ELSEVIER

Journal of Molecular Graphics and Modelling 21 (2002) 119–128

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

An efficient method for reconstructing protein backbones from α-carbon coordinates

Yoriko Iwata, Atsushi Kasuya, Shuichi Miyamoto*

Exploratory Chemistry Research Laboratories, Sankyo Co. Ltd., Tokyo, Japan

Received 10 February 2002; accepted 28 June 2002

Abstract

We present an approach for building protein backbones from α-carbon (Cα) coordinates. The approach is analytical and based on the information of favored regions in the Ramachandran map. The backbone construction consists of three parts: prediction of (ϕ, ψ) angle pairs from the Cα trace, generation of atomic coordinates with these (ϕ, ψ) angles, and refinement by subsequent energy minimization. Tests on several known protein structures show that the root mean square deviations in reconstructed backbones are 0.25–0.48 Å for coordinates and 14–34° for ϕ and ψ angles. The results indicate that our method is one of the best methods proposed in terms of accuracy. It has also been revealed that the approach is not only robust against errors in Cα coordinates but is also capable of providing equivalent or more reasonable models compared to other known methods. Furthermore, backbone structures were found to be built accurately by using the (ϕ, ψ) angles from a different structure of the same protein. This suggests that the approach could be effective and useful in homology modeling studies.

Keywords: Protein backbone; Backbone construction; Cα coordinate; Main chain dihedral angle; Ramachandran map; Molecular modeling

1. Introduction

Much of the 3D information for proteins and nucleic acids is compiled in the protein data bank (PDB) [1] and the number of entries has been increasing exponentially during the past few years in line with the current pace of X-ray crystallographic technology as well as the advance of the human genome project. For a fraction of important proteins, however, the data entries in the PDB have been limited to only α-carbon (Cα) coordinates. Therefore, several methods for modeling protein backbone structures from Cα coordinates have been investigated in recent years. These approaches can also be used to fit protein structures to X-ray crystallographic data or to generate complete protein structures from lattice representations for tertiary structure prediction. In the interpretation of crystallographic electron density maps, assignment of likely Cα coordinates is an important early step, which is followed by the model building of backbones. Since the main chain is often represented by only Cα coordinates in the lattice simulations, the generation of complete protein structures usually begins with the construction of the main chain coordinates.

Most methods for modeling protein backbones from Cα coordinates are classified into three categories: analytical methods [2–4], fragment search methods [5–9], and molecular dynamics or Monte Carlo simulations [10,11]. One of the early analytical approaches was presented by Purisima and Scheraga more than 15 years ago [12]. This method propagated the backbone reconstruction from an initial guess of a (ϕ, ψ) pair for an inner residue. Unfortunately, the assumed ideal rigid protein geometries led to numerical instabilities and did not yield a very good fit for the real α-carbon trace. Rey and Skolnick determined the position of the β-carbon, first, by using the trigonometric relations among protein coordinates and then rebuilt the main chain structures [2]. They also recently reported a more rapid method [13]. Two other independent groups successfully constructed protein backbones by rotating peptide groups around the virtual axes that link successive Cα atoms. Luo and Tang accepted all peptide plane orientations that satisfy the constraint for the N–Cα–C bond angle. The backbone with the largest portion of (ϕ, ψ) pairs in favored regions of the Ramachandran map was selected [3]. In contrast, the other group, Payne converted the statistical data of (ϕ, ψ) distributions to potentials of mean force based on 61 protein crystal structures [4]. Peptide group orientations that give the minimum value of the semiempirical Hamiltonian function, which includes the potentials of mean force, were chosen. Very

* Corresponding author. Tel.: +81-3-3492-3131; fax: +81-3-5436-8570.
E-mail address: miya@shina.sankyo.co.jp (S. Miyamoto).

1093-3263/02/$ – see front matter
PII: S1093-3263(02)00142-0

recently, Scheraga and coworkers reported the energy-based reconstruction by using a Monte Carlo method [14].

From the pioneering study of Jones and Thirup [15], several investigators have reconstructed protein backbones using a variety of peptide backbone structures. The Cα trace is split into many fragments, and for every one of them a reference database is searched to identify the segment of Cα atoms that have the best overlap with the target fragment. To reduce conformational discontinuities at fragment boundaries, smoothing procedures are usually applied. The rebuilding of flavodoxin using this approach afforded a backbone with an atomic root mean square deviation (RMSD) of 0.51 Å [5]. This fragment search approach sometimes fails due to a lack of fragments in the protein database that happen to have the same geometries as those of the target structure. In both the analytical and fragment search methods, energy minimization using molecular mechanics calculations are generally carried out in the final step.

One example of the last category was reported by Correa, who rebuilt protein backbones by the sequential addition of a Pro, Gly, or Ala residue followed by a molecular dynamics refinement. Their model of flavodoxin, for example, had an RMSD of 0.49 Å [10]. Recently, Mathiowetz and Goddard built a protein main chain one residue at a time from the N-terminal using a dihedral probability grid Monte Carlo method [11].

A different approach for the building of protein backbones from Cα coordinates is taken here. It is an analytic method consisting of three parts: prediction of (ϕ, ψ) pairs from the Cα trace, calculation of backbone atomic coordinates based on the predicted (ϕ, ψ) angles, and refinement by subsequent energy minimization. The prediction of (ϕ, ψ) pairs was carried out by selecting two pairs of (ϕ, ψ) dihedrals that are in favored regions of the Ramachandran map and that satisfy the constraint for two virtual-bond angles (VBA) and one virtual-bond dihedral angle (VBDA) among four consecutive α-carbon atoms (Fig. 1). Our approach is similar to Tang's in that we use the favored regions in the Ramachandran plot, but ours differs in the candidate selection process. Not only two bond angles but also one dihedral angle derived from the four successive α-carbon atoms were included in our selection of two adjacent (ϕ, ψ) pair combination whereas Tang's group chose the peptide plane orientations that fit the restraint of the N–Cα–C bond angle.

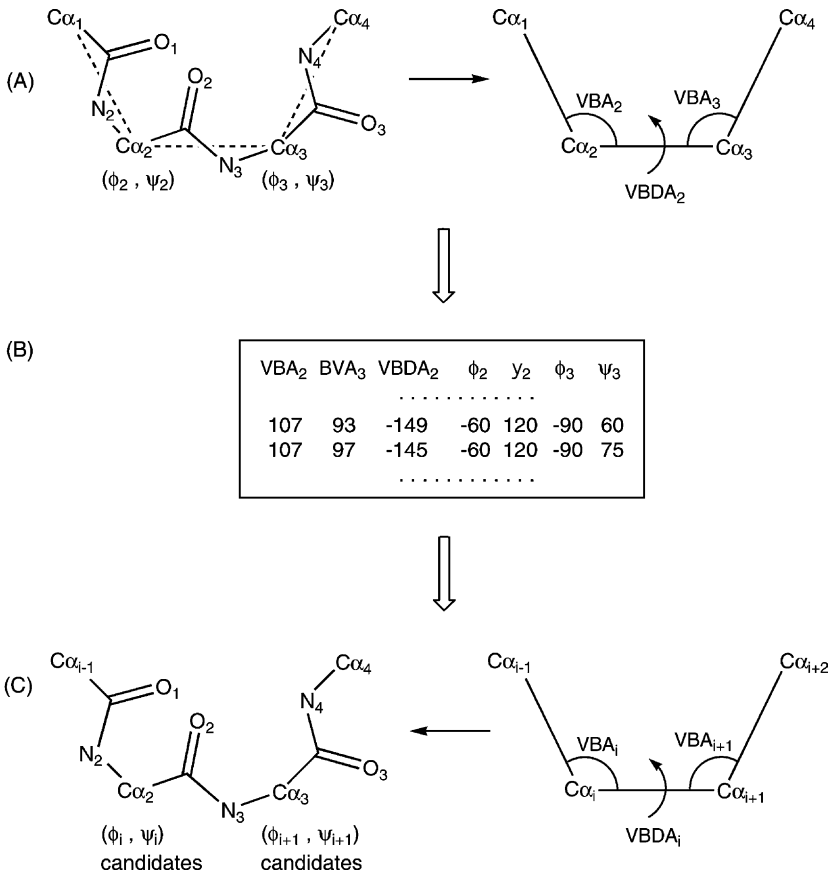

Fig. 1. Schematic drawing of the geometric basis for the (ϕ, ψ) prediction. (A) With respect to three consecutive peptide units with standard geometry, dihedral angles of two central residues ((ϕ_2, ψ_2) and (ϕ_3, ψ_3)) were rotated systematically within the favored regions of the Ramachandran map. Based on the resultant atomic coordinates of four Cα (Cα$_1$–Cα$_4$), VBA$_2$, VBA$_3$, and VBDA$_2$ were calculated. (B) A reference table of the two VBA as well as one VBDA and corresponding two adjacent (ϕ, ψ) angles was generated. (C) With regard to four successive α-carbons in the target Cα trace, VBA$_i$, VBA$_{i+1}$, and VBDA$_i$ were calculated and possible candidates of (ϕ_i, ψ_i) and (ϕ_{i+1}, ψ_{i+1}) were listed based on the corresponding table.

We also examined how our method can be applied in cases where the Cα coordinates contain some errors. We reported on the development of the ReconstrCα program which generates 3D coordinates of α-carbon atoms from a pair of stereographic figures [16]. Thus, the generated coordinates always carry a certain amount of estimation errors. The Cα positions derived from an electron density map also contain some experimental errors. Therefore, it would be useful to evaluate the effect of such errors in Cα coordinates on the constructed backbones. Lastly, it was found that the backbone structure could be built accurately by using the (ϕ, ψ) angles from a different reference structure instead of estimating these values.

In this article, we present a detailed algorithm on how to reconstruct protein backbones from Cα coordinates. The results obtained by applying this method to several protein structures are also shown and these are compared with those provided by other investigators. Application of this method to Cα coordinates containing some errors is then described and finally the results gained by making use of the (ϕ, ψ) angles from a different structure determination are presented.

2. Methods

We adopted an analytic method that was based on the information of favored regions in the Ramachandran map. The calculation flow chart, presented in Fig. 2, is divided into three parts (phases) as mentioned above. Each phase is described below in detail. The standard peptide geometry taken from the parameter of alanine in the ECEPP program [17] was employed and the peptide group was considered to be in a straight plane.

2.1. Phase I: prediction of (ϕ, ψ) angle pair

Step 1. With respect to four consecutive residues with standard geometry, dihedral angles of the two central residues ((ϕ_2, ψ_2) and (ϕ_3, ψ_3)) were rotated systematically at certain intervals within the favored regions of the Ramachandran map (Fig. 1). Based on the resultant atomic coordinates of four Cα (Cα$_1$–Cα$_4$) atoms, the angles VBA$_2$, VBA$_3$, and VBDA$_2$ were calculated and the relationships between these angles and (ϕ, ψ) angles were obtained. A reference table of the two VBA as well as one VBDA and the corresponding two adjacent (ϕ, ψ) angles was then generated. VBA and VBDA were tabulated at certain grid intervals. We employed the favored (ϕ, ψ) regions presented by Wilmot and Thornton [18] and they are shown in Fig. 3. The regions allowed for only Gly were marked as such in the reference table and used for Gly in step 2. The favored regions shown in Fig. 3 were for a *trans* peptide. The (ϕ, ψ) of residues that include or are adjacent to a *cis* peptide, which are easily identified based on Cα–Cα distances, are estimated in phase II.

Step 2. With regard to the four successive α-carbons (Cα$_{i-1}$–Cα$_{i+2}$) in the target Cα trace, VBA$_i$, VBA$_{i+1}$, and VBDA$_i$ were calculated and the possible combinations of (ϕ_i, ψ_i) and (ϕ_{i+1}, ψ_{i+1}) were listed based on the reference table generated above (Fig. 1). In the case of Pro, a set of dihedral pair angles with ϕ angles which were significantly deviated (>30°) from the canonical 68°, were eliminated. Multiple (ϕ, ψ) pairs generally correspond to a given set of two VBA and one VBDA. By repeating this process from the N-terminal straight through to the C-terminal, two kinds of (ϕ, ψ) candidates were obtained for each residue except for the terminal residues. Each of the two kinds was obtained from the combination with the preceeding or succeeding residue. Since the correct (ϕ, ψ) pair should be included in both series of candidate lists, the common (ϕ, ψ) pairs were retained and the rest were discarded. At this point, the (ϕ_i, ψ_i) sets, i.e. (ϕ_{i-1}, ψ_{i-1}) and (ϕ_{i+1}, ψ_{i+1}) that were paired with the discarded (ϕ_i, ψ_i) were also eliminated. For each residue, two series of (ϕ, ψ) candidates were again examined to keep only the common ones. This elimination process was repeated until no more candidates could be eliminated.

Step 3. Through the previous step, many (ϕ, ψ) candidates generally remained for each residue and these candidates were found clustered in the Ramachandran plot. In this step, a representative (ϕ, ψ) angle pair for each cluster was determined as the revised (ϕ, ψ) candidate as follows. If the difference in ϕ and ψ angles of the original (ϕ, ψ) candidates was within one grid spacing, these candidates were automatically regarded to be in the same cluster. The representative (ϕ, ψ) pair at the weight center of the cluster members was then determined by taking the periodicity of the angles into account. Most cases have only one cluster, and thus, a single representative (ϕ, ψ) candidate was obtained for each residue. When more than one representative (ϕ, ψ) pair (multiple candidates) were obtained, one was selected in step 4. In the case where no candidate remained, the most probable (ϕ, ψ) pair was estimated in phase II.

Step 4. The validity of the representative (ϕ, ψ) candidates was examined in terms of the Cα trace. Given the representative (ϕ_i, ψ_i) candidate of Cα$_i$, orientations of the peptides for Cα$_{i-1}$–Cα$_i$ and Cα$_i$–Cα$_{i+1}$ were calculated (Fig. 4A). By repeating this calculation from the N-terminal to the C-terminal, two estimated orientations of each peptide unit (Cα$_i$–Cα$_{i+1}$) were generally obtained from the (ϕ_I, ψ_i) and (ϕ_{i+1}, ψ_{i+1}) pairs. The conversion of the peptide plane orientations to atomic coordinates was performed on a residue basis with residues in their standard geometry as shown in Fig. 4B. The positional deviations of the doubly generated oxygen atoms were calculated. The deviations of the N–Cα–C bond angle ($A_{\mathrm{N\alpha C}}$) from the standard (109.3°) were also computed based on the average atomic positions derived from the two peptide orientations. If the difference

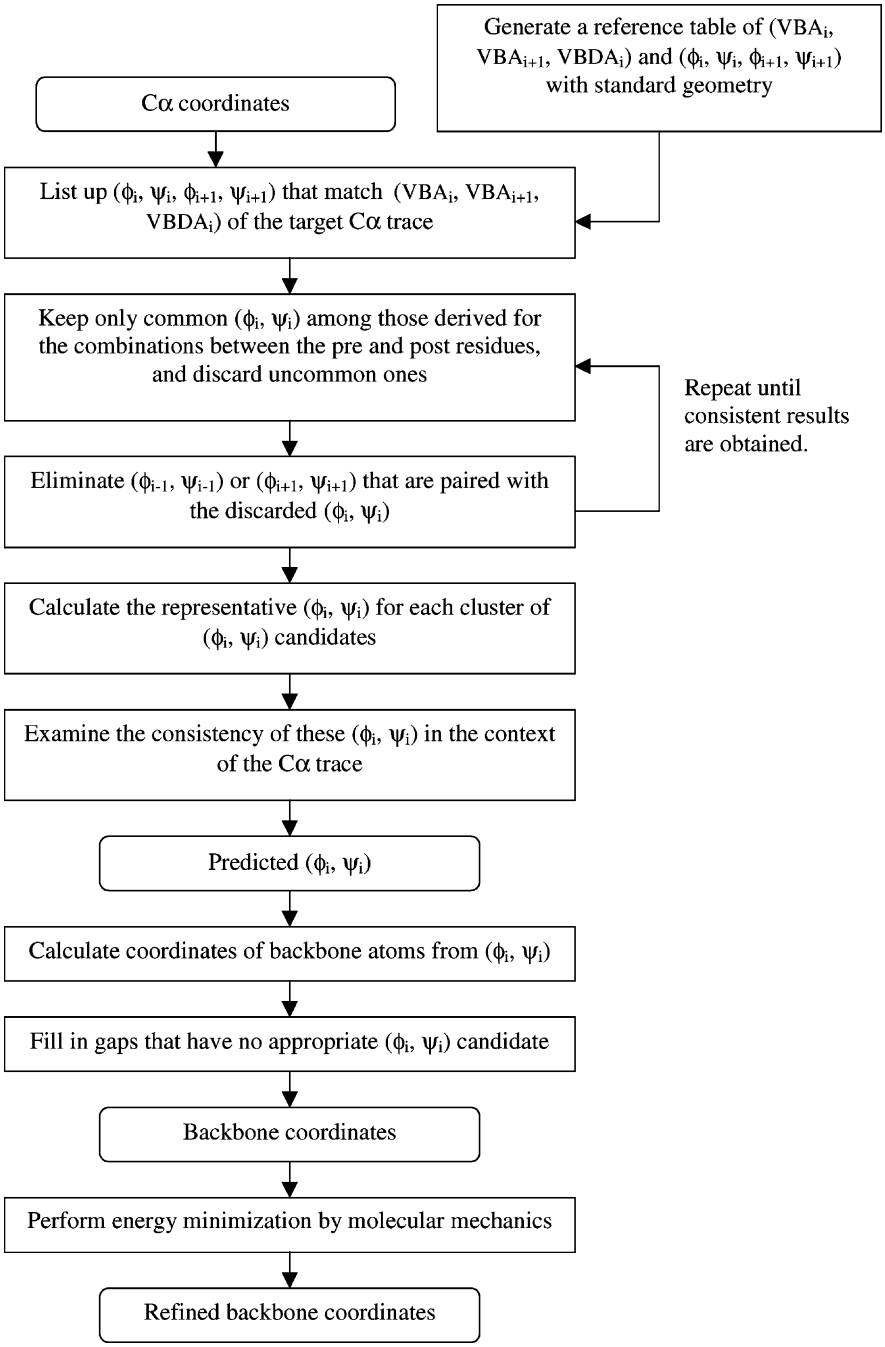

Fig. 2. The flow chart of the backbone reconstruction.

between the two O_i was less than 1.0 Å and the average deviation from ideal of N–Cα_i–C and N–Cα_{i+1}–C bond angles was less than 5°, the combination of (ϕ_i, ψ_i) and (ϕ_{i+1}, ψ_{i+1}) candidates was regarded as being consistent. When multiple candidates were obtained for Cα_i and/or Cα_{i+1}, all combinations were evaluated and the most probable (ϕ, ψ) candidate was chosen manually based on the deviations. By repeating this examination along the Cα trace, one appropriate (ϕ, ψ) pair could generally be determined for each residue except for the terminal residues.

2.2. Phase II: calculation of backbone atomic coordinates from the predicted (ϕ, ψ) angles

Step 5. In this step, most of the calculations in the previous step were repeated. With respect to the four α-carbons, two estimated orientations of the central peptide plane (Cα_i–Cα_{i+1}) were calculated one from the predicted (ϕ_i, ψ_i) and one from the (ϕ_{i+1}, ψ_{i+1}) angles. Taking the middle of the two orientations as the estimated conformation, the coordinates of the backbone atoms (N, C, and O)

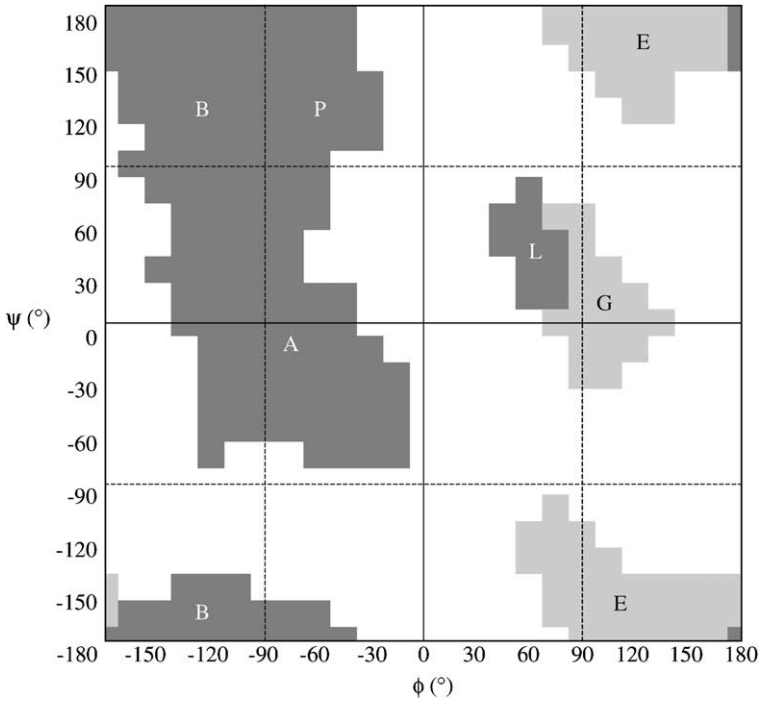

Fig. 3. The favored regions in the Ramachandran map. These regions were originally presented by Wilmot and Thornton [18] and used in this study with a grid spacing of 15°. The lightly shaded regions are allowed only for Gly. Secondary structure designations are as follows: A, α_R; B, β_E; P, β_P; L, α_L; G, γ; E, ε.

were calculated. This calculation was repeated from the N-terminal straight through to the C-terminal.

Step 6. With regard to the residues for which no appropriate (ϕ, ψ) candidate was obtained in the previous phase, they were constructed one by one from the neighboring residue whose atomic coordinates had been built. By rotating a peptide group around the virtual axis, the (ϕ, ψ) angle and N–Cα–C bond angle were evaluated to select the best peptide orientation. In other words, the orientation with the smallest deviation of N–Cα–C bond angle from the standard and with the (ϕ, ψ) angle in favored regions of the Ramachandran plot was chosen. With respect to both terminals, the ψ_1 angle of the N-terminal and ϕ_n angle of the C-terminal were set to 180° by default. In this step, all backbone coordinates were calculated including both ends.

2.3. Phase III: energy minimization with molecular mechanics

Step 7. The backbone structure obtained in the previous phase was subjected to 300 iterations of energy minimization using the Adopted-Basis Newton Raphson algorithm [19] with a distance-dependent dielectric constant of 4*r*. The Cα positions were fixed in the original position, if they had been taken from the PDB. A positional constraint of 10 kcal/(mol Å^2) was applied to Cα if errors had been introduced from the original coordinates. The CHARMm force field [20] implemented in the QUANTA system [21] was used for the molecular mechanics calculations with a united-atom representation.

3. Results and discussion

3.1. Application to five proteins

So far, several investigators have applied their various methods to proteins such as pancreatic trypsin inhibitor (4PTI [22]), carboxypeptidase A (5CPA [23]), flavodoxin (5NLL [24], 3FXN [25]), citrate synthase (2CTS [26]), and triose phosphate isomerase (1TIM [27]) for tests on backbone construction from Cα coordinates. We have therefore applied our method to these five proteins and Table 1 shows the deviations found in the resulting coordinates compared with the reference crystal structures. With respect to the final constructed models, the coordinate RMSD of the backbone structures (CRMSD) were 0.31–0.48 Å and the dihedral RMSD of the ϕ and ψ angles (DRMSD) were 19–34°, indicating accurate model constructions. The superimposition of the constructed and crystal backbone structures of 4PTI is shown in Fig. 5, which indicates that the reconstruction was satisfactory. No peptide flip (a peptide unit rotated by more than 90° from its actual position in the crystal structure) was observed in 4PTI. As can be seen from Table 1, energy minimization with molecular mechanics improved

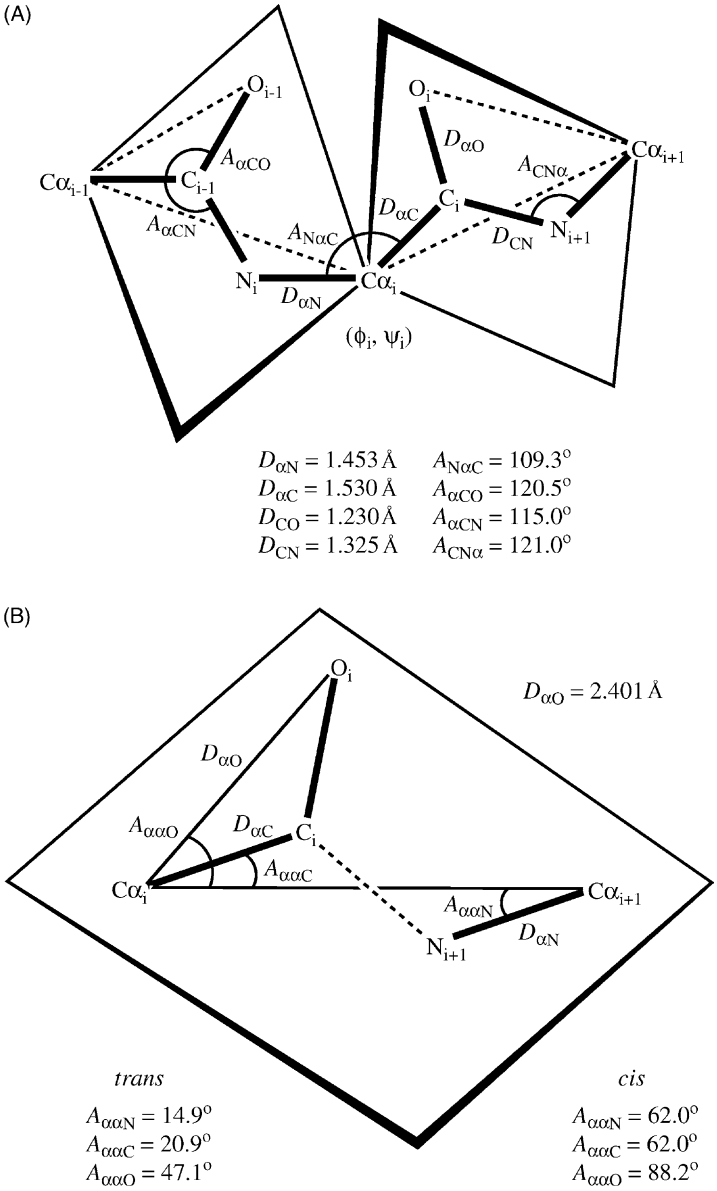

Fig. 4. Schematic representation of the geometrical basis and parameters employed in this study. The distance and bond angle parameters shown above are taken or derived from the standard planar peptide geometry in the ECEPP program [16]. (A) Calculation of peptide orientations of two adjacent peptides. Given one specific (ϕ_i, ψ_i) pair and the standard peptide geometry, the atomic coordinates of two adjacent peptide units are generated. The orientations of the two peptide planes can be defined by the torsional angles $C\alpha_{i-1}$–$C\alpha_i$–$C\alpha_{i+1}$–O_i and $C\alpha_{i+1}$–$C\alpha_i$–$C\alpha_{i-1}$–O_i. These torsional angles are regarded to correspond to those of the target structure with the (ϕ_i, ψ_i) pair in question. (B) Generation of backbone atomic coordinates. After the peptide plane orientations are estimated, the atomic coordinates are calculated on a residue basis by superposing the Cα–Cα virtual-bond directions of the target structure and the standard peptide unit. O_i is placed on the estimated peptide plane with distance, $D_{\alpha O}$, and bond angle, $A_{\alpha\alpha O}$. Similarly, N_i and C_i are located based on the $C\alpha_i$ position. Since the distance and bond angle parameters shown above are fixed, fluctuations in the Cα–Cα distances of the target structure were adjusted by the deviation of the amide bond geometries, i.e. the linkage of residues.

the CRMSD from 0.39–0.54 to 0.31–0.48 Å. The DRMSD of the final models was also improved compared to those of the initially predicted dihedral angles (26–37°). Naturally, as the estimation of ϕ and ψ angles was improved, a more accurate model was constructed.

Only ca. 4% of the total residues including those involved in *cis* peptides had no (ϕ, ψ) candidate by step 3, indicating the high effectiveness of our method. With respect to residues with multiple (ϕ, ψ) candidates, they were often found to be grouped within the sequence. As shown in Fig. 6, the number of consistent candidate combinations (paths) for the grouped residues was rather small and, in most cases, the most appropriate candidate combination was selected by comparing the deviations of dual O positions from each other and N–Cα–C bond angles from ideal. As for 1TIM or 2CTS, one or two groups of residues had two comparable

Table 1
RMS deviations between analytically reconstructed backbone structures and crystal structures

Protein	PDB code[a]	Number of residues	Resolution (Å)	Number of res. with unass. (ϕ, ψ)[b]	Number of models[c]	RMSDs			
						(ϕ, ψ) (°)		Coordinates (Å)	
						Phase I	Phase III	Phase II	Phase III
Pancreatic trypsin inhibitor	4PTI	58	1.5	2	1	26	19	0.39	0.31
Carboxypeptidase A	5CPA	307	1.54	10	1	27	22	0.42	0.31
Flavodoxin	5NLL	138	1.9	5	1	35	28	0.47	0.39
Citrate synthase	2CTS	437	2.0	10	4	27–28	22–23	0.41–0.42	0.32–0.34
Triose phosphate isomerase	1TIM	249	2.5	10	2	35–37	32–34	0.53–0.54	0.46–0.48

[a] Protein code in the PDB.
[b] Number of residues whose (ϕ, ψ) was not assigned in phase I.
[c] Number of reconstructed backbone models.

paths with respect to deviations, respectively. Two models from the two comparable paths were built for 1TIM and four (2×2) models were constructed from the combinations of the two comparable paths for 2CTS. The results of all of these models are included in Table 1. As can be seen in Table 1, the RMSD of these models was similar, indicating that the overall accuracy was not dependent on the selected path.

There are several variables that affect the efficiency of this method. These include grid spacing or resolution of the ϕ, ψ, VBA, and VBDA angles that are used for generating the reference table in step 1. We carried out test calculations with grid intervals of 10 and 15° for both ϕ and ψ angles, 5 and 7° for VBA, and 8 and 10° for VBDA by using 4PTI and 5NLL as models. Comparable results were obtained for both grid intervals of (ϕ, ψ) and VBDA. In contrast, it was found that the 7° interval of VBA gave more accurate models than that of 5°. Although the number of (ϕ, ψ) candidates selected in step 2 varied depending on these variables, the final predicted (ϕ, ψ) values were rather consistent since the candidates were selected from the center of the cluster in step 3. The data shown in this article were obtained with a grid spacing of 15, 7, and 10° for (ϕ, ψ), VBD, and VBDA, respectively.

The thresholds of deviations for validating reasonable combinations of two successive (ϕ, ψ) dihedrals in step 4 are also variables. The current deviation thresholds of 1.0 Å and 5° for the oxygen atom position and the bond angle, respectively, were derived from the test calculations on 4PTI and 5NLL. They might be somewhat loose but we considered that they were acceptable since their primary use is to exclude inconsistent combinations of (ϕ, ψ).

Table 2 represents the comparison of our results with those obtained by other investigators. Our method gave the most accurate models except for 1TIM, which has substantially poorer resolution than the other structures. We have therefore applied our approach to a new crystal structure (1TPH [28]) of triose phosphate isomerase. The 1TPH was also from chicken muscle but determined at 1.8 Å resolution. Both CRMSD and DRMSD of the newly reconstructed model of the protein were highly improved and were 0.33 Å and 23°, respectively. Furthermore, the reconstructed backbone from 1TIM matched the improved 1TPH structure slightly better than 1TIM matches 1TPH. Table 2 also indicates that even for the structures obtained before energy minimization, our RMSD were rather small compared to those of previous studies with respect to 5CPA and 2CTS. It is, therefore, inferred that our method is one of the best methods proposed in terms of accuracy.

3.2. Application to structures with errors in Cα coordinates

We reported on the development of the program ReconstCα, which generates 3D coordinates of α-carbon atoms from a pair of stereographic figures, and described the results of this application [15]. The current procedure calculates 3D coordinates of the backbone structure from 3D coordinates of α-carbon atoms while our previously

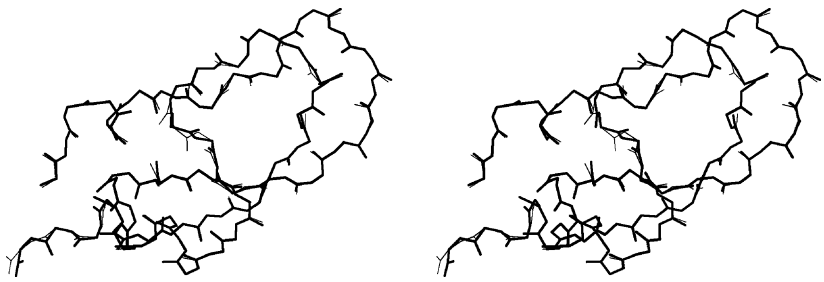

Fig. 5. A stereogram of the superimposition of the constructed (thin) and original crystal (thick) backbone structures of 4PTI.

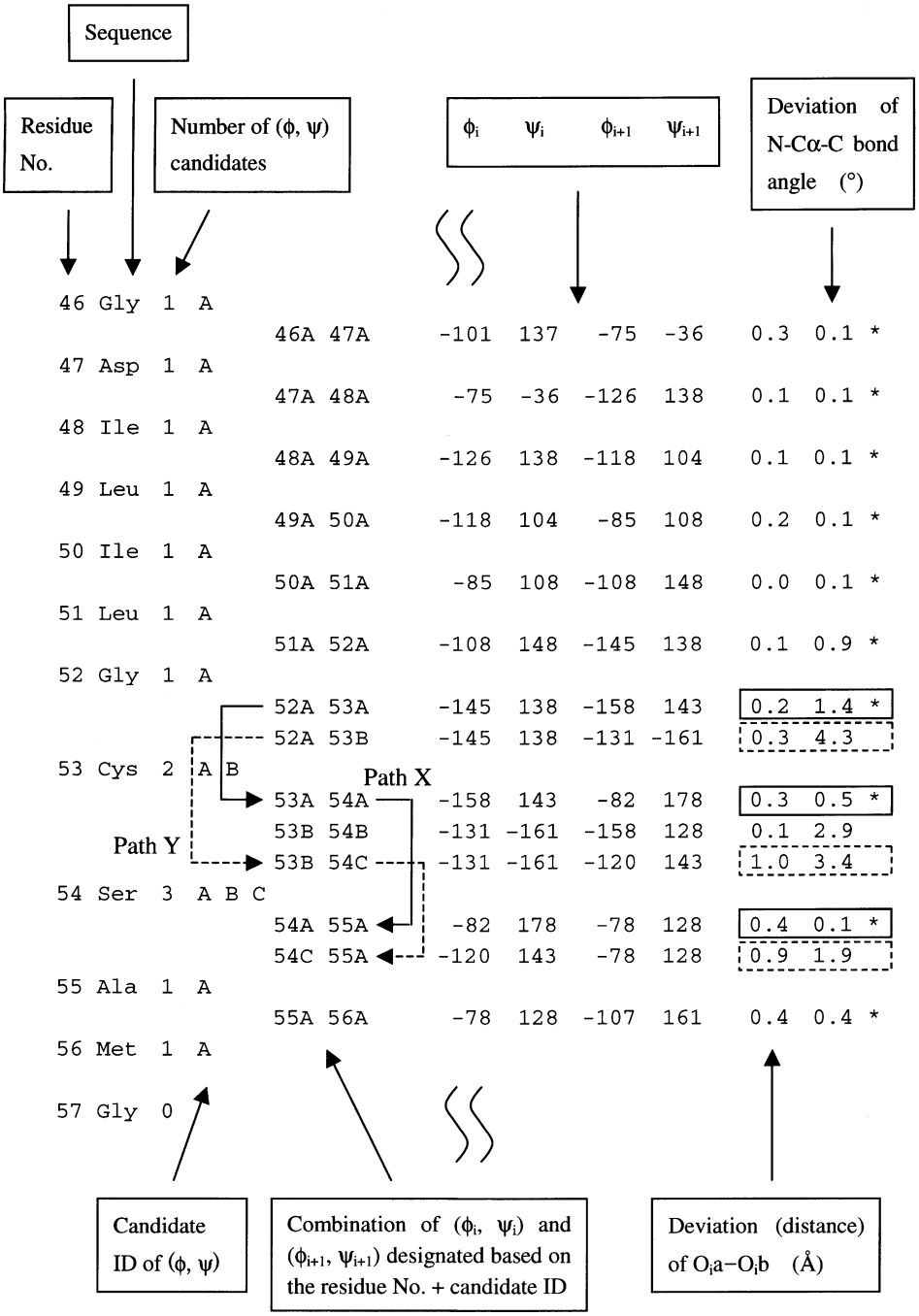

Fig. 6. An example of the output produced in step 4 for 5NLL. The output shows the consistent combinations of (ϕ, ψ) candidates of two consecutive residues. The residues with multiple (ϕ, ψ) candidates are often clustered in sequence. Here, two residues 53 and 54 have multiple candidates. Although six combinations of (ϕ, ψ) candidates are possible for those two residues, only two of them are shown to be consistent. Furthermore, path X is considered to be more preferable compared to path Y due to the smaller deviations between the two estimates of oxygen positions (O_ia–O_ib) and smaller non-ideality of N–Cα–C bond angle. In most cases the paths do not form a tree structure but a simple linear structure and the most appropriate path (combination) can be chosen based on the deviations as shown here.

published approach estimates 3D coordinates of α-carbon atoms from their stereographic 2D coordinates. The present method may be applied to such rebuilt Cα coordinates, which contain a certain amount of errors. The Cα trace generated in the early stage of the X-ray crystallographic analysis also includes some uncertainty. Therefore, it would be useful to know how such errors in Cα coordinates affect the backbone construction. For this purpose, the reconstructed Cα coordinates of 1SN3 [29] with an RMSD of 0.36 Å were examined in one case. As another test, Cα positions in 4PTI have been modified by applying random shifts of up to 0.5 Å to the *x*, *y*, and *z* coordinates independently according to the

Table 2
Comparison of RMSD between our approach and other methods (1–7)

Protein	RMSD (Å) for eight different methods[a]										
	1 (Skolnick)		2 (Tang)		3 (Payne)	4 (Wodak)	5 (Sander)	6 (Mandal)	7 (Goddard)	This study	
	Before EM[b]	After EM[b]	Before EM[b]	After EM[b]						Before EM[b]	After EM[b]
4PTI	0.63	0.43	0.36	0.36	0.32[c]	–	–	–	0.61	0.39	**0.31**
5CPA	–	–	–	0.51	0.33	0.64	0.48	0.44	–	0.42	**0.31**
5NLL[d]	0.71[e]	0.50[e]	–	–	**0.39**[f]	–	0.48[e]	0.46[e]	0.59[e]	0.47	**0.39**
2CTS	–	–	–	0.48	–	0.56	0.45	0.43	–	0.41–0.42	**0.32–0.34**
1TIM	0.63	**0.39**	0.46	0.44	0.50	0.63	0.59	0.53	–	0.53–0.54	0.46–0.48

[a] The figures in bold represent the smallest difference obtained among the methods for each protein. Other seven methods presented are as follows: 1, [2]; 2, [3]; 3, [4]; 4, [6]; 5, [7]; 6, [9]; 7, [11].
[b] Energy minimization.
[c] 6PTI was used.
[d] 5NLL was used in this study because the corresponding former file 3FXN, on which previous studies were done, was no longer registered in the current PDB database.
[e] 3FXN was used.
[f] 4FXN was used.

previous studies [3,6]. The shifts were calculated from the following expression: shift = random − 0.5, where shift is the shift applied in Å, random is a uniformly random real number between 0 and 1. Three sets of modified Cα traces were prepared by changing the random seed.

The CRMSD of the generated structure from the reconstructed Cα coordinates (1SN3) was 0.50 Å while that rebuilt from the original Cα coordinates in the PDB entry was 0.25 Å (DRMSD = 14°). The CRMSD of the reconstructed backbones from the shifted Cα coordinates (4PTI) were 0.46–0.63 Å. These CRMSD derived from the deformed Cα trace were therefore only somewhat larger than the regular CRMSD (0.31 Å) obtained from the canonical Cα coordinates and the initial RMSD (∼0.41 Å) of the shifted Cα coordinates. With respect to 4PTI, Luo and Tang performed the same test and reported a CRMSD of 0.64–0.84 Å [3], which is larger than our results. Payne also carried out a similar test and pointed out that the increase in CRMSD approximately corresponded to the errors introduced to the guiding Cα positions [4]. Thus, our approach is not only relatively robust against errors in Cα coordinates but is also capable of providing equivalent or more accurate models compared to other known methods. Moreover, the current method together with the ReconstCα program provides a way to reconstruct backbone coordinates from a pair of stereographic figures of the Cα trace.

3.3. Reconstruction of backbones with (ϕ, ψ) angles of reference structures

This is the first method which calculates backbone atomic coordinates from predicted (ϕ, ψ) angles and Cα coordinates to the best of our knowledge. We have initially confirmed that accurate backbone structures with a CRMSD of 0.14 and 0.06 Å were constructed for 5CPA and 5NLL, respectively, if the original (ϕ, ψ) angles are given. The reason that the rebuilt coordinates do not completely coincide with the original coordinates may be due to variation of bond lengths and angles, either in the model or in the real protein backbone structure. As can be seen in Table 1, the reconstruction of the backbones with a CRMSD of around 0.3 Å can be achieved, if the (ϕ, ψ) angles are estimated with an RMSD of ca. 30°. We therefore considered that if a different crystal structure of the target protein is available, the (ϕ, ψ) angles of this structure can be used instead of estimating those values.

In order to examine the efficacy in using a different reference structure, backbones of 5CPA and 5NLL were reconstructed with the (ϕ, ψ) angles of 1F57 [30] and 2FOX [24], respectively, which are different crystal structures of the same proteins, i.e. carboxypeptidase A and flavodoxin, respectively, but with some conformational changes induced by ligand binding or oxidation. The RMSD of Cα between the target and reference structures was 0.36 Å for carboxypeptidase A and that of flavodoxin was 0.25 Å. The DRMSD between the target and reference structures was 18° for carboxypeptidase A and that of flavodoxin was 15° although, normally, the DRMSD is not known beforehand. The CRMSD and DRMSD of the constructed model of 5CPA were 0.31 Å and 21°, respectively, which were comparable to those obtained by our approach described in the beginning. The CRMSD and DRMSD of the rebuilt model of 5NLL were 0.23 Å and 16°, respectively, which were better than those (0.39 Å and 28°) obtained by the regular approach, indicating that the use of the reference structure was effective. No large deviation of the constructed structure of 5NLL from the target was observed as well. To what extent the approach with reference structures is effective depends on the similarity between the target and reference structures. As far as it can be seen in the above example, it would be more advisable to use a reference structure if the RMSD of Cα is less than 0.4 Å.

In this post genome era, homology modeling studies will be actively carried out. In the case of lower levels of homology, the direct transfer of main chain coordinates may be

difficult, but one can still apply the Cα trace from the reference structures. The above results suggest that our approach might be effective in such a case.

In summary, the three kinds of application results presented above clearly indicate the effectiveness of the approach. Nonetheless, our approach could be further refined. For instance, (ϕ, ψ) angles of residues involved in *cis* peptides could be estimated in phase I if the Ramachandran map for the *cis* peptide is used in combination with that for *trans*. An automatic selection of the most appropriate (ϕ, ψ) pair among multiple candidates is also preferable. These are the tasks that we are currently working on.

4. Conclusion

We have developed a simple, but effective, approach for building protein backbones from Cα coordinates. This analytical approach is based on the information of favored regions in the Ramachandran map. Tests on six known protein structures including 1SN3 show that the CRMSD are within 0.25–0.48 Å and the DRMSD are within 14–34°. In terms of accuracy, these results indicate that our method is one of the best methods among those proposed for the same goal. It has also been revealed that the approach can be reasonably applied to the cases in which Cα coordinates contain some errors. Together with the ReconstCα program, the current method provides the way to reconstruct backbone coordinates from a pair of stereographic Cα trace. Lastly, backbone structures can be built by using the (ϕ, ψ) angles of reference structures such as those from different crystal structures of the same protein.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[2] A. Rey, J. Skolnick, Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates, J. Comput. Chem. 13 (1992) 443–456.
[3] Y. Luo, Y. Tang, Building protein backbones from Cα coordinates, Protein Eng. 5 (1992) 147–150.
[4] P.W. Payne, Reconstruction of protein conformation from estimated positions of the Cα coordinates, Protein Sci. 2 (1993) 315–324.
[5] L.S. Reid, J.M. Thornton, Rebuilding flavodoxin from Cα coordinates: a test study, Proteins 5 (1989) 170–182.
[6] M. Claessens, E.V. Custem, I. Lasters, S. Wodak, Modelling the polypeptide backbone with spare parts from known protein structures, Protein Eng. 2 (1989) 335–345.
[7] L. Holm, C. Sander, Database algorithm for generating protein backbone and side-chain coordinates from a Cα trace application to model building and detection of coordinate errors, J. Mol. Biol. 218 (1991) 183–194.
[8] M. Levitt, Accurate modeling of protein conformation by automatic segment matching, J. Mol. Biol. 226 (1992) 507–533.
[9] C. Mandal, D.S. Linthicum, PROGEN: an automated modelling algorithm for the generation of complete protein structures from the α-carbon atomic coordinates, J. Comput.-Aided Mol. Design 7 (1993) 199–224.
[10] P. Correa, The building of protein structures from α-carbon coordinates, Proteins 7 (1990) 366–377.
[11] A.M. Mathiowetz, W.A. Goddard III, Building proteins from Cα coordinates using the dihedral probability grid Monte Carlo method, Protein Sci. 4 (1995) 1217–1232.
[12] E. Purisima, H.A. Scheraga, Conversion from a virtual-bond chain to a complete polypeptide backbone chain, Biopolymers 23 (1984) 1207–1224.
[13] M. Milik, A. Kolinski, J. Skolnick, Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates, J. Comput. Chem. 18 (1997) 80–85.
[14] R. Kamierkiewicz, A. Liwo, H.A. Scheraga, Energy-based reconstruction of a protein backbone from its α-carbon trace by a Monte-Carlo method, J. Comput. Chem. 23 (2002) 715–723.
[15] T.A. Jones, S. Thirup, Using known substructures in protein model building and crystallography, EMBO J. 5 (1986) 819–822.
[16] Y. Iwata, A. Kasuya, S. Miyamoto, Reconstruction of the 3D coordinates of α-carbon atoms of proteins from a pair of stereographic figures, J. Comput.-Aided Mol. Design 10 (1996) 558–566.
[17] M.J. Browman, L.M. Carruthers, K.L. Kashuba, F.A. Momany, M.S. Pottle, S.P. Rosen, S.M. Rumsey, QCPE 11 (1975) 286.
[18] C.M. Wilmot, J.M. Thornton, β-Turns and their distortions: a proposed new nomenclature, Protein Eng. 3 (1990) 479–493.
[19] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, CHARMm: a program for macromolecular energy, minimization, and dynamics calculations, J. Comput. Chem. 4 (1983) 187–217.
[20] CHARMm, Version 23, Accelerys Inc., San Diego, CA, http://www.accelerys.com.
[21] QUANTA, Version 98, Accelerys Inc., San Diego, CA, http://www.accelerys.com.
[22] M. Marquart, J. Walter, J. Deisenhofer, W. Bode, R. Huber, The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors, Acta Cryst. Sect. B 39 (1983) 480–490.
[23] D.C. Reeds, M. Lewis, W.N. Lipscomb, Refined crystal structure of carboxypeptidase A at 1.54 Å resolution, J. Mol. Biol. 168 (1983) 367–387.
[24] M.L. Ludwig, K.A. Pattridge, A.L. Metzger, M.M. Dixon, M. Eren, Y. Feng, R.P. Swenson, Control of oxidation-reduction potentials in flavodoxin from *Clostridium beijerinckii*: the role of conformation changes, Biochemistry 36 (1977) 1259–1280.
[25] W.W. Smith, R.M. Burnett, G.D. Darling, M.L. Ludwig, Structure of the semiquinone form of flavodoxin from *Clostridium* MP. Extension of 1.8 Å and some comparisons with the oxidized state, J. Mol. Biol. 117 (1997) 195–226.
[26] S. Remington, G. Wiegand, R. Huber, Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 Å resolution, J. Mol. Biol. 158 (1982) 111–152.
[27] D.W. Banner, A.C. Bloomer, G.A. Petsko, D.C. Phillips, I.A. Wilson, Atomic coordinates for triose phosphate isomerase from chicken muscle, Biochem. Biophys. Res. Commun. 72 (1976) 146–155.
[28] Z. Zhang, S. Sugio, E.A. Komives, K.D. Liu, J.R. Knowles, G.A. Petsko, D. Ringe, Crystal structure of recombinant chicken triosephosphate isomerase-phosphoglycolohydroxamate complex at 1.8 Å resolution, Biochemistry 33 (1994) 2830–2837.
[29] R.J. Almassy, J.C. Fontecilla-Camps, F.L. Suddath, C.E. Bugg, Structure of variant-3 scorpion neurotoxin from *Centruroides sculpturatus* ewing, refined at 1.8 Å resolution, J. Mol. Biol. 170 (1983) 497–527.
[30] D.M. Van Aalten, C.R. Chong, L. Joshua-Tor, Crystal structure of carboxypeptidase A complexed with D-cystein at 1.75 Å—inhibitor-induced conformational changes, Biochemistry 39 (2000) 10082–10089.