# LIMITATION OF CONFORMATIONAL SPACE FOR PROTEINS – EARLY STAGE FOLDING SIMULATION OF HUMAN $\alpha$ AND $\beta$ HEMOGLOBIN CHAINS

MICHAŁ BRYLIŃSKI[1,2], WIKTOR JURKOWSKI[1,2],
LESZEK KONIECZNY[3] AND IRENA ROTERMAN[1]

[1] *Department of Bioinformatics and Telemedicine,*
*Collegium Medicum, Jagiellonian University,*
*Kopernika 17, 31-501 Cracow, Poland*
*myroterm@cyf-kr.edu.pl*

[2] *Institute of Chemistry, Jagiellonian University,*
*Ingardena 3, 30-060 Cracow, Poland*

[3] *Institute of Medical Biochemistry,*
*Collegium Medicum, Jagiellonian University,*
*Kopernika 7, 31-034 Cracow, Poland*

(Received 31 March 2004; revised manuscript received 15 April 2004)

**Abstract:** The starting structure of *ab initio* protein structure prediction methods is problematic as the energy minimization procedure stops searching for an optimal structure of the function's local minimum. The method presented in the paper helps to find the starting structure. Although it is based on the known native protein structure, it seems to deliver a key to the formation of a common universal starting structure. The limited conformational sub-space, defined on the basis of a geometrical model of the polypeptide backbone with the side chain-side chain interaction excluded, seems to deliver the original structure of the polypeptide, which is modified step by step as the role of the side chain interactions increases during the energy minimization procedure. Here, the method is applied to human hemoglobin chains $\alpha$ and $\beta$ to test the applicability of the method to proteins with a high content of helical forms and lacking disulphide bonds.

**Keywords:** early-stage folding, structure prediction, conformational space

## 1. Introduction

The problem of early-stage protein folding has been widely investigated by experimental biologists as well as those engaged in protein folding simulations. *Ab initio* methods require a definition of the starting conformation for the energy minimization procedure. For a polypeptide 27 amino acids long, the number of possible starting conformations has been estimated to be about $10^{16}$ [1]. The starting structures discussed by Dobson seem to have been arbitrarily selected. The possible pathway (or pathways) leading from random unfolded structures to the native structure, as has been suggested, may be traced according to the number of non-bonding contacts [1–3].

A mechanism has been proposed to limit the conformational search to particular regions of the conformational space or to narrow the pathways between unfolded and native states [4]. A number of simplified models have been proposed to solve this problem based on a simplified scheme of amino acid representation [5, 6] and limitation of the conformational space to four low-energy basins [7–9].

The approach presented in this paper introduces a model for a common definition of protein structures which may be treated as a possible *in silico* early-stage form of the polypeptide chain [10–12].

The early-stage forms (as may be assumed for folding *in silico*) have appeared as a result of a simplified presentation of structures based on two geometric parameters: $V$ – the angle which expresses the dihedral angle between two sequential peptide bond planes and $R$ – the radius of curvature, which appears to depend on angle $V$ in a parabolic relation (for $R$ on a log scale) [10, 11]. The structures selected from the whole Ramachandran map according to the model parabolic relation ($\ln R$ versus $V$, $0^\circ$ for a helical form with low $R$, to $180^\circ$ for a $\beta$-structural form with very large $R$) have revealed an area on the map which appeared to be ellipse-shaped. This elliptic path fragment of the Ramachandran map is assumed to represent the early-stage polypeptide structure. The structures obtained according to this criterion are based on the backbone conformation but excluding the side chain-side chain interaction, since only the mutual orientation of the peptide bond planes have been considered. This assumption is in accordance with suggestions that the backbone structure dominates in early-stage polypeptide structure formation and that the side chain–side chain interaction occurs at the later stages of polypeptide folding [13].

The elliptic path links all structurally significant areas (the right-handed helix, the C7eq energy minimum and the left-handed helix) and simultaneously reveals the simplest path for structural changes allowing the $\alpha$-to-$\beta$-structural transformation [11].

At the same time, a quantitative analysis of the information stored in the amino acid sequence in relation to the amount of information necessary to predict particular $\Phi$, $\Psi$ angles as they appear in real proteins, has revealed that these two quantities become equilibrated in the elliptic path approach [12].

Hemoglobin has been selected in this paper as an example to prove the correctness of the assumed model. The $\Phi$, $\Psi$ angles as they appear in the crystal structure of both $\alpha$ and $\beta$ chains have been moved on the Ramachandran map toward the elliptic path according to the shortest distance criterion. The energy minimization procedure applied to the structural forms so obtained has been aimed at estimating the degree of approach to native-like structures.

## 2. Materials and methods

### 2.1. Creating an elliptic path-derived structure

The $\alpha$ and $\beta$ chains of human hemoglobin (3HHB – PDB identification) were analyzed to check the usefulness of the model delivering the starting point (polypeptide conformation) for the energy minimization procedure.

The $\Phi$, $\Psi$ angles were calculated for each amino acid in the polypeptide chains as they appear in the native form of this protein. The $\Phi_{ell}$ and $\Psi_{ell}$ angles belonging to the elliptic path were found according to the shortest distance versus the $\Phi$, $\Psi$ angles observed in the native structure (Figure 1).

The ellipse-derived structure for the $\alpha$ and $\beta$ chains of hemoglobin was created using the ECEPP/3 program. The $\Omega$ dihedral angles were taken as $180^{\circ}$ for all amino acids. The side chain structures were created according to ECEPP/3 standards.

### 2.2. The energy minimization procedure

The energy minimization procedure was performed using the ECEPP/3 program for the ellipse-derived structure of both chains of human hemoglobin [14]. The $\Phi$, $\Psi$ angles were calculated for post-minimization structures. An unconstrained minimization solver with analytical gradient [15] was used. The values of absolute and relative function convergence tolerances were set at $1\cdot 10^{-3}$ and $1\cdot 10^{-5}$, respectively. The energy minimization procedure was carried out both with and without properly defined disulfide bonds. The coordinates and values of the backbone dihedral angles were saved for analysis at 10-step intervals. The energy minimization procedure was done for each chain on an SGI Origin 2000 at the TASK computing center in Gdansk.

### 2.3. Structural comparison

Different criteria were applied to compare the analyzed structures (native, elliptical, post-energy minimization):

1. The distances between the geometric center of the molecule and the sequential $C\alpha$ atoms (called $D_{\text{center}-C\alpha}$ in this paper) in the polypeptide chain were calculated. This plot revealed fragments completely different from the native structural forms and those formed in a similar way (although oriented in space in similar or different ways). The polypeptide fragments distinguished according to the profile were also characterized using an RMS-D calculation. RMS-D values were calculated for selected fragments after overlapping the distinguished fragments taken from the native and the ellipse-derived structural forms. RMS-D values were calculated as averages for particular polypeptide fragments.
2. A cut-off distance of 12Å was used to estimate the number of non-bonding interactions in all structural forms under consideration.
3. Since the ellipse-derived structure is assumed to represent the early stage of folding, the change of molecular size is traced during the approach to the native form of the polypeptide. The size is expressed as the size of a box large enough to contain the whole molecule.

The longest $C\alpha$-$C\alpha$ distance was taken as the $D_Z$ measure (distance along the $Z$-axis), the longest $C\alpha$-$C\alpha$ distance in the $XY$-plane was taken as the $D_Y$ measure

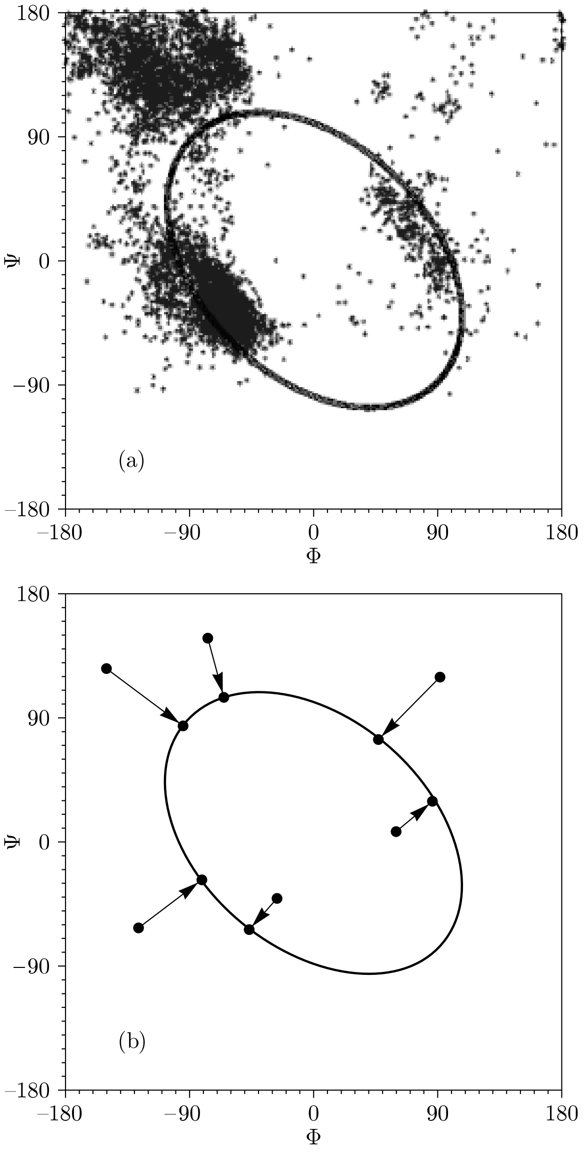

**Figure 1.** The ellipse-path-limited conformational sub-space: (a) relation to Φ, Ψ angle distribution as it appears in real proteins; (b) the shortest distance between particular Φ, Ψ angles and the point on the ellipse represents the way in which the Φ, Ψ angles can be found on the ellipse

(distance along the $Y$-axis), and the difference between the highest and lowest values of $X$ was taken as the measure of the $D_X$ box edge.

## 3. Results

### 3.1. *Φ, Ψ dihedral angle changes*

The Φ, Ψ angle distribution exposing the range of dihedral angle change for all the discussed structural forms is presented in Figure 2a and Figure 2d for the native

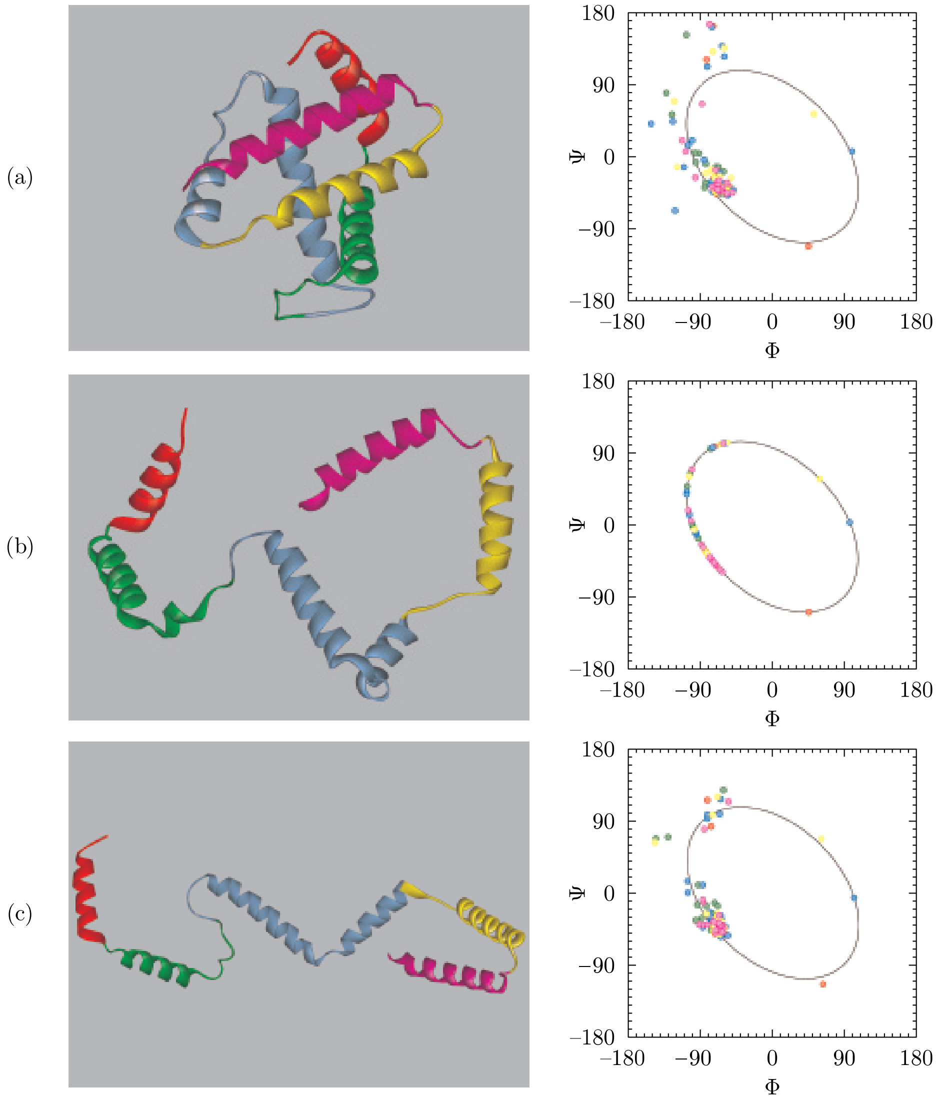

**Figure 2.** Structures of $\alpha$ and $\beta$ hemoglobin chains and their $\Phi$, $\Psi$ angle distributions versus the elliptic path. $\alpha$ hemoglobin chain: (a) the native form, (b) the ellipse-based structure, (c) the post-energy-minimization structural form

forms of $\alpha$ and $\beta$ chains of hemoglobin, respectively, in Figures 2b and 2e for the ellipse-derived structures, and in Figures 2c and 2f for the post-energy-minimization structural forms. A similarity of $\Phi$, $\Psi$ angle distributions can be seen between the native and post-energy minimization structures. The chain fragments are color-coded according to the results of analysis presented in Figure 3.

### *3.2. Spatial distribution of $C\alpha$ atoms versus the geometrical center ($D_{\mathbf{center}-C\alpha}$)*

The profile of the size of vectors linking the geometrical center with the sequential $C\alpha$ atoms offers insight into the three-dimensional relative displacements versus

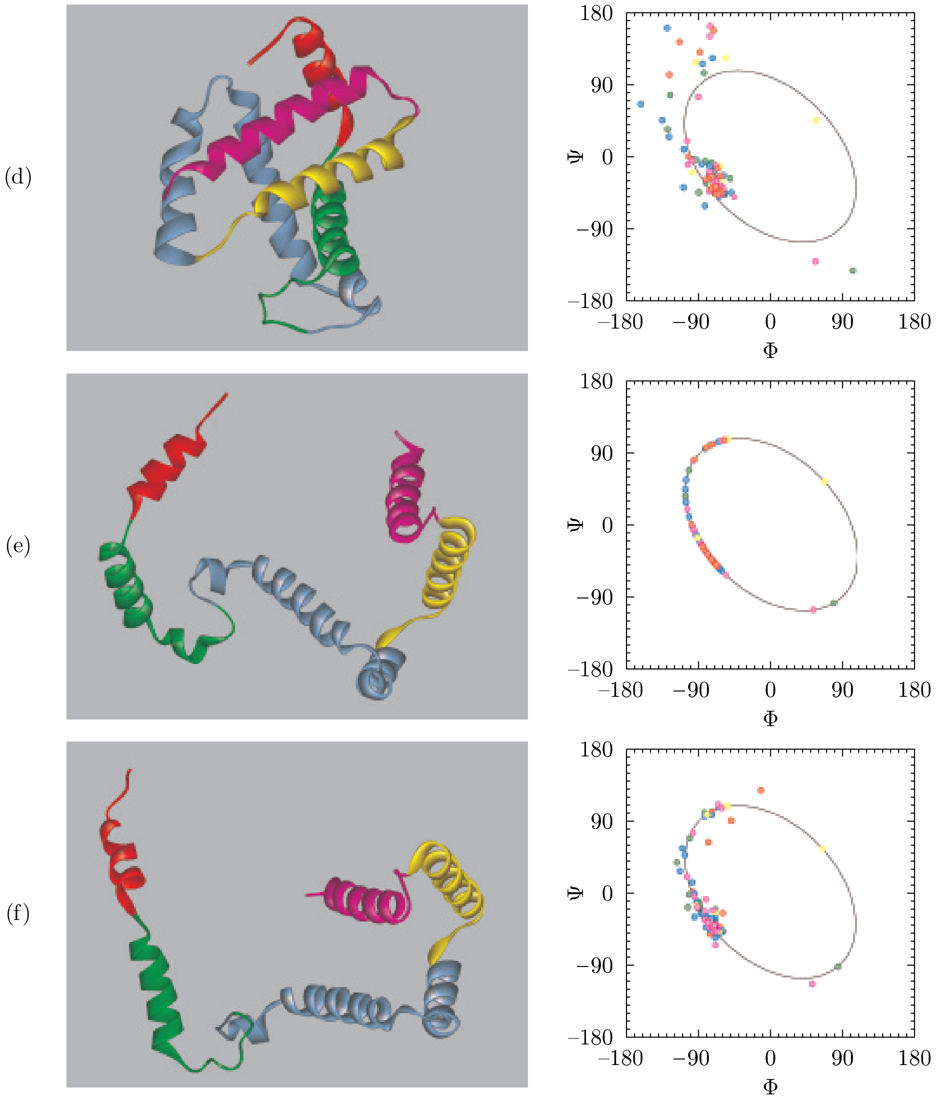

**Figure 2 – continued.** Structures of $\alpha$ and $\beta$ chains of hemoglobin and their $\Phi$, $\Psi$ angle distributions versus the ellipse path. $\beta$ hemoglobin chain: (d) the native form, (e) the ellipse-based structure, (f) the post-energy-minimization structural form

the native-form protein [16]. Structural similarity may be disclosed by overlapping of the lines representing the two compared structures, while a parallel orientation of profiles represents a similarity of structural forms in the compared molecules differently oriented in space. An increase of vector length in respect to the native structure is obviously due to an extension of the structure that is always associated with the transformation from a native to an ellipse-path-delimited structure. The profiles for all the structural forms of $\alpha$ and $\beta$ hemoglobin chains discussed in this paper are presented in Figure 3.

Five fragments can be distinguished in the $\alpha$ chain (Figures 3a and 3b). The central one, containing amino acids 47–95 aa, characterized by a very similar regularity

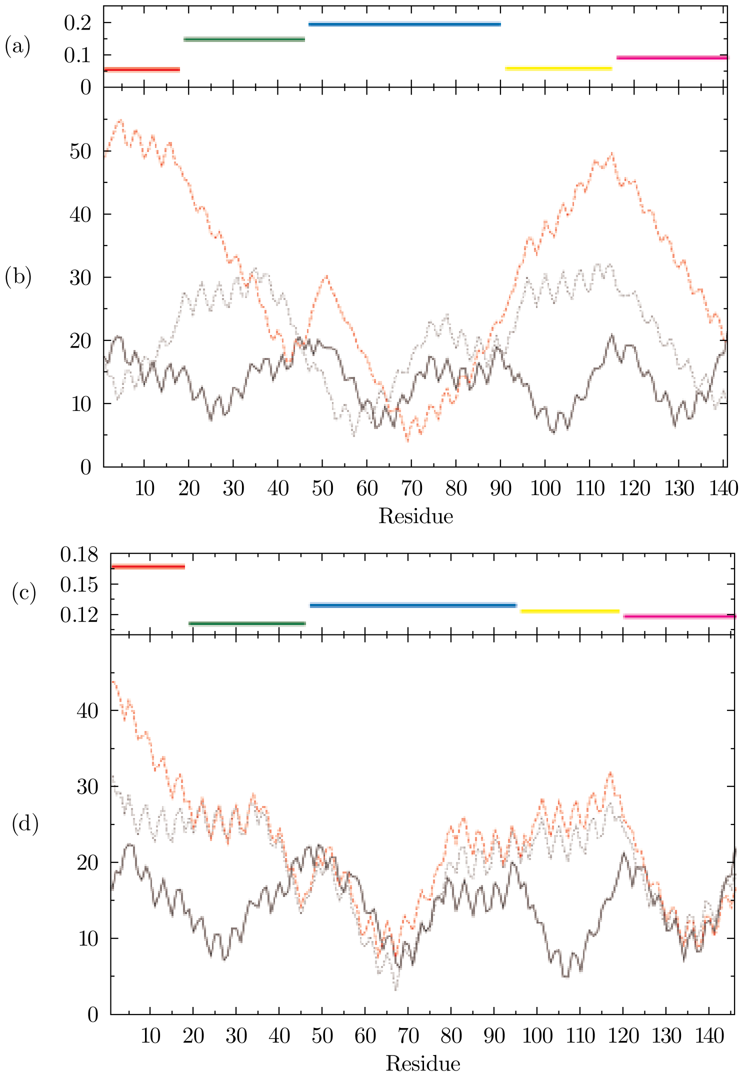

**Figure 3.** Comparison of structural forms of: (a)–(b) $\alpha$ and (c)–(d) $\beta$ hemoglobin chains. (a), (c) RMS-D [Å] (per residue) calculated for structurally differentiated polypeptide fragments; the fragments were defined according to the profile presented in (b), (d); parallel curve fragments represent a correct spatial orientation of the polypeptide, while dissimilar curve regularities represent fragments with little similarity of spatial orientation of the particular polypeptide fragment; (b), (d) profile representing the distribution of distances linking the geometrical center of the molecule with the sequential $C\alpha$ atoms; continuous line – the native form, dotted line – the ellipse-derived structure, dashed line – the post-energy-minimization structure; the color notation of all the graphic presentations in this paper is in accordance with the fragments distinguished in this figure

of $D_{\text{center}-C\alpha}$ in the native and post-energy-minimization structures, appeared to represent the highest RMS-D of about 0.2 Å. RMS-D values are averages calculated for each distinguished fragment. In the $C$-terminal fragment (116–141 aa) the profiles are represented by parallel lines, which suggests that the structural forms are similar, while the spatial orientations differ. The $N$-terminal part (1–18 aa) seems to represent conformations of a quite similar post-energy-minimization structure, with opposite orientations versus the geometrical center. These overlapped fragments show the lowest RMS-D values.

Five fragments can also be distinguished in the $\beta$ chain (Figures 3c and 3d), but their RMS-D values are significantly lower for the $\beta$ than those for the $\alpha$ chain. The lowest RMS-D was obtained for the 19–46 aa fragment and the $C$-terminal fragment (120–146 aa), the $D_{\mathrm{center}-C\alpha}$ profiles of which overlap for all the analyzed structural forms. The central fragment (47–90 aa), which is characterized by great similarity of $D_{\mathrm{center}-C\alpha}$ profiles, reveals a higher RMS-D, although a value close to 1.2Å seems to express quite good accordance.

### 3.3. Visual analysis

Figure 2 visualizes the structural changes in the $\alpha$ and $\beta$ chains of hemoglobin under different conditions. A color notation differentiates the particular polypeptide fragments and distinguishes them according to their similarity as measured by the $D_{\mathrm{center}-C\alpha}$ vector profiles. The same color notation is used for the $\Phi$, $\Psi$ angle distributions and RMS-D fragments.

### 3.4. Non-bonding interactions

The native non-bonding interactions present in all the discussed structural forms are shown in Figure 4. The percentage of native non-bonding interactions present in an ellipse-derived structure is 47.88% in the $\alpha$ chain and 49.74% in the $\beta$ chain. The post-energy minimization structures represent the following native non-bonding interactions: 48.28% for the $\alpha$ chain and 48.49% for the $\beta$ chain.

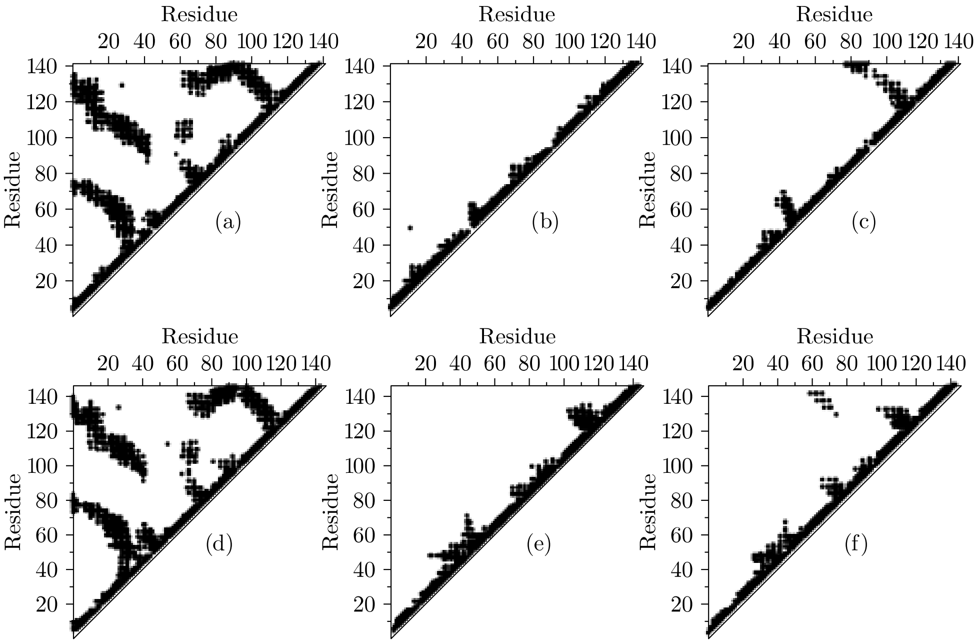

**Figure 4.** Non-bonding contacts; in the $\alpha$ hemoglobin chain: (a) the native form, (b) the ellipse-derived structure, (c) the post-energy-minimization; in the $\beta$ hemoglobin chain: (d) the native form, (e) the ellipse-derived structure, (f) the post-energy-minimization

The numbers of non-bonding contacts are calculated for a 12Å cut-off distance. This quantity obviously depends on the molecule under study [17] and is strongly related to the percentage of helical forms in the analyzed structure. The elliptic path goes through the region attributed to the helical forms on the Ramachandran map. It seems that contacts in helical fragments are present in all the discussed structural forms of hemoglobin chains. The decrease of non-bonding contacts in the post-energy-minimization structural form of the $\alpha$ chain is probably due to the disappearance of a short helical fragment (green).

### *3.5. Change of molecule size*

The most critical problem concerning the relation between the ellipse-derived structures and the native structures is the question of how much the size of the molecule is changed, that is, what degree of compactness of ellipse-derived structures is necessary to obtain the native form.

The increase in size (volume) of the box containing the protein molecule (the size of which is calculated according to the procedure given in Methods) relative to the native form appeared to be 4.09 and 3.92 for the ellipse-derived $\alpha$ and $\beta$ chains, respectively, and 4.22 and 3.48 for the post-energy-minimization $\alpha$ and $\beta$ chains. The $\beta$ chain decreases in molecular size after the energy minimization procedure, while the $\alpha$ chain increases its size during the search for an optimal structure. Visual analysis shows this to be caused by the elongation of the whole molecule. The inter-helical fragments (two $C$-terminal fragments – pink and yellow) found the inter-helical contacts, causing an increase of non-bonding contacts with a simultaneous increase of molecular size.

## 4. Discussion

The aim of this paper has been to assess whether an ellipse-limited conformational sub-space can be used to deliver the starting structure for the energy minimization procedure used particularly in *ab initio* methods. The degree of similarity between the native form and the post-energy-minimization structure of hemoglobin chains seems unsatisfactory, although an approach toward the native structure is obvious.

The same model of early-stage structural form presentation works much better for proteins with SS bonds, BPTI [18] and ribonuclease [12]. The BPTI molecule is small and contains three disulphide bonds. The energy minimization step including SS bonds in the list of procedure options has led to a significant approach toward the native structure [17]. An approach toward the native form with SS bonds absent from the energy minimization procedure applied to ribonuclease and BPTI has yielded structures far from the native one, representing degrees of similarity comparable to those presented in this paper. Hemoglobin chains have been selected here in order to assess the model for a specific protein with no SS bonds. It seems that the interactions present in the ECEPP force field are insufficient to reach the final form of protein structure.

This paper presents the first step of a comprehensive model of a protein folding procedure. The second step – traditionally called the hydrophobic collapse – is under consideration. Currently, the initial ellipse-limited polypeptide structure is created on the basis of a back step from the native structural form. A separate procedure to predict the initial Φ, Ψ angles belonging to the ellipse-limited conformational sub-space has been completed [19].

## Acknowledgements

Many thanks to Prof. Marek Pawlikowski (Faculty of Chemistry, Jagiellonian University) for helpful discussions.

This work has been supported by Collegium Medicum under grant no. 501/P/133/L.

## References

[1] Dobson C M 2001 *Phil. Trans. R. Soc. Lond.* **B 356** 133

[2] Chan H S and Dill K A 1998 *Proteins Struc. Func. Gen.* **30** 2

[3] Dill K A and Chan H S 1997 *Nature Struct. Biol.* **4** 10

[4] Alonso D O V and Daggett V 1998 *Prot Sci.* **7** 860

[5] Liwo A, Czaplewski C, Pillardy J and Scheraga H A 2001 *J. Chem. Phys.* **115** 2323

[6] Liwo A, Arfukowicz P, Czaplewski C, Ołdziej S, Pillardy J and Scheraga H A 2002 *Proc. Natl. Acad. Sci. USA* **99** 1937

[7] Fernàndez A, Kostov K and Berry R S 1999 *Proc. Natl. Acad. Sci. USA* **96** 12991

[8] Fernàndez A, Colubri A, Aqpigmanesi G and Burastero T 2001 *Physica* **A 293** 358

[9] Sosnick T R, Berry R S, Colubri A and Fernàndez A 2002 *Proteins Struct. Func. Gen.* **49** 15

[10] Roterman I 1995 *J. Theoretical Biol.* **177** 283

[11] Roterman I 1995 *Biochimie* **77** 204

[12] Jurkowski W, Byliński M, Konieczny L, Wiśniowski Z and Roterman I 2004 *Proteins Struct. Func. Bioinf.* **55** 115

[13] Baldwin R L 2002 *Science* **295** 1657

[14] Scheraga H A 1992 *Rev. Comput. Chem.* **3** 73

[15] Gay D M 1983 *ACM Trans. on Mathematical Software* **9** 503

[16] Orengo C A, Bray J E, Hubbard T, LoConte L, Sillitoe I 1999 *Proteins Struct. Func. Gen. Suppl.* **3** 149

[17] Fersht A R and Daggett V 2002 *Cell* **108** 573

[18] Byliński M, Jurkowski W, Konieczny L and Roterman I 2004 *Bioinformatics* **20** 199

[19] Byliński M, Konieczny L and Roterman I 2004 *Early Stage Folding in Proteins – Structure to Sequence Relation* (submitted)