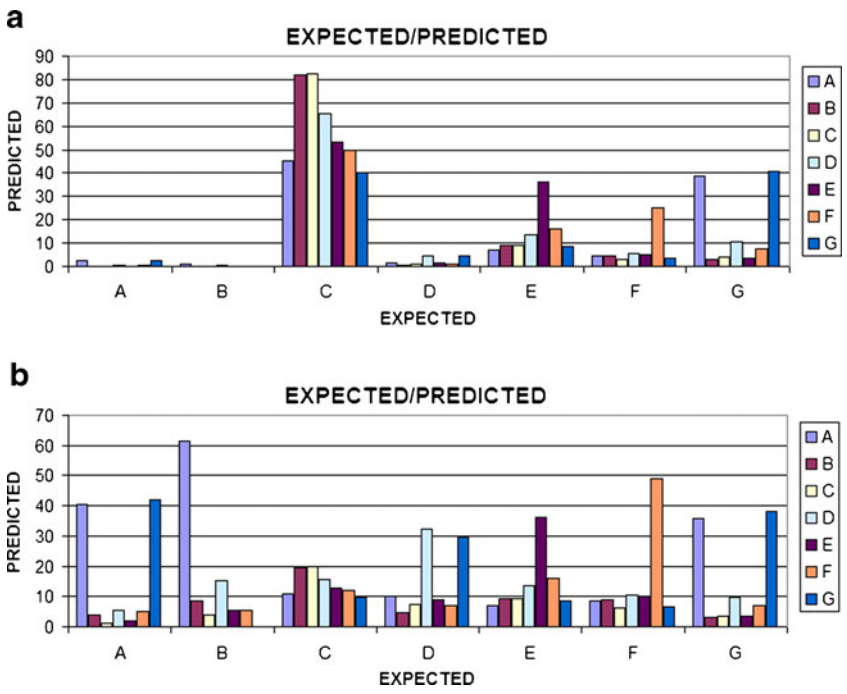

Fig. 4 Comparison of results for the entire testing dataset. The figure lists the aggregate frequency of correct and erroneous predictions for all amino acids. **a** normalized versus *step back* (Table 3). **b** normalized versus the *step forward* (Table 4)

Table 4 Frequency (in percentage value) of correct structural code predictions for algorithms based on the contingency table. The table lists the similarities and differences in results obtained using the *step back* (treated as golden standard) and *step forward* approaches for the entire protein data set, grouped by individual structural codes

		Predicted—*step forward*							Versus *step back*
		A	B	C	D	E	F	G	
Observed—*step back*	A	4.61	0.56	36.36	2.37	12.44	3.64	40.0	100
	B	0.28	1.11	67.10	1.57	13.09	9.68	7.19	100
	C	0.09	0.02	85.85	0.76	8.28	2.97	2.03	100
	D	0.11	0.06	70.33	5.97	14.50	5.81	3.21	100
	E	0.09	0.05	51.87	0.96	40.26	4.78	1.98	100
	F	0.14	0.12	51.16	1.28	19.90	24.82	2.60	100
	G	1.00	0.33	44.15	1.97	9.48	3.19	39.87	100

Helical proteins

The protein 2BA2 (PDB code) is a sample helical protein for which both presented approaches provide consistent predictions. Its native three dimensional structure, as well as the outcomes of *step-forward* and *step-back* simulations, are visualized in Fig. 6.

Proteins with differentiated secondary structure motifs

The protein 1PCZ (PDB code) is representative of a class of proteins which contain α-helical and β-sheet motifs. In spite of this fact, this protein exhibits relatively high prediction accuracy, as illustrated in Fig. 7. Of note is the satisfactory accuracy of β-structure prediction; traditionally a difficult task in ab initio algorithms [19]. Both models (*step back* and *step forward*) appear to correctly identify the location of loops.

Proteins with poor prediction consistency

From among the analyzed proteins, one of the poorest ES prediction consistency was noted for the protein 2RQW (PDB code), with diverse structural characteristics. Visual inspection reveals significant variations between theoretical models for this protein: *step back* predictions differ greatly from *step forward* simulations (Fig. 8). The reason of failure

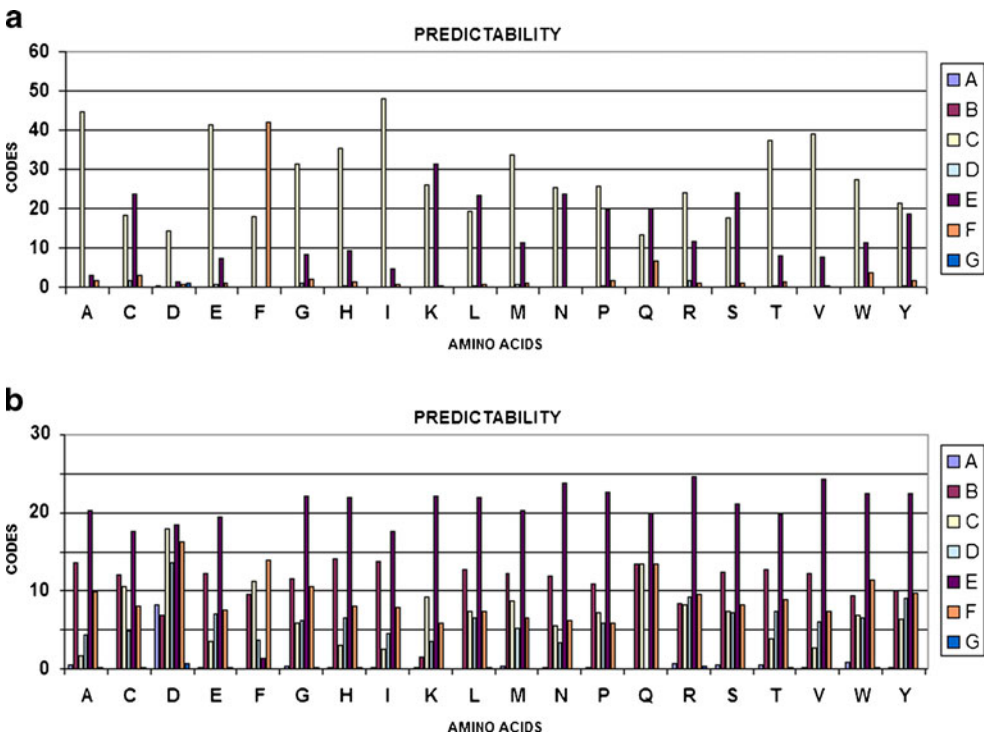

Fig. 5 Amino acids predictability. **a** correct predictions, **b** false predictions

Table 5 Accuracy (expressed in percentage values) of structural code predictions for individual amino acids (NA means that the given amino acid has not been observed in a particular zone)

	Predicted—*step forward*						
Amino acid	A	B	C	D	E	F	G
Ala	0.0	2.20	96.04	1.42	13.38	14.37	0.0
Cys	25.0	0.0	71.37	16.48	65.07	30.0	21.87
ASP	5.87	3.60	52.68	2.96	16.35	5.97	64.34
Glu	0.0	0.0	93.22	6.67	30.32	13.79	8.86
Phe	0.0	2.67	63.04	0.55	2.04	74.54	3.03
Gly	0.0	5.0	85.60	15.84	30.66	16.79	11.70
His	14.28	2.08	92.03	7.43	27.98	13.99	6.12
Ile	0.0	1.35	95.69	2.49	21.01	10.99	1.94
Lys	0.0	0.0	78.04	1.06	61.75	5.62	0.0
Leu	0.0	0.0	79.80	5.69	50.84	10.57	8.70
Met	0.0	7.69	86.44	0.92	46.79	19.85	10.34
Asn	0.0	3.33	83.76	0.65	53.19	2.68	0.0
Pro	NA	0.0	80.19	1.37	59.75	30.93	27.78
Gln	NA	0.0	78.35	13.19	52.68	21.60	18.64
Arg	NA	1.47	79.71	16.23	37.86	17.05	24.91
Ser	NA	0.0	79.43	11.90	51.54	9.50	2.22
Thr	NA	0.0	93.52	4.56	31.53	12.63	4.10
Val	NA	0.0	94.91	2.21	26.74	8.95	0.0
Trp	NA	1.33	85.13	2.24	32.98	22.40	1.22
Tyr	NA	2.0	77.55	5.07	48.58	18.88	3.03

lies in the algorithm by which Φ and Ψ angles are projected onto the conformational subspace: the *step back* model seeks the nearest point along the elliptical path, while the *step forward* approach adopts the coordinates of the relevant probability maximum corresponding to a given structural code. Thus, while the *step back* algorithm can be characterized as continuous, the *step forward* model is inherently discrete (of course, additional differences may result from the somewhat arbitrary assignment of structural motifs to sequence fragments, as previously discussed).

Seeking the reasons behind the observed differences

In order to explain the reasons behind the mismatched predictions provided by both theoretical models the authors have focused on the involvement of amino acids in external interactions (other than short-range interactions with immediate neighbors and steric effects). Since the presence of external molecules may affect the resulting conformation of the polypeptide chain, it is worthwhile to assess the link between prediction accuracy and the involvement of specific

Table 6 Best- and worst-case scenarios from the point of view of structural accordance between *step back* and *step forward* predictions. The best results are further subdivided into entirely helical and non-helical structures

Best accordance						Worst accordance		
Helical structures			α + β + RC structures					
PDB ID	AA	%	PDB ID	AA	%	PDB ID	AA	%
2BA2	128	81.08	2VBL	152	68.24	3C05	59	22.73
1FPO	171	77.91	1S6M	288	67.54	2RQW	24	25.27
1ZHC	320	75.36	2O5E	634	63.55	1I16	130	25.98
2X04	77	70.27	1PCZ	191	62.68	1TQ5	234	28.14
3LJW	118	68.47	1CR9	219	60.26	3O0X	347	29.87

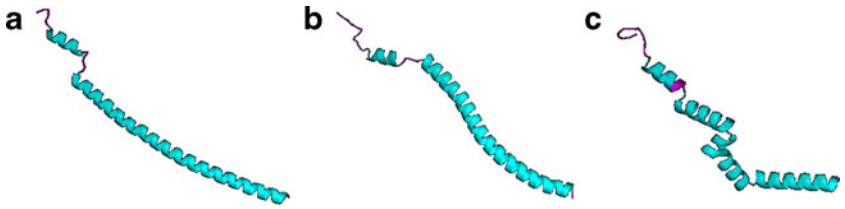

Fig. 6 Three dimensional structures of the protein 2BA2 (PDB code) which exhibits the highest structural prediction consistency among purely helical proteins. **a** the native structure obtained from the PDB, **b** the *step back* model, **c** the *step forward* model. Fragments forming α-helixes and loops in native structure of the protein (determined using the DSSP algorithm) are marked in cyan and magenta respectively in all three images.

residues in interactions with ligands, ions or other proteins. The following tables illustrate prediction accuracy as a function of such involvement.

Applying chi-square criteria to the values listed in Tables 7 and 8 reveals a causal link between the presence (or absence) of external ligands and the accuracy of theoretical predictions. Although the relation between the status of residue (engagement in any external interactions) and accuracy of its prediction appears to be significant, the engagement in ion binding or nucleic acid complexation was revealed as the opposite case. The residues engaged in ligand binding appear to represent the strong dependency between their status and accuracy of prediction taking into consideration all parameters measuring the dependency between effects of analysis and engagement in external interaction. The dependence was found for residues engaged in protein-protein interaction. However the OR and RR analysis suggests also (besides ligand binding) the ion complexation as influencing the status of particular residue in respect to presented analysis. The surprising result is lack of correlation (and influence) of nucleic acid complexation and the structural predictability of residue in respect to presented method.

Steric clashes

It should be noted that both *step back* and *step forward* prediction results involve steric clashes (i.e., the chains loop back upon themselves or are packed too tightly). A special algorithm has been devised to resolve such problems by adjusting the values of Ψ_e and Φ_e angles within the limits imposed by the partitioning of the elliptical path into structural zones. Adjustments are performed in a hierarchical fashion, starting with zones *A*, *B* and *D*, then proceeding to zones *F* and *G*. Zones *C* and *E* are not affected. The convergence criterion is that no two atoms in the molecule may be brought closer than within 4 Å of each other.

Conclusions and Discussion

The goal of our analysis was to pinpoint the greatest problems associated with protein structure prediction. The model presented in this work is assumed to avoid as much as possible the random search for initial structural forms for further energy optimization procedures. Model avoids also the technique based on pasting the short polypeptides fragments preliminarily recognized as preferable for particular fragment. The technique defining the "consensus" sequence of structural codes in the overlapping system introduces the smoothing of structural elements without analogy to particular examples identified in proteins available in crystal forms. The accordance level received for presented technique seems to be satisfactory assuming that the detailed definition of the final structure is the result of late stage folding process which is able to introduce the local corrections to the structure defined in the early step. The elimination of clashes (as they appear in

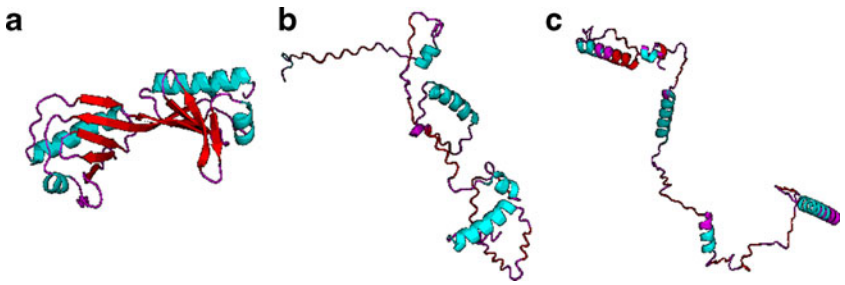

Fig. 7 Three dimensional structures of protein 1PCZ (PDB code) which exhibits the highest structural prediction consistency from among proteins not dominated by helical motifs. **a** the native structure obtained from the PDB, **b** the *step back* model, **c** the *step forward* model. Fragments forming α-helixes, β-sheets and loops in native structure of the protein (determined using the DSSP algorithm) are marked in *cyan*, *red* and *magenta* respectively in all three images

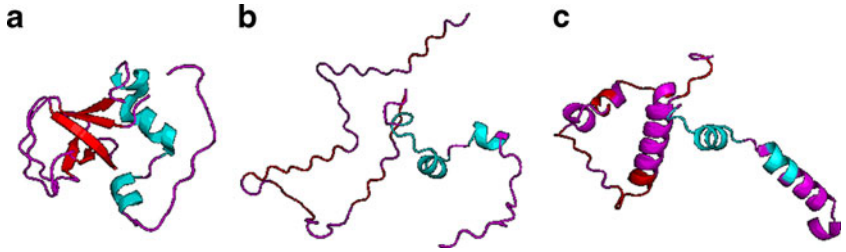

Fig. 8 Three dimensional structures of the protein 2RQW (PDB code), for which the *step forward* predictions were the most inaccurate in the study group. **a** the native structure obtained from the PDB, **b** the *step back* model, **c** the "step forward" model. Fragments forming α-helixes, β-sheets and loops in native structure of the protein (determined using the DSSP algorithms) are marked in *cyan*, *red* and *magenta* respectively in all three images

the structures generated according to presented model) although keeping the structural code is aimed on the verification of the model particularly in respect to defined structural zones. Introduction of random search for Φ and Ψ angles (outside the ellipse path) could deprive the model of its heuristic character.

The main idea of the model is to be able to trace the folding process in the sense of monitoring the steps introducing the mismatch to be able to recognize its source and the conditions. The main question concerning the protein structure prediction is not "How to predict the correct structure" but rather "Why do they fold the way they do". It is expected that our model at least attempts to find the answer to this question.

Structural forms of early step folding is not available although the description of some rare cases can be found in literature [2]. Search for mechanistic model mimicking the folding process seems to be required. The introduction of limited conformational sub-space postulated earlier [12]. Model of ellipse path (constructed on the basis of backbone geometry) limits the size of conformational space to the extent of balancing the amount of information carried by amino acid sequence with the amount of information sufficient to predict the structure of early stage intermediate. It was shown that such balance is achieved for the presented model [8, 11]. The accordance received for presented model (a little bit below 50 %) is in the range of expectations. The high specificity of biological function requires highly differentiated structural motifs. The classification used for the construction of contingency table does not take into account the status of particular residue in respect to its participation in biological activity. The ellipse path was generated assuming the relaxed form of backbone. The balance between specificity and general model may be of the range comparable to the level of predictability of presented model. The construction of the late steps is assumed to be in strong relation to the construction of specific structural motifs related to function. The "late stage model" seems to collaborate well with the "early stage model" presented in this paper [13, 20]. Results shown in Tables 6 and 7 may suggest the important role of external factors which limit the conformational freedom of the backbone, what was assumed in the model presented. The construction of contingency table for residues not engaged in any external interaction is planed. The influence of external factors (ligands, ions, nucleic acids, and protein

Table 7 Summarized view of the effect of external interactions on incorrect predictions (total number of residues analyzed: 56,836). Each cell lists (respectively) the number of residues involved in protein complexation (P-P), ligand complexation (L), ion complexation (I), nucleic acid complexation (NA) and any form of complexation (ALL)

			Prediction	
			Correct	Failure
Involvement in external interaction	No	P-P	21547	25374
		L	23034	27115
		I	23607	27946
		NA	23592	27974
		ALL	19492	22898
	Yes	P-P	2130	2679
		L	642	938
		I	69	107
		NA	84	79
		ALL	4184	5155

Table 8 Results of statistical analysis showing the values of chi-square test, OR, RR and D. Additionally to the values of particular parameter the [1]95 % OR confidence interval and [2]95 % RR confidence interval is given. The protein-protein interaction (P-P), ligand binding (L), ion complexation (I) and nucleic acid complexation (NA) were taken into consideration. ALL represents engagement in any form of external complexation

	Statistical analysis			
External interaction	Chi-square	OR[1]	RR[2]	D
P-P	0.031	1.07 (1.01–1.13)	1.04 (1.00–1.07)	0.02
L	0.000	1.24 (1.12–1.38)	1.13 (1.07–1.20)	0.05
I	0.080	1.31 (0.96–1.74)	1.17 (0.98–1.43)	0.07
NA	0.140	0.79 (0.58–1.09)	0.89 (0.77–1.05)	0.38
ALL	0.038	1.05(1.00–1.10)	1.03 (1.00–1.05)	0.01

complexed) was the object of analysis presented in detail in [21–23].

So far the presented model was used to predict ES intermediates for the following proteins: lysozyme [6], ribonuclease [8], hemoglobin [7] and BPTI [9]. Early stage structures (generated with the use of structural codes) were fed into late stage (LS) model calculations. Results suggest the validity of the resulting structures; thus the accuracy of the presented ES model is difficult to be estimated due to very low number of structures generated experimentally [2]. The validity of ES results is possible after performing the simulation of folding on the basis of LS model [20]. The structure created by LS model produces the structure easily comparable with the crystal structures deposited in PDB [15].

In spite of the above, comparative analysis of ES structures in the serpin family suggests that such proteins do indeed possess structural components facilitating biological function [24]. An important advantage of the proposed model is that it introduces a clear division of RC secondary structural motifs into several categories referenced by letters *A*, *B*, *D* and *F*. This distinction enables better differentiation of secondary structural characteristics. Another positive aspect of our approach is its relatively high accuracy in modeling such structures.

The early stage intermediate generation algorithm discussed here is a direct counterpart of fold recognition methods as defined in the CASP4 nomenclature [19]. In addition to predicting secondary structural characteristics for specific fragments of the polypeptide chain, it also models loops (fragments connecting different structural motifs) by distinguishing zones *A*, B, *D*, *F* and *G*. A distinct advantage of the presented approach is that it enables clear identification of the reasons behind incorrect predictions, whereas for Monte Carlo methods such analysis can only be performed at the final stage of simulation and does not enable the researcher to identify the origin of errors. In the context of our algorithm, erroneous predictions in the LS intermediate are found to correspond to the presence of external ligands which distort the structure of the polypeptide chain [21–23]. Such interactions are not to be confused with natural folding preferences of the polypeptide backbone (Φ and Ψ angles). The work also presents the impact of external interactions on the folding process for specific types of amino acids.

The early stage intermediate appears in literature suggesting the increased interest in step-wise folding process [25–33]. The one step folding model assuming the generation of 3D structure for known amino acid sequence solely was examined as impossible on the basis of information theory [34]. The analysis of early-stage geometrical motifs present in crystal structure suggest the reliability of the presented model despite relatively low however satisfactory level of the structural predictability [35–38].

Acknowledgments The research presented in this paper was partly funded by Collegium Medicum grants UJ K/ZDS/001531. We wish to thank Piotr Nowakowski for editorial work and Anna Śmietańska for technical support.

References

1. Creighton TE (1990) Protein folding. Biochem J 270:1–16
2. Religa TL, Markson JS, Mayor U, Freund SM, Fersht AR (2005) Solution structure of a protein denatured state and folding intermediate. Nature 437:1053–1056
3. Bystroff C, Shao Y (2004) Modeling protein folding pathways. In: Bujnicki JM (ed) Practical bioinformatics. Springer, Heidelberg, pp 97–122
4. Rohl CA, Strauss CE, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93
5. Roterman I (1995) Modelling the optimal simulation path in the peptide chain folding—studies based on geometry of alanine heptapeptide. J Theor Biol 177:283–288
6. Jurkowski W, Brylinski M, Konieczny L, Roterman I (2004) Lysozyme folded in silico according to the limited conformational sub-space. J Biomol Struct Dyn 22(2):149–58
7. Brylinski M, Jurkowski W, Konieczny L, Roterman I (2004) Limitation of conformational space for proteins—early stage folding simulation of human α and β hemoglobin chains. TAS Conformational K Quarterly 8:413–422
8. Jurkowski W, Brylinski M, Konieczny L, Wiśniowski Z, Roterman I (2004) Subspace in simulation of early-stage protein folding. Proteins 55(1):115–127
9. Bryliński M, Jurkowski W, Konieczny L, Roterman I (2004) Limited conformational space for early-stage protein folding simulation. Bioinformatics 20(2):199–205
10. Bryliński M, Konieczny L, Czerwonko P, Jurkowski W, Roterman I (2005) Early-stage holding In proteins (in silico)—sequence-to-structure relation. J Biomed Biotechnol 2:65–79
11. Jurkowski W, Baster Z, Dułak D, Roterman I (2012) The Elary stage intermediate. In: Roterman-Konieczna I (ed) Protein folding in Silico. Woodhead, Cambridge, pp 1–20
12. Alonso DO, Daggett V (1998) Molecular dynamics simulations of hydrophobic collapse of ubiquitin. Protein Sci 7:860–874
13. Banach M, Konieczny L, Roterman I (2012) Use of the "fuzzy oil drop" model to identify the complexation area in protein homodimers. In: Roterman-Konieczna I (ed) Protein folding in Silico. Woodhead, Cambridge, pp 95–122
14. Agresti A (2007) Contingency tables. In: Introduction to categorical data analysis, 2nd edn. Wiley, New York, pp 21–64
15. http://www.ebi.ac.uk/pdbsum. Accessed 15 Dec 2012
16. StatSoft, Inc. (2011) STATISTICA (data analysis software system), version 10. www.statsoft.com
17. Król M, Konieczny L, Stąpor K, Wiśniowski Z, Ziajka W, Szoniec G, Roterman I (2012) Misfolded proteins. In: Roterman-Konieczna I (ed) Protein folding in Silico. Woodhead, Cambridge, pp 141–164
18. Zhou AQ, O'Hern CS, Regan L (2011) Revisiting the Ramachandran plot from a new angle. Prot Sci 20:1166–1171
19. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I (1999) Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins Suppl 3:149–70
20. Konieczny L, Bryliński M, Roterman I (2006) Gauss-function-based model of hydrophobicity density In proteins. In Silico Biol 6:5–22

21. Banach M, Konieczny L, Roterman I (2012) Can the structure of hydrophobic core determine the complexation site? In: Roterman-Konieczna I (ed) Identification of ligand binding site and protein-protein interaction area. Springer, Heidelberg, pp 41–54
22. Alejster P, Banach M, Jurkowski W, Marchewka D, Roterman I (2013) Comparative analysis of techniques oriented on the recognition of ligand binding area in proteins. In: Roterman-Konieczna I (ed) Identification of ligand binding site and protein-protein interaction area. Springer, Heidelberg, pp 55–86
23. Marchewka D, Jurkowski W, Banach M, Roterman I (2013) Prediction. In: Roterman-Konieczna I (ed) Identification of ligand binding site and protein-protein interaction area. Springer, Heidelberg, pp 105–134
24. Bryliński M, Konieczny L, Kononowicz A, Roterman I (2008) Conservative secondary structure motifs already prezent In early-stage folding (in silico) as fund In the serpine family. J Theor Biol 251:275–285
25. Feng H, Zhou Z, Bai Y (2005) A protein folding pathway with multiple folding intermediates at atomic resolution. Proc Natl Acad Sci USA 102:5026–5031
26. Kuwajima K, Schmid FX (1984) Experimental studies of folding kinetics and structural dynamics of small proteins. Adv Biophys 18:43–74
27. Fischer B (1996) Folding of lysozyme. EXS 75:143–61
28. Eyles SJ, Radford SE, Robinson CV, Dobson CM (1994) Kinetic consequences of the removal of a disulfide bridge on the folding of hen lysozyme. Biochemistry 33(44):13038–48
29. Krishna MM, Englander SW (2007) A unified mechanism for protein folding: predetermined pathways with optional errors. Protein Sci 16(3):449–64
30. Bédard S, Krishna MM, Mayne L, Englander SW (2008) Protein folding: independent unrelated pathways or predetermined pathway with optional errors. Proc Natl Acad Sci U S A 105(20):7182–7187
31. Su ZD, Arooz MT, Chen HM, Gross CJ, Tsong TY (1996) Least activation path for protein folding: investigation of staphylococcal nuclease folding by stopped-flow circular dichroism. Proc Natl Acad Sci U S A 93(6):2539–44
32. Grantcharova VP, Baker D (1997) Folding dynamics of the src SH3 domain. Biochemistry 36(50):15685–92
33. Gardner BM, Walter P (2011) Unfolded proteins are Ire1-activating ligands that directly induce the unfolded protein response. Science 333(6051):1891–1894
34. Alejster P, Jurkowski W (2011) Roterman I (2012) Structural information involved In the interpretation of the step-wise protein folding process. In: Roterman-Konieczna I (ed) Protein folding in Silico. Woodhead, Cambridge, pp 39–54
35. Roterman I, Konieczny L, Jurkowski W, Prymula K, Banach M (2011) Two-intermediate model to characterize the structure of fast-folding proteins. J Theor Biol 283:60–70
36. Roterman I, Konieczny L, Banach M, Jurkowski W (2011) Intermediates in the protein folding process: a computational model. Int J Mol Sci 12:4850–4860
37. Banach M, Prymula K, Jurkowski W, Konieczny L, Roterman I (2012) Fuzzy oil drop model to interpret the structure of antifreeze proteins and their mutants. J Mol Model 18:229–237
38. Jurkowski W, Kułaga T, Roterman I (2011) Geometric parameters defining the structure of proteins-relation to early-stage folding step. J Biomol Struct Dyn 29(1):79–104