4-1-2004

Conformational Subspace in Simulation of Early-Stage Protein Folding

Wiktor Jurkowski
Uniwersytet Jagielloński w Krakowie

Michał Brylinski
Uniwersytet Jagielloński w Krakowie

Leszek Konieczny
Uniwersytet Jagielloński Collegium Medicum

Zdzisław Wiíniowski
Uniwersytet Jagielloński Collegium Medicum

Irena Roterman
Uniwersytet Jagielloński Collegium Medicum

Recommended Citation

Jurkowski, W., Brylinski, M., Konieczny, L., Wiíniowski, Z., & Roterman, I. (2004). Conformational Subspace in Simulation of Early-Stage Protein Folding. *Proteins: Structure, Function and Genetics, 55* (1), 115-127. https://doi.org/10.1002/prot.20002

PROTEINS: Structure, Function, and Bioinformatics 55:115–127 (2004)

Conformational Subspace in Simulation of Early-Stage Protein Folding

Wiktor Jurkowski,[1,3] Michał Brylinski,[1,3] Leszek Konieczny,[2] Zdzisław Wiíniowski,[3] and Irena Roterman[3]*
[1]*Institute of Chemistry, Jagiellonian University, Krakow, Poland*
[2]*Institute of Medical Biochemistry, Collegium Medicum-Jagiellonian University, Krakow, Poland*
[3]*Department of Bioinformatics and Telemedicine, Collegium Medicum-Jagiellonian University, Krakow, Poland*

ABSTRACT **A probability calculus was used to simulate the early stages of protein folding in ab initio structure prediction. The probabilities of particular ϕ and ψ angles for each of 20 amino acids as they occur in crystal forms of proteins were used to calculate the amount of information necessary for the occurrence of given ϕ and ψ angles to be predicted. It was found that the amount of information needed to predict ϕ and ψ angles with 5° precision is much higher than the amount of information actually carried by individual amino acids in the polypeptide chain. To handle this problem, a limited conformational space for the preliminary search for optimal polypeptide structure is proposed based on a simplified geometrical model of the polypeptide chain and on the probability calculus. These two models, geometric and probabilistic, based on different sources, yield a common conclusion concerning how a limited conformational space can represent an early stage of polypeptide chain-folding simulation. The ribonuclease molecule was used to test the limited conformational space as a tool for modeling early-stage folding. Proteins 2004;55:115–127.**

Key words: protein structure prediction; structural entropy; information

INTRODUCTION

Homology-based protein modeling uses the nearest similar protein structure for prediction. In ab initio methods the early-stage structures are the crucial step in protein structure prediction. Despite many years of struggle with this problem, the results are still not satisfactory. A call for new approaches appeared in the final report of CASP 2000.[1]

The rapid growth of databases in the postgenomic era allows researchers to use probability calculation in protein structure prediction studies.[2] The hidden Markov model has been applied in developing a tool for massive postgenomic databases to analyze as many structures as possible, particularly in the context of nucleic acid sequences on the one hand and the biological activity of proteins on the other.[3]

The probability-based GOR algorithm[4] was designed to predict the allowed secondary structure state in a given sequence context.

Information theory was applied recently in a work[5] intended to verify the role of the local folding code and to identify specific amino acids critical in the formation of local structure. The authors of the model suggested that short amino acid (3–7 aa) sequences would cover the necessary amount of information defining the structure of short polypeptide fragments.

This article introduces a measurable scale [expressed in information entropy units (bits)] of structure predictability for each amino acid. The scale is based on the distribution of ϕ, ψ angles found for a particular amino acid as it appears in proteins of known structure. Analysis of this scale reveals the need to limit the conformational space (conformational subspace) to simulate early-stage protein folding. The results show that a simplified geometrical model based on the notion of the conformational subspace accords with the entropy-based analysis. The subspace appeared as the conformational space limited to the ellipse path.[6] It offers a way forward in attempts to predict protein folding.

MATERIALS AND METHODS

Amino Acid-Dependent ϕ, ψ Angle Distribution

The ϕ, ψ angle distribution in the proteins was analyzed by taking two sets of protein structures, one formed of all the proteins in the PDB,[7] and the second formed of nonredundant proteins. A set of nonredundant protein sequences was selected by using the BLAST[8] algorithm. From the set accessible on the BLAST ftp server (version 2002.10.02), only those with the highest nonredundancy (*p* value 1-E-7) were chosen. This subset was narrowed down to a group of structures with known three-dimensional (3D) structures (solved with experimental techniques) deposited in the PDB. The ϕ, ψ angles were divided into 20 groups for each amino acid separately. The probability of ϕ_i and ψ_j occurrence for a particular amino acid is denoted p_{ij}.

Grant sponsor: Collegium Medicum; Grant number: 501/P/133/L.

*Correspondence to: Irena Roterman, Department of Bioinformatics and Telemedicine, Collegium Medicum, Jagiellonian University, Kopernika 17, 31-501 Kraków, Poland. E-mail: myroterm@cyf-kr.edu.pl

Received 13 May 2003; Accepted 19 September 2003

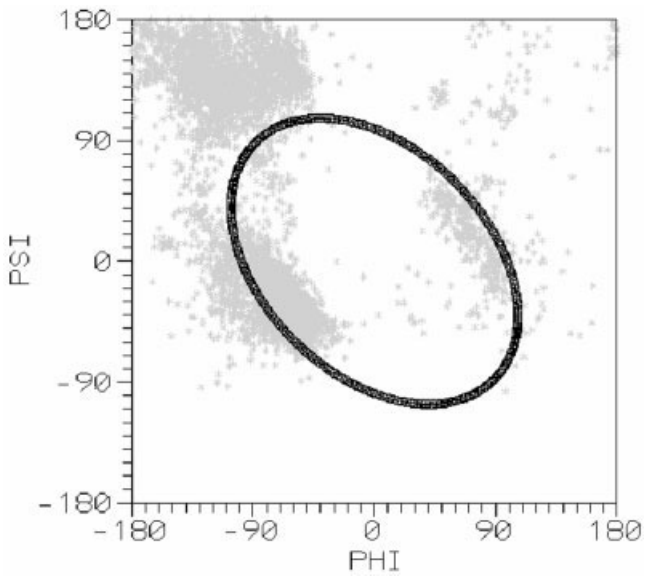

Fig. 1. ϕ, ψ angle distribution as it appears in proteins (selection presented in Methods) versus the ellipse path (black line).

Amino Acid Frequency

Both databases (complete PDB and nonredundant subbase) were also used to measure the frequency of occurrence of a particular amino acid. The probability of a particular amino acid's occurrence in the analyzed proteins is denoted p_f.

ϕ, ψ Angle Distribution for Ellipse Presentation

Proteins representing different structures were selected: mostly helical protein (α- and β-chains of hemoglobin: 3HHB), mostly β-structure (light and heavy chains of Fab fragment of IgG: 2FB4), and structurally mixed proteins belonging to the serpine family (uncleaved ovalbumin: 1OVA; cleaved bovine antithrombin: 1ATT; cleaved human α-1-antichymotrypsin: 2ACH; human antithrombin chain I: 1AZX; human antithrombin chain L: 2ANT; α-1-antitrypsin: 7API) to show the relation between the ϕ, ψ angle distribution on the Ramachandran map and the ellipse path (Fig. 1).

Energy Distribution

Maps representing the energy distribution all over the Ramachandran map were calculated by using the ECEPP force field.[9] The structures were created on a 5° grid for ϕ and ψ angles. Energy minimization was performed for each grid point for the molecule ACE-X-MNE (X represents the amino acid under consideration), with the ϕ and ψ angles constrained at the appropriate grid point, whereas the rest of the molecule was allowed to change its structure.

The energy distribution was transformed to the probability scale by a two-step procedure: 1) the energy distribution was standardized (0–1 scale) with the energy integral equal to 1 and 2) the values were transformed to the reverse form in the opposite relation: the higher the energy for a particular grid point, the lower the probability for this structure to occur.

Structural Entropy Scale

All types of maps (energy-based and both versions of the ϕ, ψ angle distribution-based map) were transformed to the normalized probability scale.

The Shannon definition[10] treating the amount of information (in bits) as probability-dependent was adopted to measure the amount of information carried by a particular amino acid:

$$SI(p_f) = -\log_2 p_f[\text{bit}] \quad (1)$$

The Shannon entropy[10] expressing the mean level of uncertainty in ϕ ψ prediction (using n degree step) can be calculated as follows:

$$S_k = \sum_{i=1}^{360/n} \sum_{j=1}^{360/n} -p_{ij}\log_2(p_{ij}) \text{ [bit]} \quad (2)$$

where: S_k is informational entropy, p_{ij} is the probability for the k-th amino acid to represent the i-th ϕ and j-th ψ dihedral angles (the ij-th grid point), n is step size.

S_k can evaluate the level of uncertainty in ϕ, ψ selection for a particular amino acid. The higher the S_k for the amino acid, the more difficult it is to predict its structure expressed by ϕ, ψ angles (for an assumed precision).

Limitation of the Conformational Space

The conformational space was limited to an ellipse path on the Ramachandran map based on a previously described model[6] (presented in abbreviated form in the Appendix).

Ribonuclease as a Test Protein

Ribonuclease (5RAT according to PDB identification) was taken to test the usefulness of the model. The ϕ, ψ angles were calculated for each amino acid in the polypeptide chain as they appear in the native form of the protein. The criterion of the shortest distance between the observed ϕ, ψ angles and those belonging to the ellipse was used to find the ϕ_e, ψ_e angles.

The ellipse-derived structure for ribonuclease was created by using the ECEPP/3 program. The Omega dihedral angles were taken as 180° for all amino acids. The side-chain structures were formed according to ECEPP/3 standards and were free to rotate during the energy minimization procedure.

Energy Minimization Procedure

The energy minimization procedure was performed by using the ECEPP/3 program for the ellipse-derived structure of ribonuclease.[11] The ϕ, ψ angles were calculated for postminimization structures. The unconstrained minimization solver with analytical gradient[12] was used. The values of absolute and relative function convergence tolerances were set at 1×10^{-3} and 1×10^{-5}, respectively. The energy minimization procedure was conducted both with and without properly defined disulfide bonds. The coordinates and values of the backbone dihedral angles were

TABLE I. Probability- and Energy-Based Structural Entropy (bit)

	Probability-Based					
	5°		10°		Energy-Based	
AA	PDB	Nonredundant	PDB	Nonredundant	5°	10°
PRO	7.92	8.33	6.02	6.51	9.16	7.11
ALA	8.88	8.86	6.98	7.00	11.82	9.52
ILE	8.93	8.78	7.00	6.91	10.98	8.67
LEU	8.98	8.86	7.07	7.04	11.71	9.51
VAL	9.03	8.89	7.09	7.06	11.23	9.09
MET	9.07	8.86	7.18	7.12	11.42	9.12
GLU	9.11	9.05	7.19	7.22	11.67	9.43
TRP	9.31	9.10	7.40	7.38	11.74	9.46
GLN	9.33	9.16	7.42	7.36	11.60	9.28
ARG	9.36	9.29	7.45	7.47	11.61	9.31
PHE	9.47	9.37	7.54	7.52	11.52	9.20
LYS	9.53	9.45	7.61	7.62	11.47	9.17
THR	9.67	9.49	7.88	7.66	11.24	8.94
TYR	9.56	9.33	7.63	7.53	11.47	9.18
ASP	9.81	9.68	7.90	7.81	11.63	9.29
HIS	9.82	9.67	7.91	7.92	11.80	9.46
SER	9.86	9.69	7.94	7.81	11.89	9.58
CYS	9.96	9.71	8.04	7.94	11.73	9.47
ASN	10.05	9.90	8.13	8.11	11.65	9.28
GLY	10.72	10.60	8.80	8.87	12.12	9.79

S_k (structural entropy [bit]) expressing the potential predictability of amino acids based on the ϕ, ψ angle distribution (for complete set of proteins in PDB and selected nonredundant subset) and energy-based distribution (after transformation to the probability scale) for 5° and 10° step precision.

saved for analysis at 10-step intervals. The energy minimization procedure for ribonuclease was done on an SGI Origin 2000 in the computing center of TASK in Gdansk.

Structure Comparison

The structures (native, elliptical, and postenergy minimization) were compared by using different criteria:

1. The distances between the geometric center of the molecule and the sequential Cα atoms in the polypeptide chain were calculated. This plot revealed a rough degree of similarity. The polypeptide fragments distinguished according to the profile were also characterized by using root-mean-square deviation (RMSD) calculation. The RMSD values for selected fragments were calculated after overlapping the fragments taken from native and ellipse-derived structural forms. The RMSD value was calculated per one amino acid in the polypeptide fragment
2. The number of native nonbonding interactions was calculated for all structural forms, assuming a cutoff distance equal to 12 Å.
3. The box large enough to contain the whole molecule was also calculated for each structural form of the protein molecule. The box size was calculated as follows: the longest Cα-Cα distance was taken as the D_z measure (distance along z axis), the longest Cα-Cα distance in the xy plane was taken as the D_y measure (distance along y axis), and the difference between the highest and lowest values of x was taken as the measure for the D_x box edge.

RESULTS

Structural Entropy Scales

The S_k values (according to Eq. 2) calculated for each amino acid were used to produce the amino acid-dependent S_k scale (Table I) incorporating both approaches (energy and two ϕ, ψ distribution bases). The S_k values estimate the predictability of the structure for a particular amino acid. The higher the S_k, the higher the level of uncertainty. Obviously, PRO and GLY are placed at the opposite terminal positions in the ordered chain of S_k values in both approaches. Table I also illustrates predictability versus step size.

Keeping in mind that one amino acid (assuming equal probability of a particular amino acid's occurrence, $p = 1/20$) holds $SI(p) = 4.32$ bits (according to Eq. 1), one can easily judge that the S_k values (according to Eq. 2) shown in Table I, expressing the mean amount of information necessary to predict the ϕ, ψ angles, are much higher than that. The same calculation for lower precision (10 × 10° grid) proves that decreasing the precision does not solve the problem, indicating that a significant increase of the grid step—much too big to be satisfactory—is necessary.

The individual frequency of a particular amino acid's occurrence in proteins differentiates the level of information delivered by this amino acid. The information scale based on that frequency is shown in Table II and Figure 2. The S_k values are still above the individual amino acid-dependent information level: $SI(p_f)$.

TABLE II. Information: Carried by Amino Acid and Necessary to Predict Particular Phi, Psi Angle

AA	Amount of information held by amino acid		Insufficiency/excess	SE (10°)	SE (5°)	SE (1°)
	Complete PDB	Nonredundant				
GLY	3.727	3.805	−2.013	5.740	6.630	7.806
ASP	4.121	4.117	−0.895	5.016	5.950	7.073
LEU	3.549	3.492	−0.888	4.437	5.380	6.438
LYS	3.937	3.908	−0.827	4.764	5.710	6.789
ALA	3.661	3.662	−0.801	4.462	5.419	6.409
SER	4.060	4.095	−0.797	4.857	5.785	6.975
ASN	4.494	4.545	−0.692	5.186	6.126	7.267
GLU	3.905	3.833	−0.645	4.550	5.498	6.520
THR	4.107	4.196	−0.472	4.579	5.502	6.720
ARG	4.362	4.249	−0.288	4.650	5.600	6.677
VAL	3.868	3.886	−0.240	4.108	5.057	6.233
GLN	4.684	4.663	0.017	4.667	5.607	6.676
ILE	4.203	4.151	0.088	4.115	5.064	6.208
PHE	4.679	4.713	0.151	4.528	5.466	6.617
TYR	4.836	4.941	0.262	4.574	5.498	6.685
PRO	4.451	4.442	0.389	4.062	4.958	6.124
HIS	5.461	5.477	0.593	4.868	5.805	6.965
CYS	5.597	5.544	0.805	4.792	5.720	6.937
MET	5.636	5.614	1.152	4.484	5.425	6.494
TRP	6.091	6.236	1.579	4.512	5.444	6.581

SE (structural entropy [bit]) calculated according to Eq. 2 for the conformational subspace limited to the ellipse path probability distribution at 1° (column 5), 5° (column 6) and 10° grid step size (column 7) precision. SI, the amount of information (bits) carried by a particular amino acid calculated (according to Eq. 1) on the basis of individual frequency of occurrence in the whole PDB (column 2) and nonredundant subset (column 3). The insufficiency/excess of information versus the 10° stepsize ellipse path probability distribution (column 4) is the result of subtracting the values from columns 2 and 5.

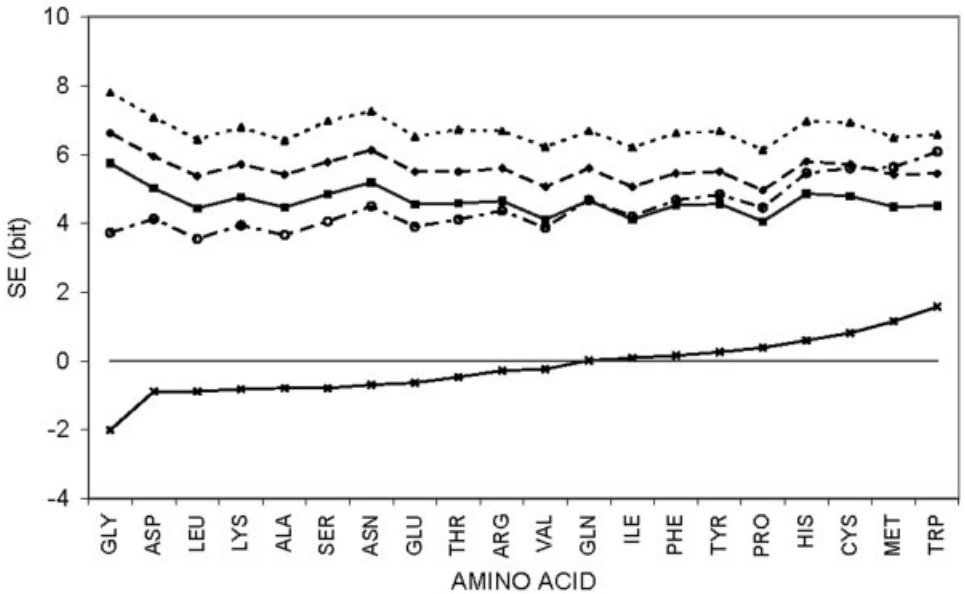

Fig. 2. Plot representing the amount of information (bits) necessary for 10° step prediction in relation to the amount of information carried by an individual amino acid. ▲, SE (1°); ♦, SE (5°); ■, SE (10°); ○, SI (p_f) carried by amino acid; ×, difference between SI(p_f) and SE (10°).

The Ellipse Path

A previously introduced model was based on a geometrical representation of the polypeptide chain according to two parameters: the V-angle, expressing the dihedral angle between two sequential peptide bond planes and the R-radius of curvature related to the V-angle (details in Ref. 6 and in Appendix). It showed that the ellipse path can characterize the polypeptide chain structure, assuming

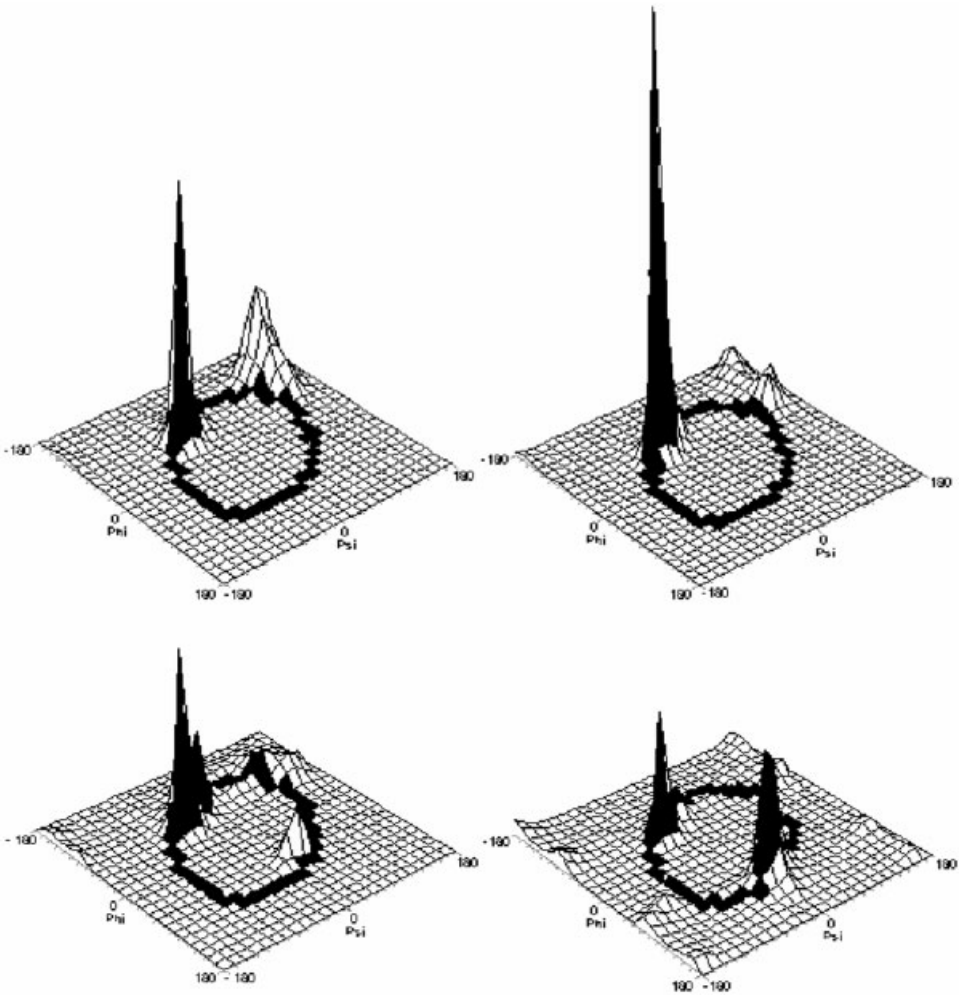

Fig. 3. 3D representation of probability distribution calculated for 5° grid size. The black fields distinguish the ellipse path to show the amino acid-dependent relation of the ϕ, ψ angle distribution versus the ellipse path for VAL, ALA, ASN, and GLY, respectively.

that only peptide bond plane orientation determines the structure of the polypeptide backbone. This model does not incorporate side-chain–side-chain interaction. Figure 1 shows the ellipse path found to be the optimal path for the early search for polypeptide chain structure versus experimentally measured distributions of ϕ, ψ angles as they appeared in selected proteins (see Methods). The standard ϕ, ψ distributions usually presented in protein crystallographic data show this specific elliptical distribution.

The relation between the ϕ, ψ angle distribution characteristic for a particular amino acid and the ellipse path is also shown in Figure 3 for selected amino acids (the criteria for selection are given later in this article).

A search of the whole conformational space may be replaced by one limited to the ellipse path subspace. The effectiveness of such a limited subspace can be proved as follows. The probability distribution along the ellipse path can be calculated after moving all points (ϕ, ψ angles on the Ramachandran map) to the closest point on the ellipse (shortest distance criterion). The probability distribution along the ellipse path is shown in Figure 4. The x axis of the profiles presented in Figure 4 expresses the t parameter of the ellipse equation. The starting point ($t = 0$) is $\phi = 90°$ and $\psi = -90°$. The clockwise movement along the ellipse path is represented by the increase in t values (1° step). The right-handed helical region is reached for t values in the range of 90–120°, and β-structural forms are represented by t values in the range of 180–220°. Left-handed helical forms are represented by t values close to 270° and above.

Two profiles are presented: the probability-based ϕ ψ distribution (black line: complete set of proteins in PDB) and the energy-based distribution: transformed to the probability scale (gray line) calculation.

Table II presents the *SE* values (*SE* denotes the S_k values calculated for the ellipse path)—the sum of all the t values (see Eq. 2 in Appendix and Eq. 2)—for the set of structures (with 1°, 5°, and 10° precision for the t parameter) limited to the ellipse path versus the amount of information stored in a particular amino acid according to its frequency of occurrence in the proteins (Fig. 2).

Energy- and Probability-Based Profiles

The energy-derived and probability-derived profiles differ (Fig. 4). The main differences are related to the two

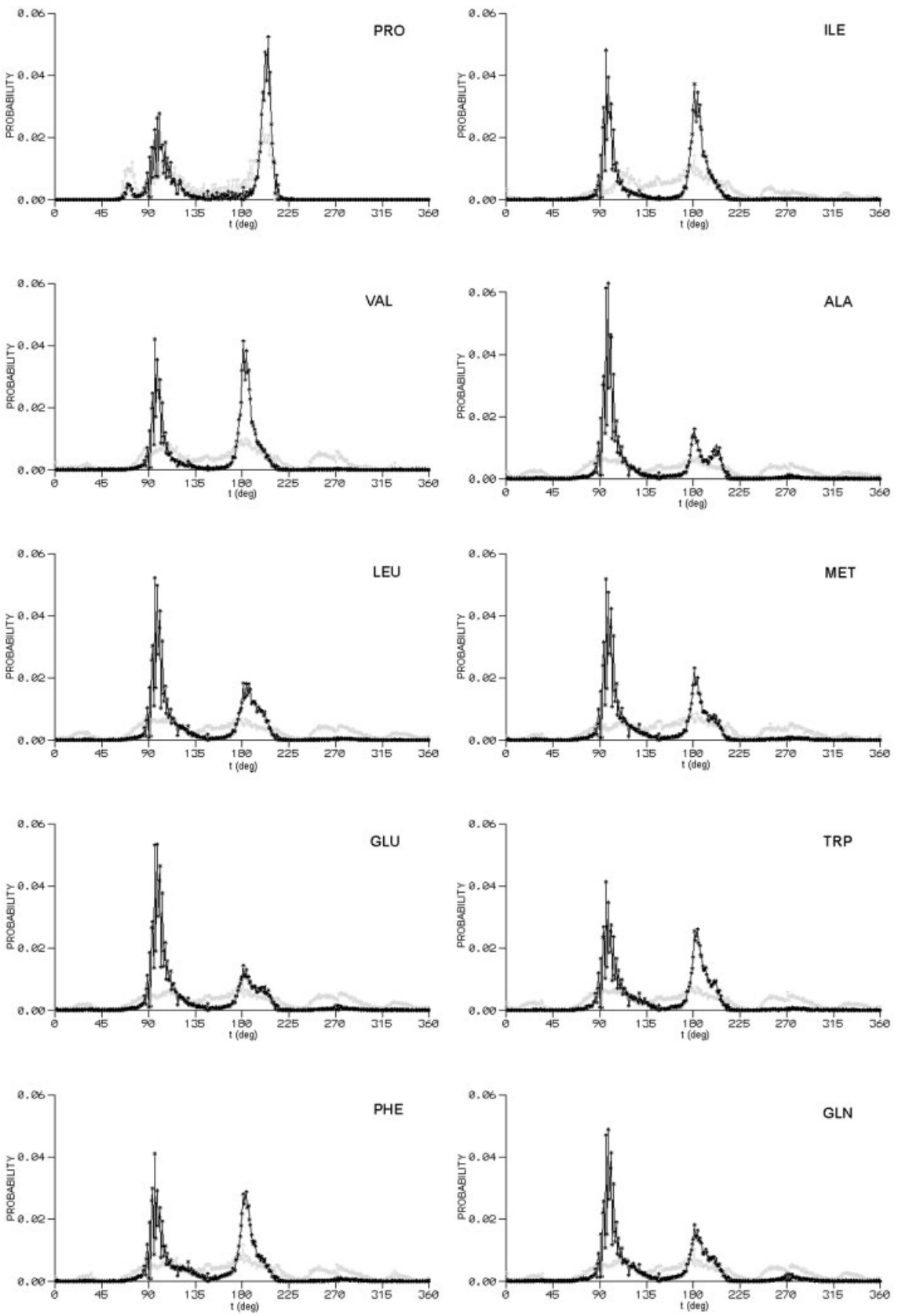

Fig. 4. Amino acid-dependent probability profiles as they appear after moving all ϕ, ψ angles to the nearest point on the ellipse path in order of increasing SE. The *t* parameter (ellipse equation, Eq. 2 App.) equal to 0° represents the starting point at $\phi = 90°$ and $\psi = -90°$, then going clockwise along the ellipse by 1° steps. Black line, probability distribution-based profile; gray line, energy distribution-based profile.

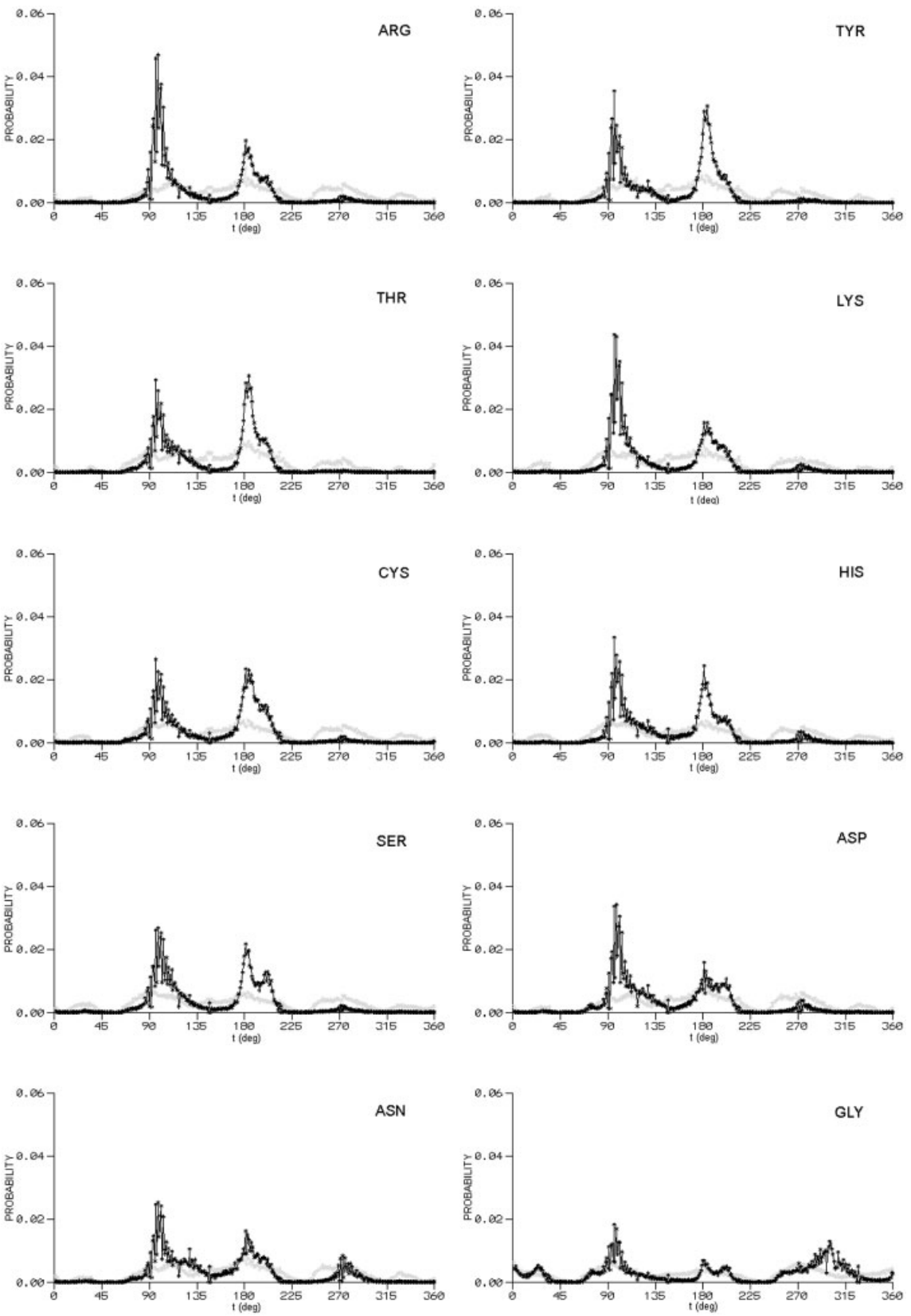

Figure 4. (Continued.)

most significant energy minima: α-helical and β-structural. It is obvious that these two energy minima cannot be observed for dipeptides, where no hydrogen bonds significant for these two structures are available. However, there are some exceptions, including ASN and GLY, where these two profiles (energy- and ϕ, ψ angle distribution-based) are quite similar.

The ellipse path probability distribution profiles characterize amino acids in respect to their preferred structural differentiation. There are amino acids with an almost "binomial" distribution, with two forms well dominating (VAL). There are amino acids with quite a differentiated profile of very low predictability, such as GLY and ASN. To illustrate these differences, these amino acids are presented in 3D diagrams in Figure 3.

The quantitative relations between the SE for different degrees of precision (1°, 5°, and 10°) and the information stored in a particular amino acid are shown in Table II and Figure 2.

Particularly interesting are the profiles in the region of t values between 95° (α_R helical area) and 180° (β-structural area). This fragment represents the area on the Ramachandran map with significantly lower occupation (see Fig. 1). The differentiation of SE values is caused mostly by the differences in this area (excluding PRO and GLY for obvious reasons). TYR and ASP even show an additional small maximum in this fragment of the ellipse. The β-structure region is also represented in different forms. One maximum is observed in the case of VAL. Some other amino acids show the β-structural probability maximum differentiated by splitting into two local maxima (e.g., ALA) (see also Fig. 3).

The relation between the amount of information coded by an amino acid and the SE value for the 10° step reveals that almost half (n = 9) of the amino acids have excess information, the structure of which can be predicted (Table II; Fig. 3.). The others (n = 11) represent lower levels of information and their predictability is much less. The highest unpredictability is attributed to GLY.

Structural Analysis of Ribonuclease as a Model Test

ϕ, ψ *dihedral angle changes*

ϕ, ψ angle distributions exposing the range of dihedral angle change are presented for all discussed structural forms: for the native form of ribonuclease [Fig. 5(A)], for the ellipse-derived structure [Fig. 5(B)], for the postenergy minimization form with SS-bonds present [Fig. 5(C)], and for the postenergy minimization structure with SS-bonds absent [Fig. 5(D)].

Spatial distribution of Cα atoms versus the geometrical center

The profile of the size of vectors linking the geometrical center with sequential Cα atoms gives insight into the 3D relative displacements versus native-form protein.[13] The structural similarity may be disclosed by overlapping of lines representing two compared structures.

The overlapped fragment of $D_{\text{center-C}\alpha}$ profiles reveals identical fragments. Parallel orientation of profiles represents similar structural forms in the two compared molecules oriented in space differently. The generally higher values of the vector length in respect to native structure are obviously due to the extension of the structure that is always associated with the transformation from the native to ellipse path-delimited structure. Dissimilar fragments of the profile show different structural forms in the protein molecule.

The profiles for all analyzed structures of ribonuclease are presented in Figure 6. Four fragments can be distinguished. The fragment containing amino acids 46–80 was distinguished because of the similarity of profiles in all structural forms. This fragment, individually overlapped, gives an RMSD value < 1 Å. Two N-terminal fragments, 1–17 aa and 18–45 aa, were selected as representing parallel orientation of $D_{\text{center-C}\alpha}$ profiles, with low RMSD values also. The C-terminal fragment (81–124 aa) is represented by quite different forms of $D_{\text{center-C}\alpha}$ profiles and described by a rather high RMSD value.

Visual Analysis

Figure 5 visualizes structural changes in ribonuclease in different conditions. The color notation differentiates particular polypeptide fragments and distinguishes them according to their similarity as measured by the $D_{\text{center-C}\alpha}$ vector profiles (Fig. 6.). The same color notation is used for ϕ, ψ angle distributions and RMSD fragments (Fig. 5).

The ϕ, ψ angle distribution in the postenergy minimization procedure with SS-bonds present seems not to change much. Two energy minima, helical and C7eq, seem to stabilize the structure. No signs of the presence of β-structural forms (C5, β-parallel, β-antiparallel) can be observed in that structure. Generally, β-structure is difficult to reach in any ab initio protein structure prediction.[1]

Quite a good approach (the characteristic boat-shaped form) was reached for the same procedure with SS-bonds defined according to the natural system present in this molecule.

Nonbonding Interaction

The nonbonding interactions present in all discussed structural forms are shown in Figure 7. Significant similarity of nonbonding contact distributions can be seen between native and post energy minimization forms with SS-bonds taken into account during energy minimization. The ellipse-derived structure also reveals the presence of one part of the contact map present in the native form of ribonuclease.

The percentage of native nonbonding interactions present in the ellipse-derived structure is equal to 32.98%, and 41.50% and 33.60% for postenergy minimization structures with and without SS-bonds, respectively. This quantity obviously depends on the molecule under study[14] and is strongly related to the percentage of helical forms in the analyzed structure. The ellipse path goes through the region attributed to helical forms on the Ramachandran map; this is probably caused by the presence of all helical nonbonding contacts in ellipse-derived and postminimization structures. However, the appearance of the characteristic distribution of nonbonding contacts [Fig. 7(C)] makes the model promising.

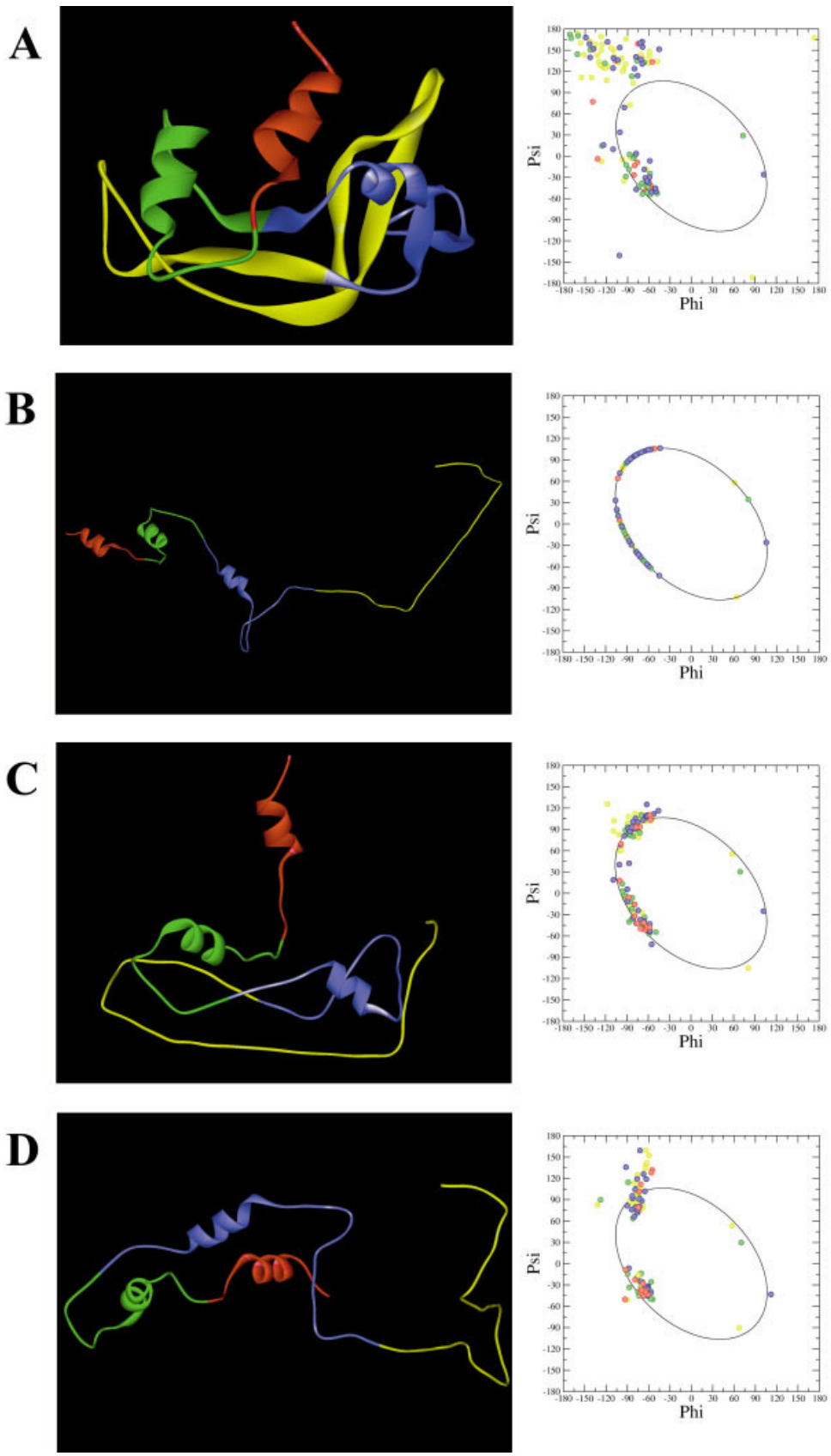

Fig. 5. The structure of ribonuclease and its ϕ, ψ angle distribution versus the ellipse path. **A:** Native form. **B:** ellipse-based structure. **C:** Postenergy minimization structural form of ribonuclease with SS-bonds present in the energy minimization procedure. **D:** Postenergy minimization structural form of ribonuclease with SS-bonds absent in the energy minimization procedure.

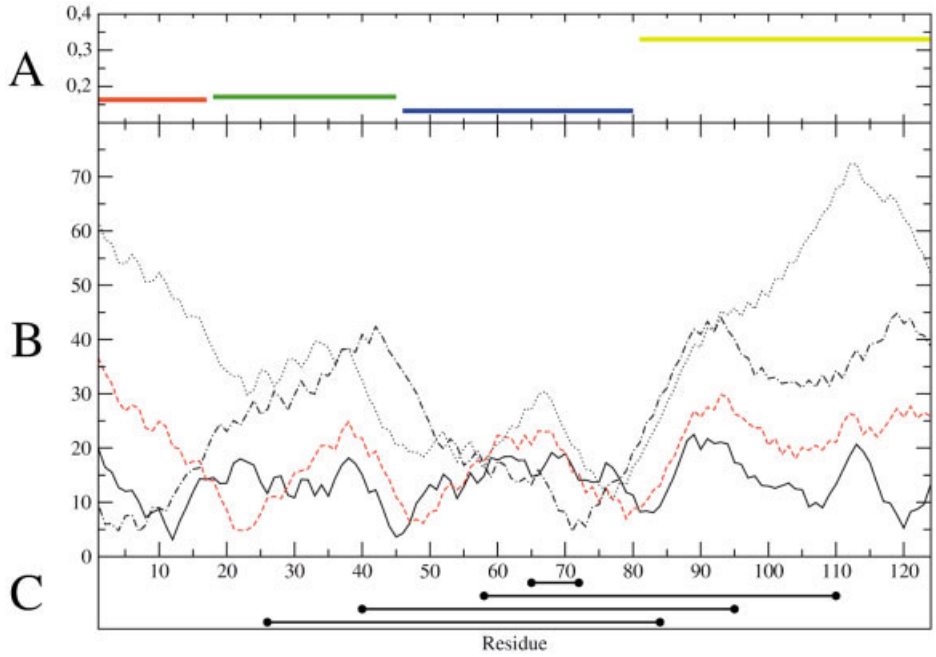

Fig. 6. Comparison of structural forms of the ribonuclease molecule. **A:** RMSD (per residue) calculated for structurally differentiated polypeptide fragments. The fragments were defined according to the profile presented in B. The parallel fragments of curves represent the correct spatial orientation of the polypeptide, whereas the dissimilar regularity of the curve represents fragments with low similarity of the spatial orientation of the particular polypeptide fragment. **B:** Profile representing the distribution of distances linking the geometrical center of the molecule with sequential Cα atoms (called $D_{center\text{-}C\alpha}$ in the text). Continuous (solid) line, native form; Dotted line, ellipse-derived structure; dashed line, postenergy minimization structure with SS-bonds present; dotted/dashed line, post-energy minimization structure with SS-bonds absent; C, SS-bond system in ribonuclease. The color notation introduced in all graphic presentations in this article corresponds to the fragments distinguished in this figure.

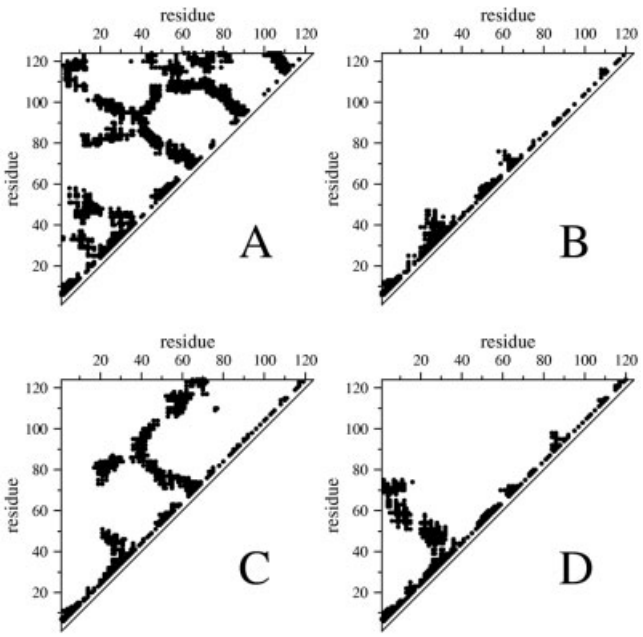

Fig. 7. Nonbonding contacts in ribonuclease. **A:**Native form. **B:** Ellipse-derived structure. **C:** Postenergy minimization with SS-bonds present. **D:** Postenergy minimization with SS-bonds absent.

Molecule Size Change

The most critical problem concerning the relation between ellipse-derived structures and native structures is the question of how much the size of the molecule is changed (i.e., what degree of compactness of ellipse-derived structures is necessary to reach the native form. The relative (vs the native form of protein) increase of box size (volume) containing the whole protein molecule (the size of which is calculated according to the procedure given in Methods) appeared to be 9.59 for ellipse-derived and 4.96 for postenergy minimization with SS-bonds present in the energy minimization procedure. The box edge (see Methods) proportions are as follows: 29.99:27.42, 63.20:27.32, and 81.99:41.82. The postenergy minimization structure (with SS-bonds) reached one edge length quite well, whereas the two others are still too large.

DISCUSSION

The search for the global energy minimum is the main problem in protein structure prediction.[15] Analysis of the conformational space, discussed extensively,[16] led to the energy landscape perspective. Here we approached it by introducing the idea of the conformational subspace to limit the multidimensional energy surface. The previously introduced geometrical model using an ellipse as the optimal path for the energy minimization procedure may also be treated as an early stage of polypeptide chain-folding simulation.[6,17] It was found that the ellipse pathway—the outcome of structural analysis (Appendix Fig. 1.)—satisfies the following conditions:

1. The structures created according to the ellipse path exhaust the spectrum of polypeptide chain shapes from V-angle equal to 0° (helix: low R) to V equal to 180° (β-like structures: very high R) and can be very easily transformed into the form of ϕ, ψ angles.[18–21]
2. The ellipse pathway links all structurally important areas on the Ramachandran map (α_R helix through the C7eq energy minima and then to the α_L-helix). It may be treated as controlled passage through the energy barriers.[19–21] It may also be seen in helix melting simulations.[22,23]
3. It represents the conformational subspace to the extent that the structure can be predicted based only on the amino acid sequence. The measurable scale of amino acid structural predictability reveals that amino acids are structurally more or less predictable. The problem of the specific sequences that represent characteristic structural motives has been discussed.[24–27]

To overcome the problem of the multidimensional energy surface, a simplified model of polypeptide chain representation has been suggested.[28–30] Our alternative solution to this problem is to limit the conformational space. The S_k values calculated for amino acids characterize each of them very well. Using the flat surface all over the Ramachandram map (assuming equal probabilities for each ϕ, ψ value with 5° step size), the S_x value (x denotes a virtual amino acid with no particular structure preferable) is equal to 12.311 bits. The amino acid representing the closest S_k value is GLY. The dispersion of ϕ, ψ angles for this amino acid represents the lowest level of determination. The S_x value is useful as a relative scale to characterize the set of amino acids.

The selected protein molecule—ribonuclease—appeared to prove the model, particularly because simple energy minimization was performed to obtain the final structure. The influence of molecular dynamics simulation added to the procedure delivers much better results (the size increase ratio vs the native form is only 2.1, unpublished result).

The method presented in this article was assumed to deliver the model for early-stage folding structures. The stages distinguished as partitioning the protein-folding process proposed by Ferguson and Fersht[31] are as follows: 1) specific or nonspecific chain collapse; 2) formation of secondary and tertiary structure, according to the balance of local and nonlocal, native and non-native interactions; and 3) desolvation of the chain as it folds to a lower energy conformation. The proposed model is assumed to describe the first two steps in an event sequence oriented toward reaching the native structure. All examined structures analyzed according to the presented model—BPTI[32] (small molecule used very frequently as the model for structural analysis), lysozyme[33] (the influence of molecular dynamics simulation is discussed), hemoglobin chains[34] (a protein with no disulphide bonds) and the whole serpine family[17,35] (large molecules with differentiated structural forms demonstrating high similarity)—were proved in regard to hydrophobic center creation. The analysis indicated that the hydrophobicity distribution in the protein molecules obtained by the presented model should be corrected. The next step of the folding model, complementary to the ellipse path-derived structure and describing the creation of the hydrophobic center, is under consideration.

REFERENCES

1. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins 2001;Suppl 5:2–7.
2. Hargbo J, Elofsson A. Hidden Markov models that use predicted secondary structures for fold recognition. Proteins 1999;36:68–76.
3. Wrabl JO, Larson SA, Hilser VJ. Thermodynamic propensities of amino acids in the native state ensemble: implications for fold recognition. Protein Sci 2001;10:1032–1045.
4. Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 1996;266:540–553.
5. Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. Proteins 2000;38:149–164.
6. Roterman I. Modeling of optimal simulation path in the peptide chain folding—studies based on geometry of alanine heptapeptide. J Theo Biol 1995;177:283–288.
7. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242 (release #99 January 2002).
8. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410
9. Scheraga HA. Predicting three-dimensional structures of oligopeptides. In: Lipkowitz KB, Boyd DB, editors. Reviews in computational chemistry. Vol. 3. New York: VCH Publishers Inc, 1992. p. 73–130
10. Shannon CEA. Mathematical theory of communication. Bell Syst Tech J 1948;27:379–423.
11. Scheraga H.A. Predicting three-dimensional structures of oligopeptides. Rev Comput Chem 1992;3:73–142.
12. Wang J, Cieplak P, Kollman PA. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?. J Comput Chem 2000;21:1049–1074.
13. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of Ab initio three-dimensional prediction, secondary structure and contacts prediction. Proteins 1999;Suppl 3:149–170.
14. Fersht AR, Daggett V. Protein folding and unfolding at atomic resolution. Cell 2002;108**:**573–582.
15. Zhang Y, Kohara D, Skolnick J. Local energy landscape flattering: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 2002;48:192–201.
16. Chan H-S, Dill KA. Protein folding in the landscape perspective: chevron plots and non-Arhenius kinetics. Proteins 1998;30:2–33.
17. Leluk J, Konieczny L, Roterman I. Similarity search in proteins. Bioinformatics 2003;19:117–124.
18. Flory PJ. Principles of polymer chemistry. Ithaca, NY: Cornell University Press; 1953.
19. Higo J, Ito N, Kuroda M, Ono S, Nakajima N, Nakamura H. Energy landscape of a peptide consisting of alfa-helix, 310 helix, beta-turn, beta-hairpin and other disordered conformations. Protein Sci 2001;10:1160–1171.
20. Hayward S. Peptide-plane flipping in proteins Protein Sci 2001;10: 2219–2227.
21. Garcia P, Serrano L, Durand D, Rico M, Bruix M. NMR and SAXS characterization of the denatured state of the chemotactic protein CheY: implication for protein folding initiation. Protein Sci 2001; 10:1100–1112.
22. Huo S, Straub JE. Direct computation of long time processes in peptides and proteins: reaction path study of the coil-to-helix transition in polyalanine. Proteins 1999;36:249–261.
23. Daggett V, Levitt M. Molecular dynamics simulation of helix denaturation. J Mol Biol 1992;223:1121–1138.
24. De Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 2000;41:271–287.
25. Geourion C, Combet C, Blanchet C, Deleage G. Identification of related proteins with weak sequence identity using secondary structure information. Protein Sci 2001;10:788–797.
26. Solis AD, Rackovsky S. Optimally informative backbone structural propensities in proteins. Proteins 2002;48:463–486.
27. Salwiński Ł, Eisenberg D. Motif-based fold assignment. Protein Sci 2002;10:2460–2469.

28. Lewitt M. A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 1976;104:50–107.
29. Liwo A, Czaplewski C, Pillardy J, Scheraga H. Cummulation-based expression for the multibody terms for the correction between local and electrostatic interaction in the united residue force field. J Chem Phys 2001;115:2323–2347.
30. Liwo A, Arłukowicz P, Czaplewski C, Ołdziej S, Pillardy J, Scheraga H. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape—application to the UNRES force field. Proc Natl Acad Sci USA 2002;99:1937–1942.
31. Ferguson N, Fersht AR. Early events in protein folding. Curr Opin Struct Biol 2003;13:75–81.
32. Bryliński M, Jurkowski W, Konieczny L, Roterman I. Limited conformational space for early stage protein folding . Bioinformatics 2003. Forthcoming.
33. Jurkowski W, Bryliński M, Konieczny L, Roterman I. Lysozyme folded in silico according to limited conformational sub-space. Submitted for publication.
34. Bryliński M, Jurkowski W, Konieczny L, Roterman I. Human α and β hemoglobin chains folded according to limited conformational sub-space. Submitted for publication.
35. Bryliński M, Jurkowski W, Leluk J, Piwowar M, Konieczny L, Roterman I. Consensus structure for protein families. Submitted for publication.

APPENDIX

The main assumption for the model presented below is that all structural forms of polypeptides in proteins can be treated as helical. The β-structure in this approach is a helix with a very large radius of curvature. The radius of curvature depends on the V-angle, which expresses the dihedral angle between two sequential peptide bond planes. The quantitative analysis of the relation between these two parameters (V and R) used the following procedure:

1. The structure of the alanine pentapeptide was created for each 5° grid point on the Ramachandran map. Each alanine present in the pentapetide represented the ϕ, ψ angles appropriate for a particular grid point.
2. Before the parameters (R and V) were calculated, all structures (for each grid point) were oriented in a unified way: the averaged position of the carbonyl oxygen atoms and the averaged position of carbonyl carbon atoms determined the *z* axis.
3. The radius of curvature was calculated for projections of Cα atoms on the *xy* plane. The radius of curvature for extended (and β-structural) forms is very large (theoretically infinite). This is why the natural logarithmic scale was introduced to express the magnitude of R.
4. The V angle was calculated as the difference between the tilt of the central peptide bond plane and the tilt of two (averaged) neighboring peptide bond planes.

The Ramachandran map expressing the V angle distribution and R radius of curvature (in ln scale) is shown in Figure A1.

The ln(R) dependence on the V angle for structures representing low-energy conformations is shown in Figure A2. The approximation function found for this relation is as follows:

$$\ln(R) = 0.000340\ V^2 - 0.02009\ V + 0.848 \tag{A1}$$

The distribution of ϕ, ψ angles of structures that satisfy the above equation is shown in Figure A2. The ellipse path found based on this distribution is as follows:

$$\begin{aligned} \text{Phi} &= -A\cos(t) - B\sin(t) \\ \text{Psi} &= A\cos(t) - B\sin(t) \end{aligned} \tag{2}$$

where A and B are long and short ellipse diagonals, respectively.

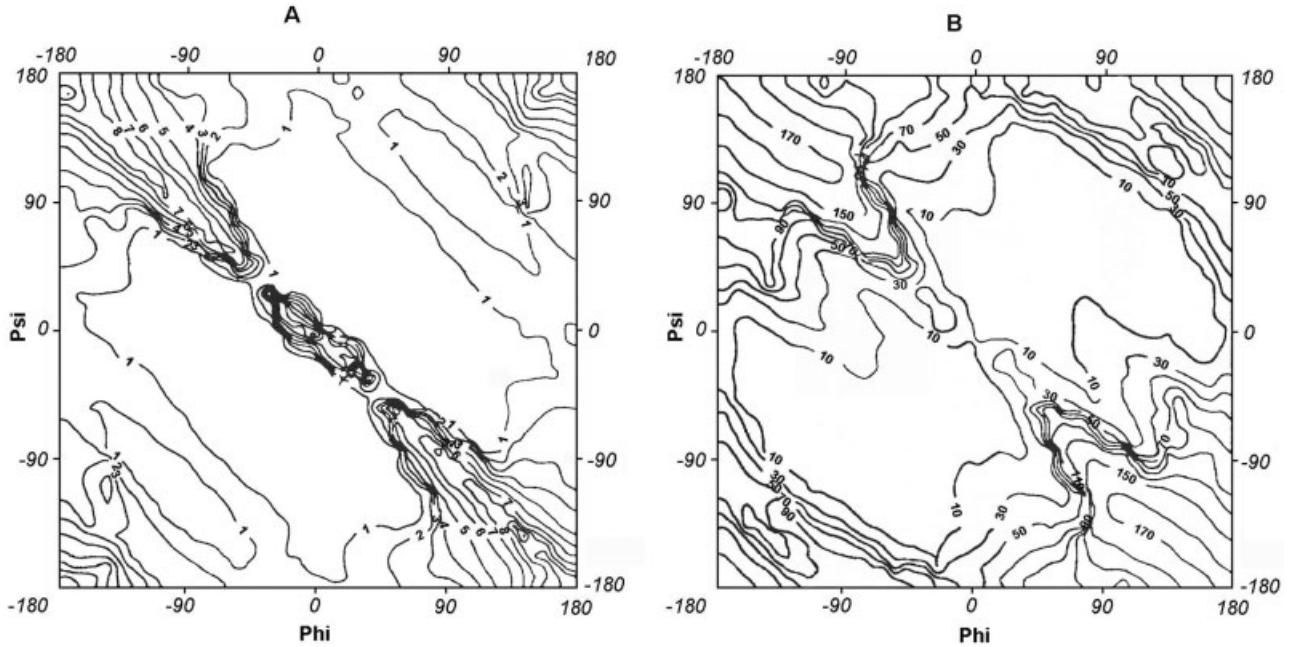

Fig. A1. Distribution of geometrical parameters all over the ϕ-ψ map. **A:** Radius of curvature R on natural logarithmic scale. **B:** Dihedral angle (V) between two sequential peptide bond planes.

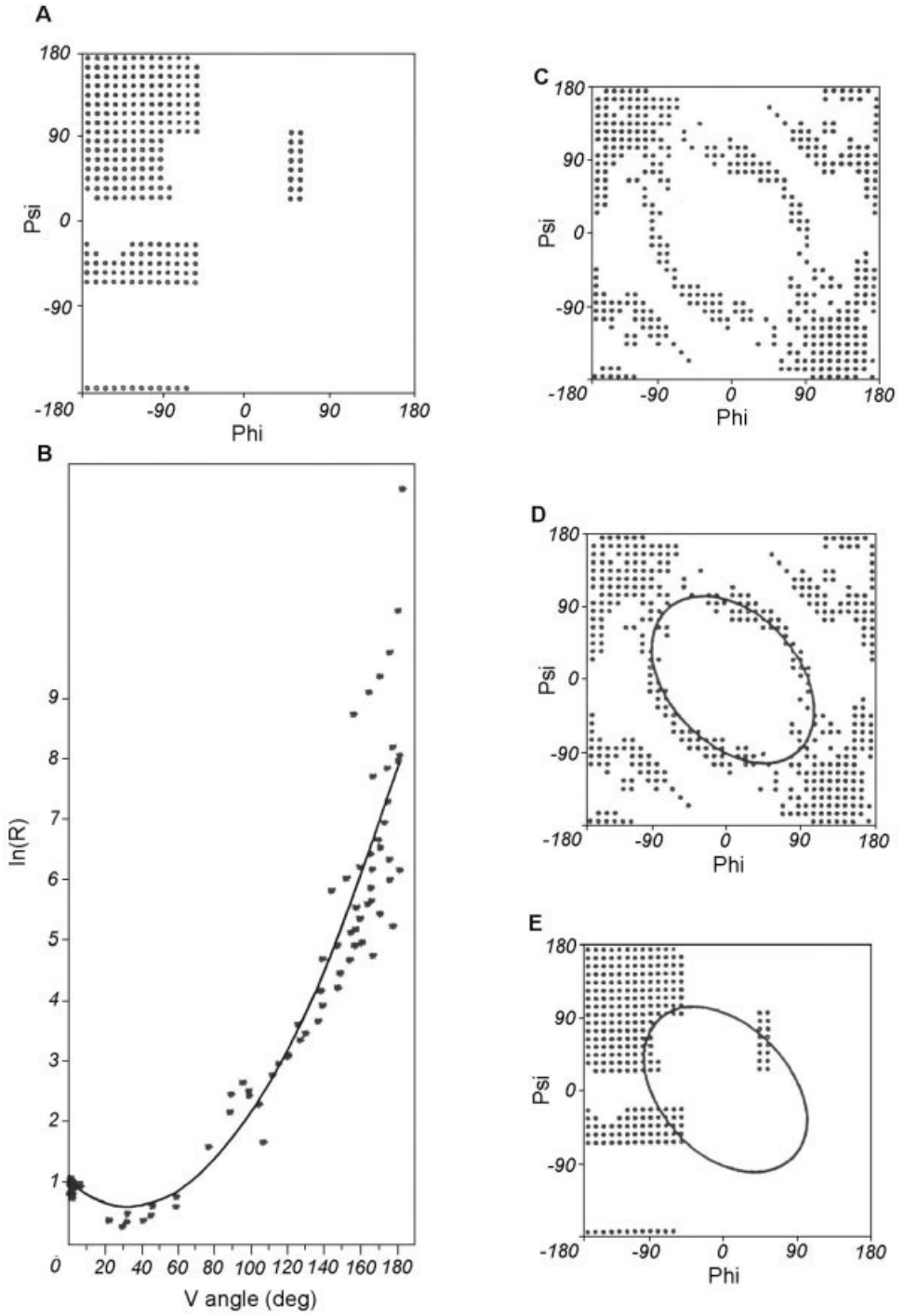

Fig. A2. Ellipse path determination. **A:** ϕ-ψ map with low-energy area distinguished. **B:** ln(R) as a function of V angle for grid points shown in A. **C:** ϕ-ψ map with grid points, where the structure satisfies Eq. 1. **D:** Proposed ellipse path. **E:** Low-energy areas linked by ellipse.