Sangyoon Lee

School of Mechanical and Aerospace Engineering
Konkuk University, Seoul, Korea

Gregory S. Chirikjian

Department of Mechanical Engineering
Johns Hopkins University
Baltimore, MD 21218, USA
gregc@jhu.edu

Pose Analysis of Alpha-Carbons in Proteins

Abstract

In this paper we present a novel method to describe the pose (position and orientation) distribution of amino acid residue pairs within a protein, which are proximal in space and distal in sequence. While the Ramachandran plot provides information of protein conformations using the ϕ and ψ angles between sequentially proximal residues, our method can offer six-dimensional relative pose information. Distribution data are visualized in the form of continuous distributions by using Gaussian distribution functions on $SO(3)$ and $\mathbb{R}^3$. Hence, we discuss how the classical Gaussian functions can be generalized to capture both positional and orientational data. The method is applied to 168 protein structures in the Protein Data Bank and results are discussed.

KEY WORDS—interaction between residues, 6D relative pose, protein data, data visualization, Gaussian function, axis-angle representation, computational tool, continuous distribution

1. Introduction

More than 30 years ago, Ramachandran and Sasisekharan (1968) showed that a sequence of amino acids comprising a protein must have certain geometries, which do not allow certain relative positions and orientations between sequentially adjacent pairs. In this formulation, the allowed/disallowed regions are represented in the *phi–psi* (ϕ–ψ) plane. (See Figure 2 for the graphical definition of ϕ, ψ angles.) In this paper, we examine a related research issue: given a protein, we first affix a frame of reference to the alpha-carbon atom (C_α) of each amino acid in the structure. Then we record all possible positions and orientations between amino acids that are proximal in space and distal in sequence, i.e., within certain spatial/sequential distance cutoffs. Hence, in essence we seek a six-dimensional Ramachandran-like plot for sequentially distant residue pairs.

There have been several studies on backbone–backbone, backbone–side-chain and side-chain–side-chain interactions in protein structures. Bahar and Jernigan (1996) studied the statistical distribution of interactions between residues in polypeptides and presented the existence of preferred distributions for a given residue type.

Banavar, Maritan, and Seno (2002) showed that the distribution of relative orientations of amino acids exhibits peaks at specific angles. The relative orientation is represented by the angle between two vectors, each of which joins next-nearest-neighbor C_α atoms along the polypeptide chain. Therefore, the vector for the ith amino acid, $C(i)$, connects $C(i-1)$ and $C(i+1)$.

In three recent papers, Buchete, Straub, and Thirumalai (2004a, 2004b, 2004c) describe orientational potentials for protein simulations. They studied three types of interactions (side-chain–side-chain, side-chain–backbone, and backbone–backbone) with local reference frames of side chains and a virtual interaction center on the backbone in the middle of the peptide link.

The significant difference between our approach and those previous related works is that we examine the three-dimensional rotational data of the rigid-body displacement relating the two local reference frames in addition to three-dimensional positional (distance and direction) data in space. We therefore provides full six-dimensional probability densities, whereas others have focused on lower-dimensional marginal densities.

Statistical probabilities using geometrical information of orientation, position, or distance in polypeptide chains could be a useful tool to develop efficient computational methods for protein fold recognition and protein structure prediction, and also for simulations of coarse-grained models of proteins.

Our statistical analysis of pose (position and orientation) data of polypeptides also may be helpful for modeling protein

The International Journal of Robotics Research
Vol. 24, No. 2–3, February–March 2005, pp. 183-210,
DOI: 10.1177/0278364905050353

structures. A relevant work by Kemp and Chen (1998) presents worm-like polymer chains which model the low-temperature protein structures. The worm-like polymer chains are used to reproduce a helix ground state (coil–helix transition). The paper discusses three parameters to measure the degree of helicity within the chain.

Trovato, Ferkinghoff-Borg, and Jensen (2003) proposed a model for a protein with two different interactions that mimic the hydrophobic effect and the angular dependence of hydrogen bonding. The results in this paper could provide a guideline for generating new models of polypeptide chains. Pose information can be extracted from new models and compared with the results in our paper.

In our analysis, the pose data appear like a cloud in the group of three-dimensional rigid-body motions, and we would like to visualize this cloud in such a way that relative pose relations can be understood clearly. In order to achieve this, we plot "two-dimensional slices" of relative position data and other slices of orientation data. Any "holes" in these plots represent poses that one amino acid does not attain relative to its neighbors. As a result, plots like Ramachandran's ϕ–ψ plot are formed. Now, however, the data are in a higher dimension than the two-dimensional ϕ–ψ plane, and the data are for sequentially distant yet spatially proximal residues rather than sequentially proximal residues. In order to apply this study broadly, a large amount of data should be taken from various proteins in the Protein Data Bank (PDB; Berman et al. 2000). Hence, interpreting the large amount of data is a significant problem.

When presented with a large set of point data, there are two issues related to smoothing or filtering of the original data. First, in order to visualize the data, it makes sense to replace the original discrete points with a continuous density or distribution. This distribution can be found by dividing up the domain on which the data are located to form a histogram, or by replacing each data point with a distribution. Then the distribution for the whole data set is the sum of distribution functions for each data point. Using this distribution method is often preferable from the point of view of data visualization because the result does not have the discontinuities that are artifacts of histogram methods. On the other hand, it can be more computationally intensive to use distribution methods. Another reason for replacing each individual data point with a distribution is that the initial data may have some associated measurement error, and replacing each point with a normalized distribution reflects this error. In contrast to the other statistical analysis approaches in Bahar and Jernigan (1996), Banavar, Maritan, and Seno (2002), Buchete, Straub, and Thirumalai (2004a, 2004b, 2004c), our approach is able to smooth data and reflect potential measurement errors in a very natural way.

The second issue related to smoothing and filtering is related to the selection of proper distributions. In the case of data on the line or in multidimensional Cartesian coordinates, the Gaussian distribution is a popular choice because of its nice properties and the physical nature of its origins. Hence, part of this paper is about how the classical Gaussian functions can be generalized to capture both positional and orientational data, and then the application of these ideas to real protein data.

2. Review of Terminology and Notation from Molecular Biophysics

Proteins are composed of 20 different amino acids: alanine (Ala); arginine (Arg); asparagine (Asn); aspartic acid (Asp); cysteine (Cys); glutamine (Gln); glutamic acid (Glu); glycine (Gly); histidine (His); isoleucine (Ile); leucine (Leu); lysine (Lys); methionine (Met); phenylalanine (Phe); proline (Pro); serine (Ser); threonine (Thr); tryptophan (Trp); tyrosine (Tyr); valine (Val). Amino acids are classified into three groups: the hydrophobic group has Ala, Ile, Leu, Met, Phe, Pro, and Val; the charged group has Arg, Asp, Glu, and Lys; the polar group has Asn, Cys, Gln, His, Ser, Thr, Trp, and Tyr (Branden and Tooze 1999).

Each amino acid can be divided into two parts: main-chain atoms and side chains. The main-chain part has a central carbon atom (C_α) which is attached to a hydrogen atom (H), an amino group (NH_2), and a carboxyl group (COOH). However, the side chain bound to the C_α atom is different for each different amino acid (Branden and Tooze 1999). See Figure 1.

A protein is a polypeptide chain consisting of amino acid residues. These residues are what remains from amino acids that have bonded by releasing a water molecule (one H and one OH from each joining pair). Figure 2 shows a method to separate a polypeptide chain into repeating units (Branden and Tooze 1999). That is, a polypeptide chain is divided into peptide units that go from one C_α atom to the next C_α atom. Two "torsion angles" called *phi* (ϕ) and *psi* (ψ) provide a way to characterize conformational information of protein backbones since bond lengths and bond angles are relatively fixed. The rotation angle around the N–C_α bond is called *phi* (ϕ) and the rotation angle around the C_α–C' bond from the same C_α atom is called *psi* (ψ). Ramachandran and Sasisekharan (1968) introduced a planar plot, now called the Ramachandran plot, where the angles, ϕ and ψ, are the axes, and allowable regions in this plane are shaded.

Although the overall structure of a protein molecule can be irregular, within each protein so-called secondary structures show regularity. The secondary structures usually consist of two types: *alpha* (α) *helices* or *beta* (β) *sheets*. They are characterized by many consecutive residues with similar phi (ϕ), psi (ψ) angles.

The alpha helix is a significant component of secondary structures. Residues comprising an alpha helix have a phi angle of about -60° and a psi angle of about -50° (Branden and Tooze 1999). The alpha helix has 3.6 residues per turn, which corresponds to 5.4 Å rise along the helical axis (1.5 Å per residue; Branden and Tooze 1999). The second impor-

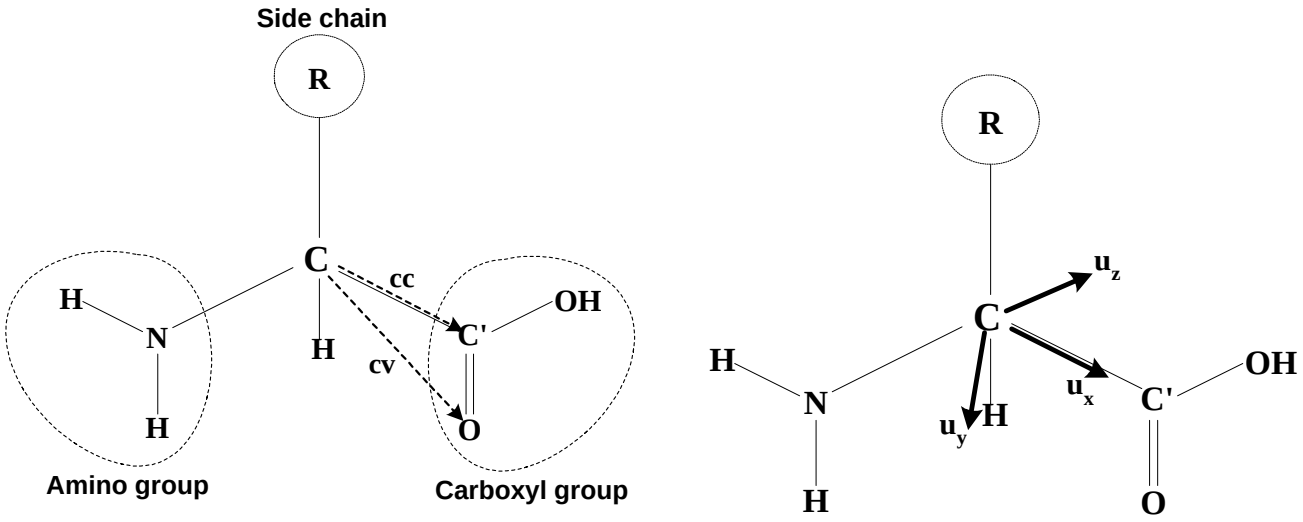


Fig. 1. Schematic diagram of an amino acid. A central carbon atom (C_α) is attached to an amino group (NH_2), a carboxyl group (COOH), a hydrogen atom, and a side chain (R). This also shows how a local reference frame $[\mathbf{u}_x, \mathbf{u}_y, \mathbf{u}_z]$ is determined using C_α, C, and O (vectors **cc** and **cv**).

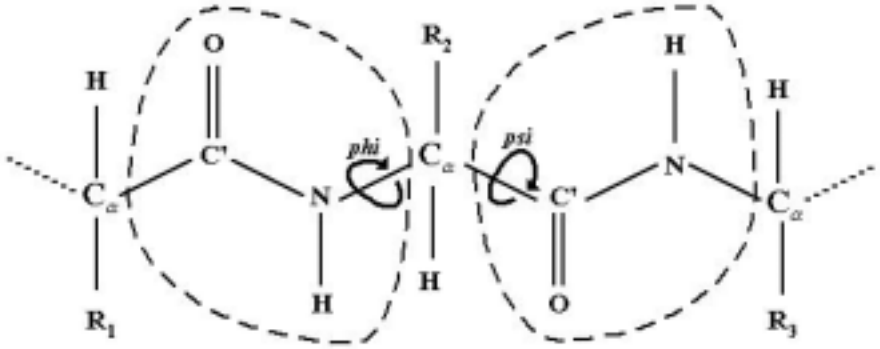


Fig. 2. Two peptide units. Each peptide unit has the C_α atom and the $C' = O$ group of amino acid n in addition to the NH group and the C_α atom of amino acid n+1. Each such unit is planar and more or less rigid.

tant secondary structure is the beta (β) sheet. This structure is constructed from a combination of several regions of the polypeptide chain. These regions are called β strands. Beta strands are generally from five to ten residues long and they are found in the upper-left quadrant of the Ramachandran plot (Branden and Tooze 1999). There are two types of β sheets: parallel and antiparallel. In parallel β sheets, the amino acids in the aligned β strands can all run in the same biochemical direction. In antiparallel β sheets, the amino acids in successive strands can have alternating directions.

Whereas the Ramachandran plot is now a standard method for describing constraints between adjacent amino acid residues, no such tool exists for examining correlations between sequentially distant but spatially proximal residues. Before attempting to generate Ramachandran-like plots with six-dimensional pose data for residues that are sequentially distant and spatially proximal, we first need to affix a local coordinate frame to each amino acid. The origin of the local frame resides at the C_α atom and the frame orientation is specified by three atoms, C_α, C, and O. In Figure 1, the x-axis of the frame is obtained from a vector **cc** that connects C_α and C. The cross product of **cc** and **cv** determines the z-axis, where **cv** is a vector connecting C_α and O. Therefore, the unit vectors pointing along the x-axis and z-axis are

$$\mathbf{u}_x = \frac{\mathbf{cc}}{||\mathbf{cc}||}, \qquad \mathbf{u}_z = \frac{\mathbf{cc} \times \mathbf{cv}}{||\mathbf{cc} \times \mathbf{cv}||}.$$

By the cross product, $\mathbf{u}_z \times \mathbf{u}_x$, the remaining y-axis is determined.

3. Gaussian Functions for $SO(3)$

In this section we present a Gaussian function for $SO(3)$, the group of rotations in three-dimensional space. This is similar to the folded normal density solution on the circle discussed in the following subsection. This presentation builds on the work of Chirikjian and Chétalet (2002) and Chirikjian and Kyatkin (2000).

3.1. Gaussian Functions on the Line and Circle

Here we examine a distribution which is useful for smoothing discrete data on the line and circle. A natural way to perform smoothing is through diffusion.

The heat equation on the real line is

$$\frac{\partial F}{\partial t} = K \frac{\partial^2 F}{\partial x^2}$$

where $F(x, t)$ is the temperature in a material. Here K is a constant $\sqrt{k/(\sigma\rho)}$ determined by the thermal conductivity k, specific heat σ, and the density ρ of the material. The solution of this equation subject to the initial condition $F(x, 0) = \delta(x - 0)$ is known as a Gaussian or normal distribution, and is given in Kreyszig (1999) by

$$F(x, t) = \frac{1}{2\sqrt{\pi K t}} \mathrm{e}^{-x^2/4Kt}. \tag{1}$$

A natural question may be how the Gaussian distribution is generalized to spaces other than the real line. The next easiest one-dimensional case is the unit circle. It may be shown that the solution to the heat equation on the circle is obtained by "wrapping" the solution of the heat equation on the line around the circle, i.e., shifting all intervals on the line of the form $[2\pi n, 2\pi(n + 1)]$ for $n \in \mathbf{Z}$ to the interval $[0, 2\pi]$, and superposing the values of the function. This is written as

$$f(\theta, t) = \sum_{n=-\infty}^{\infty} F(\theta - 2\pi n, t). \tag{2}$$

A nice feature of the expansion in eq. (2) is that when Kt is small, only one or at most a few terms in the expansion need to be retained since the Gaussian function decays so rapidly. In the next subsection we discuss an analogous folded normal distribution for $SO(3)$.

3.2. Folded Normal Density Solution for $SO(3)$

If the axis direction and the angle of a rotation are denoted as $\mathbf{n} = [n_1,\ n_2,\ n_3]^{\mathrm{T}} \in S^2$ and $\theta \in [-\pi, \pi]$, respectively, then a rotation matrix can be written as (Murray, Li, and Sastry 1994; Chirikjian and Kyatkin 2000)

$$\mathrm{ROT}[\mathbf{n}, \theta] = \exp(\theta N).$$

Here, S^2 is the unit sphere and N is the skew-symmetric matrix such that $N\mathbf{x} = \mathbf{n} \times \mathbf{x}$ for every $\mathbf{x} \in \mathbb{R}^3$ and $||\mathbf{n}|| = 1$. The vector $\mathbf{n}$ is called the dual vector of N.

A natural way to define a Gaussian function for $SO(3)$ is as the solution of the heat equation, just as was done for the line and circle in the previous subsection. That is, we seek the solution of the equation

$$\frac{\partial F}{\partial t} = K \nabla^2_{SO(3)} F \tag{3}$$

with an initial condition $F(R, 0) = \delta(R)$. The Laplacian operator for $SO(3)$ is written in the axis-angle parametrization as (Varshalovich, Moskalev, and Khersonskii 1988; Chirikjian and Kyatkin 2000)

$$\nabla^2_{SO(3)} = \frac{\partial^2}{\partial\theta^2} + \cot\theta/2 \frac{\partial}{\partial\theta} + \frac{1}{4\sin^2\theta/2}\left(\frac{\partial^2}{\partial\lambda^2} + \cos\lambda \frac{\partial}{\partial\lambda} + \frac{1}{\sin^2\nu}\frac{\partial^2}{\partial\nu^2}\right), \tag{4}$$

where λ and ν are spherical coordinates for the vector $\mathbf{n} = \mathbf{n}(\lambda, \nu)$.

We seek a solution that is a class function on $SO(3)$ since such functions have the useful property that they commute under convolution with all other functions. Since every class function for $SO(3)$ is a function only of the angle of rotation θ, eq. (3) simplifies to

$$\frac{\partial F}{\partial t} = K\left(\frac{\partial^2 F}{\partial\theta^2} + \cot\theta/2 \frac{\partial F}{\partial\theta}\right). \tag{5}$$

Chirikjian and Chétalet (2002) proposed one possible generalization of the concept of a Gaussian function for the group $SO(3)$. This solution is analogous to the folded normal density solution (2) on the circle. This candidate Gaussian function is modified in Lee (2002) as

$$F(\theta, t) = C \frac{e^{Kt/4}}{(\pi K t)^{3/2}} \frac{\theta}{\sin\theta/2} \mathrm{e}^{-\theta^2/4Kt}, \tag{6}$$

which is folded around the circle defined by $-\pi \le \theta \le \pi$, as in eq. (2). This produces the Gaussian for $SO(3)$, where θ is the angle from the axis-angle parametrization of $SO(3)$. The scaling factor C is the mass we choose to give each $SO(3)$-Gaussian distribution. Ways of choosing this value are discussed in the next subsection, as are reasons for using this function for representing orientational data.

3.3. Why Using the Usual Gaussian is Not Sufficient for Orientational Averaging

The space of all vectors $\mathbf{x} = \theta\mathbf{n}(\lambda, \nu)$, is often used to represent $SO(3)$ as a solid ball of radius π in $\mathbb{R}^3$ with antipodal

points identified. For any parametrization (q_1, q_2, q_3) of $SO(3)$ (including (x_1, x_2, x_3), (θ, λ, ν) and Euler angles (α, β, γ)), integration is performed as

$$\int_{SO(3)} f(R)\,\mathrm{d}(R) = \int_{\mathbf{q}\in Q} f(R(\mathbf{q}))\, w(\mathbf{q})\,\mathrm{d}q_1\,\mathrm{d}q_2\,\mathrm{d}q_3$$

where $w(\mathbf{q})$ is proportional to the Jacobian determinant $|\det(J(R(\mathbf{q}))|$ where the Jacobian matrix $J(R(\mathbf{q}))$ relates rates of change in $\mathbf{q}$ to angular velocity and Q is the region defined by all values of $\mathbf{q}$ required to cover $SO(3)$ once. In the context of the parametrizations discussed in the previous subsection,

$$|\det(J(R(\theta, \lambda, \nu))| = 4\sin^2(\theta/2)\sin\nu \quad \text{and}$$

$$|\det(J(R(\mathbf{x}))| = \frac{2(1-\cos\|\mathbf{x}\|)}{\|\mathbf{x}\|^2} \tag{7}$$

where $\mathbf{x} = \theta\mathbf{n}(\lambda, \nu)$ are the parameters we have used to display the data.

If one wants to display $SO(3)$ data as if they are data in $\mathbb{R}^3$, then one needs to normalize correctly. That is, if one observes a distribution $\rho_{obs}(R(\mathbf{q})) = \tilde{\rho}_{obs}(\mathbf{q})$, one needs to recognize that this has a built-in bias, and is related to the actual underlying probability density as

$$\tilde{\rho}_{obs}(\mathbf{q}) = \tilde{\rho}_{act}(\mathbf{q})\, w(\mathbf{q}).$$

When there are discrete observed data, this relationship is equivalent to the following:

$$\tilde{\rho}_{obs}(\mathbf{q}) = \frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{q}-\mathbf{q}_i) \quad \text{and}$$

$$\tilde{\rho}_{act}(\mathbf{q}) = \frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{q}-\mathbf{q}_i)/w(\mathbf{q}_i). \tag{8}$$

When smoothing orientational data, it is not sufficient to replace each Dirac delta function $\delta(\mathbf{q})$ in eq. (8) with a kernel $k(\mathbf{q})$ such as a Cartesian Gaussian function because this would not preserve the mass contributed by each of the original data points. However, a smoothing and renormalization of the form

$$\delta(\mathbf{q}-\mathbf{q}_i)/w(\mathbf{q}_i) \;\rightarrow\; \frac{k(\mathbf{q}-\mathbf{q}_i)}{\int_{\mathbf{q}\in Q} k(\mathbf{q}-\mathbf{q}_i)w(\mathbf{q})\,\mathrm{d}\mathbf{q}}$$

would preserve mass.

The $SO(3)$ Gaussian function effectively has this geometric normalization built in already, and so no additional normalization is required. It also has the added feature that when it is shifted as $f(R(\mathbf{q})) \rightarrow f(R^T(\mathbf{q}_i)R(\mathbf{q}))$ it does not distort in $SO(3)$, whereas a transformation of the form $k(\mathbf{q}) \rightarrow k(\mathbf{q}-\mathbf{q}_i)$ potentially can lead to significant distortions in $SO(3)$ as the variance of the kernel becomes large.

We now address the minor issue of how to choose C, which involves a normalization that depends on a subjective choice rather than being dictated by geometry. Integrating the folded version of eq. (6) over $SO(3)$ yields

$$\int_{\theta=-\pi}^{\pi}\int_{\nu=0}^{\pi/2}\int_{\lambda=0}^{2\pi} f(\theta, t)4\sin^2(\theta/2)\sin\nu\, d\gamma\, d\nu\, d\theta = 16C.$$

Therefore, a choice of $C = 1/16$ will ensure that the $SO(3)$-Gaussian $f(\theta, t)$ has unit mass under this definition of $SO(3)$ integral. However, often the $SO(3)$ integral is normalized so that $\int_{SO(3)} 1\,\mathrm{d}R = 1$ rather than $8\pi^2$, which is what is obtained when using $w(\theta, \nu, \lambda) = 4\sin^2(\theta/2)\sin\nu$ (Chirikjian and Kyatkin 2000). If this is done, one would use $w(\theta, \nu, \lambda) = (1/2\pi^2)\sin^2(\theta/2)\sin\nu$. In this case, one should define $C = \pi^2/2$ in order for each Gaussian to have unit mass. Of course, if one wants the contribution from n points to be a probability density, an additional division by n would be required.

4. Analysis of Protein Pose Statistics Using Generalized Gaussian Functions

The PDB (Berman et al. 2000) is a huge collection of information about the structure (x–y–z position of atoms) within thousands of different proteins. Various experimental methods are used to determine these structures, and some methods have larger error than others. The statistical analysis presented here is based on some of the most accurate data.

Table 1 lists the PDB codes for 168 structures used in our analysis. All together there are 37,971 residues in these proteins. Table 2 shows the number of residues for each amino acid type. These 168 are a subset of the structures used by Chakrabarti and Debnath (2001). The structures were chosen from the PDB at the Research Collaboratory for Structural Bioinformatics (RCSB; http://www.rcsb.org/pdb/).

The resolution of the structures is 2.0 Å or better, and the R-factor is less than 20%. The resolution of the diffraction data depends on how well ordered the crystals are. In the process of crystallographic refinement of a model, the model is changed to minimize the difference between the experimentally observed diffraction amplitudes and those calculated for a hypothetical crystal containing the model instead of the real molecule. This difference is expressed as an R-factor (Branden and Tooze 1999). In general, 2.0 Å resolution and 20% R-factor are considered sufficiently good. The maximum sequence identity between any two of the polypeptide chains is $\leq$25% (Branden and Tooze 1999). This ensures that our statistics are not biased because we sample a set of non-homologous proteins.

4.1. Distributions of Relative Orientation Between Residues

Figures 3–8 show plots of relative orientation data between two local coordinate frames affixed to the C_α of amino acids.

Table 1. PDB Codes for the Structures Used in Our Analysis of Relative Pose

153L	16PK	1A3C	1A48	1A6M	1A7S	1A8D	1A8E
1ABA	1ADS	1AK1	1AMF	1AMM	1AQB	1ARU	1AUN
1AWD	1AXN	1AYL	1AZO	1B0Y	1B6G	1BDO	1BEA
1BEC	1BFD	1BFG	1BG6	1BGF	1BJ7	1BK0	1BM8
1BRT	1BS9	1BTN	1BXA	1BY1	1BY2	1C3D	1C52
1CEO	1CEX	1CFB	1CNV	1CPO	1CPQ	1CSH	1CV8
1CVL	1DCS	1DHN	1DIN	1DUN	1ECD	1EDG	1EUS
1EZM	1FIT	1FNA	1FUS	1G3P	1GCI	1GKY	1GOF
1GSA	1HFC	1HKA	1HOE	1HXN	1IAB	1IXH	1JDW
1JER	1KNB	1KOE	1LAM	1LCL	1LIS	1LKI	1LOU
1MDC	1MLA	1MML	1MOQ	1MRJ	1MSK	1MUN	1NAR
1NIF	1NKR	1NLR	1NLS	1NOX	1NP4	1NPK	1OAA
1OPY	1PBE	1PGS	1PHF	1PLC	1PNE	1POA	1POC
1PPN	1PTY	1RCF	1REC	1RHS	1RIE	1RZL	1SFP
1SKF	1SMD	1SRA	1SUR	1SVY	1TCA	1TIB	1TML
1VHH	1VID	1VLS	1VNS	1WAB	1WHI	1WHO	1XNB
1YCC	1YGE	2A0B	2ABK	2ACY	2AYH	2CBP	2CTC
2DRI	2DTR	2EBN	2END	2GAR	2GDM	2HBG	2HFT
2ILK	2PII	2PTH	2PVB	2QWC	2RN2	2SAK	2SNS
3CHY	3CLA	3CYR	3ENG	3GRS	3LZT	3PTE	3SEB
3SIL	3TDT	3TSS	3VUB	5P21	6CEL	7RSA	8ABP

Table 2. Number of Residues for Each Amino Acid Type

Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
3218 (8.47 %)	1720 (4.53 %)	1882 (4.96 %)	2231 (5.88 %)	601 (1.58 %)	1415 (3.73 %)	2128 (5.60 %)	3029 (7.98 %)	834 (2.20 %)	1960 (5.16 %)
Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
3108 (8.19 %)	2143 (5.64 %)	729 (1.92 %)	1502 (3.96 %)	1893 (4.99 %)	2522 (6.64 %)	2315 (6.10 %)	616 (1.62 %)	1495 (3.94 %)	2630 (6.93 %)

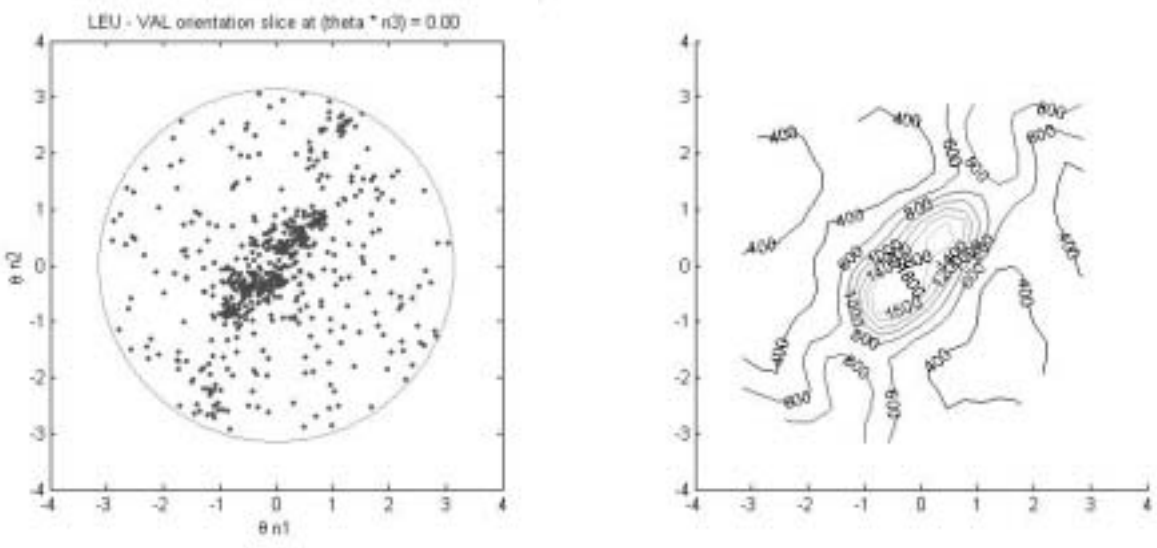
Here amino acids are sequentially distant and spatially proximal. Two cutoff values are used so that the sequential distance of residue pairs is three or higher and the spatial distance of residue pairs is less than 10.0 Å.

Each figure consists of two plots. The left plot displays relative orientation data in the form of discrete points on a planar slice. The coordinates of each point are the three components of $\theta\mathbf{n}$ where $\mathbf{n} = [n_1,\ n_2,\ n_3]^{\mathrm{T}}$ is the rotation axis and θ is the rotation angle. Both plots are planar slices that are cut at the origin and perpendicular to the axis of n_3. Note that, in general, the slice at $\theta n_3 = 0$ is the most populated one. The thickness of each slice is $\pi/10$.

In addition to visualization with points, the relative orientation data are visualized with a continuous distribution function, which is the sum of Gaussian functions for $SO(3)$ described in Section 3.2. In this approach each dot on the left plot is considered as a heat source, i.e., the initial condition in the form of the delta function. Then the distribution for all the points is the sum of distribution functions for each data point. Note that a scaled version of eq. (6) was used to produce plots and the scaling value was 0.1. Here, the parameter Kt in eq. (6) was set to 0.05. The right plot of each figure illustrates the sum of diffusions of each heat source in the form of contours. Each contour is labeled as a number. Higher numbers mean that points (heat sources) are concentrated. We observe in each figure that locations of clusters in both plots are identical.

Most hydrophobic–hydrophobic pairs appear to have multiple clusters on the slice. In particular, pairs associated with valine have clusters near the center of the slice. Figure 3 displays the distribution of relative orientation data of the leucine–valine pair.

Every charged–charged pair is found to have multiple symmetrical clusters near the center along a line. This is due to



(a) (b)

Fig. 3. Distribution of relative orientation data of the leucine–valine pair.

pairs within the same α helix, which is shown in Figures 10 and 12(d). For instance, Figure 4 is the plot for the arginine–glutamic acid pair.

No common attribute is observed in polar–polar amino acid pairs. Figure 5 shows the distribution for the glutamine–threonine pair. One larger clump is found near the center.

Many of the hydrophobic–charged pairs are found to have multiple clusters. Figure 6 displays the distribution for the alanine–glutamic acid pair where four symmetric clusters are seen along the line. This plot appears quite similar to the plot of the arginine–glutamic acid pair, which is in Figure 10(a). The alanine–glutamic acid pair will be revisited later with plots of relative orientational data for several values of θn_3, which are in Figure 9.

In polar–charged pairs, pairs associated with glutamine show two symmetrical clusters along a line. For instance, Figure 7 is for the glutamine–glutamic acid pair. This plot also looks similar to the plot of the alanine–glutamic acid pair, but the plot of glutamine–glutamic acid pair appears sparser.

No common attribute is found in polar–hydrophobic amino acid pairs. Figure 8 shows the distribution for the tyrosine–isoleucine pair where we see concentrated areas near the center of the slice.

A set of plots in Figure 9 displays how the relative orientation data of the alanine–glutamic acid pair appear as the value of θn_3 changes from -0.94 to 1.26. Clusters appear to move from upper right to lower left as θn_3 increases, and they are not seen in the slices at $\theta n_3 \leq -0.94$ or $\theta n_3 \geq 0.94$.

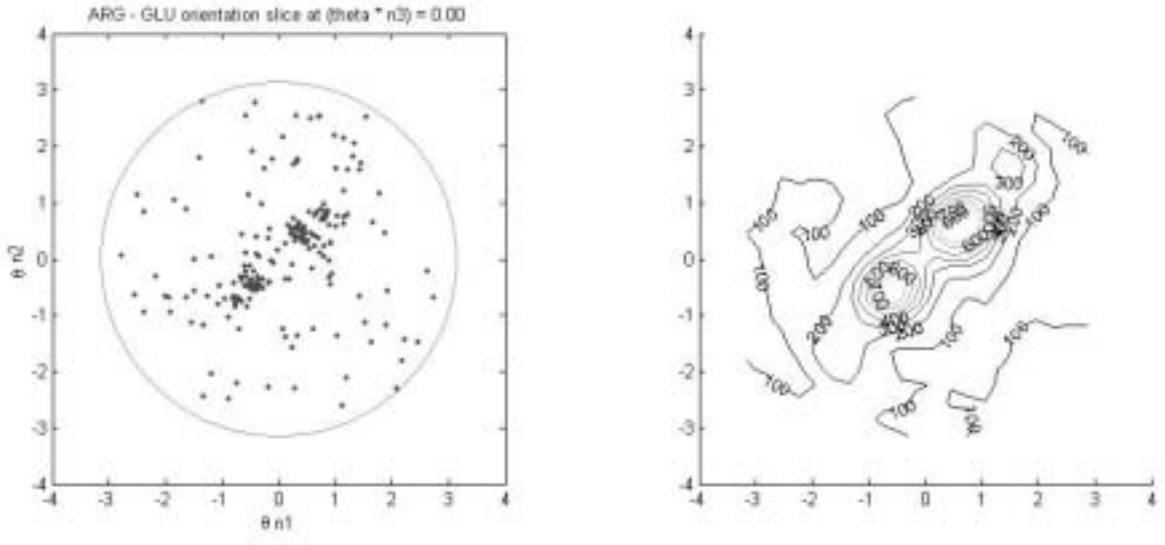
Now we discuss the sources of such clusters in distribution plots of relative orientation data in order to extract more detailed information about clusters. In particular, residues in secondary structures, i.e., helices or sheets, draw more attention. We also examine if clusters are related to the sequential distance of residue pairs. Since we use the number two for the cutoff value in the sequential distance, pairs that have the sequential distance of 3, 4, 5 were investigated more thoroughly.

For orientation data, we take the arginine–glutamic acid pair for example. Figure 10 illustrates distribution plots of relative orientation of the pair when $\theta n_3 = 0.0$. It is observed in Figure 10(a) that four clusters labeled as a–d exist near the center of the slice. In Figure 10(b), we can find that those clusters are from residue pairs within the same α helix. However, contributions of other secondary structures like 3_{10} helices or β sheets to the clusters are negligible, and thus they are omitted. If we observe Figures 10(c) and 10(d), clusters a and d are from pairs with the sequential distance of 3 and clusters b and c are from pairs with the sequential distance of 4.

Those four clusters in Figure 10(a) are examined in another way. The mean of relative orientation matrices of each cluster is calculated and is displayed in Figure 11. The mean for each cluster is obtained by finding a rotation matrix R_m to minimize the following cost function

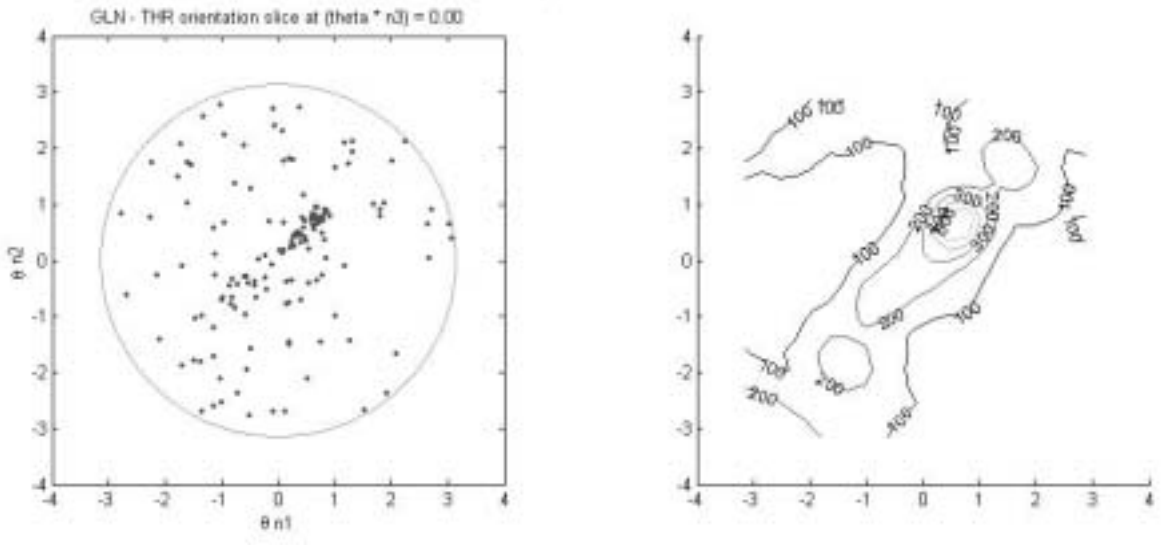
$$C(R_m) = \sum_{i=1}^{n} \| R_m - R_i \|^2$$

where $R_m, R_i \in SO(3)$. Gradient descent on $SO(3)$ is used to solve for R_m. In analogy with the definition of the partial derivative (or directional derivative) of a scalar function of $\mathbb{R}^N$-valued argument, we can define differential operators



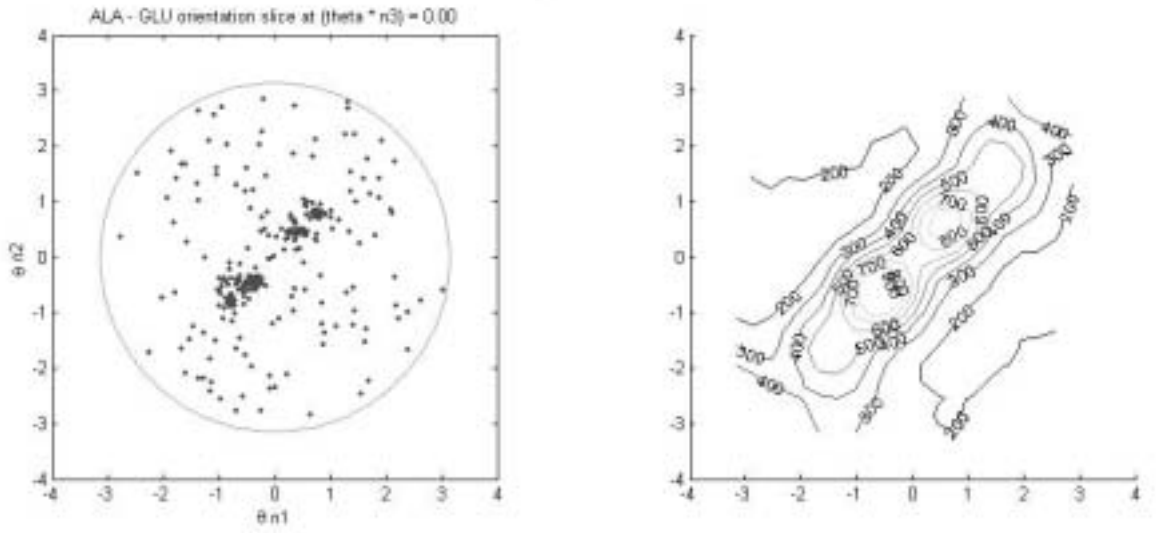
(a) (b)

Fig. 4. Distribution of relative orientation data of the arginine–glutamic acid pair.



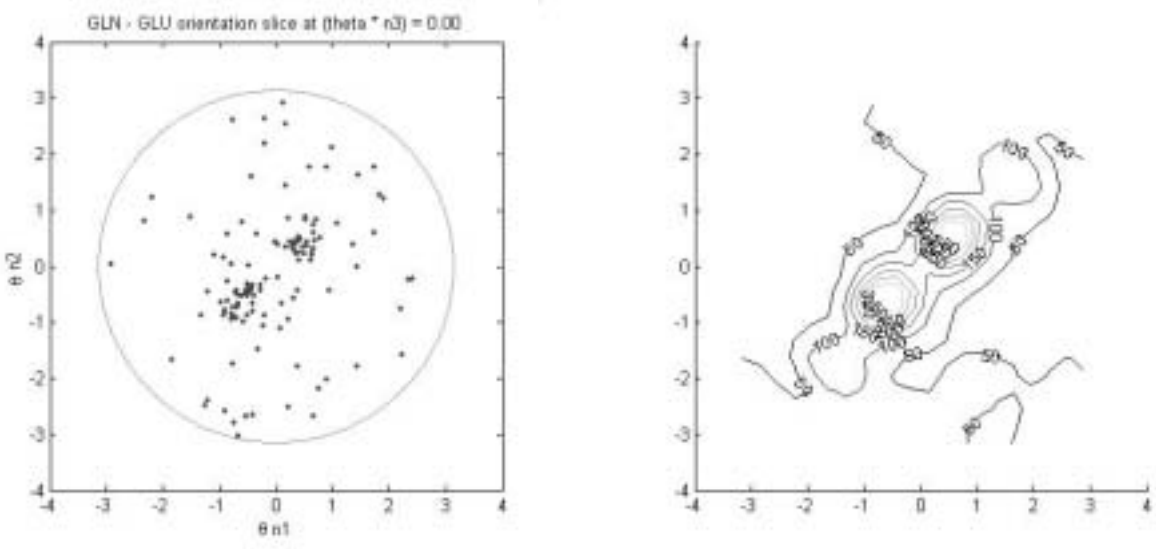
(a) (b)

Fig. 5. Distribution of relative orientation data of the glutamine–threonine pair.



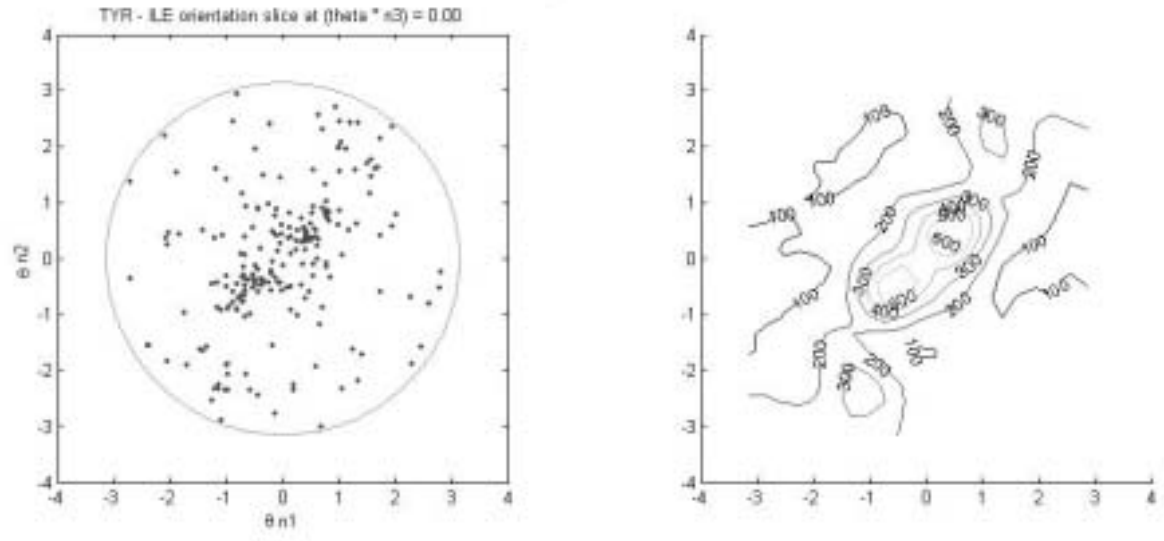
(a) (b)

Fig. 6. Distribution of relative orientation data of the alanine–glutamic acid pair.



(a) (b)

Fig. 7. Distribution of relative orientation data of the glutamine–glutamic acid pair.



(a) (b)

Fig. 8. Distribution of relative orientation data of the tyrosine–isoleucine pair.

which act on functions of rotation-valued argument. Refer to Chirikjian and Kyatkin (2000) and Lee, Fichtinger, and Chirikjian (2002) for the definition of differential operators and a specific example of the gradient descent method. The three column vectors of $R_m = [\mathbf{u}, \mathbf{v}, \mathbf{w}]$ are illustrated in Figure 11.

It is observed that R_m of cluster a and R_m of cluster b are nearly equal to the inverse of R_m of cluster d and the inverse of R_m of cluster c, respectively. This is explained by recalling that clusters a and d are made from some pairs with the same sequential distance and clusters b and c are from other pairs with the same sequential distance. In general, the pose distribution of residue pairs (i, j) that are sequentially apart by $+n$ $(i - j = +n)$ is related to the pose distribution of pairs that are sequentially apart by $-n$ by the following expression

$$f_{ij}(g) = f_{ji}(g^{-1}),$$

where $g \in SE(3)$ and $f_{ij}(g)$ is the pose probability density of a frame attached at j relative to a frame attached at i. If $g = (R, \mathbf{b})$ where $R \in SO(3)$ and $\mathbf{b} \in \mathbb{R}^3$, this can be written as $f_{ij}(R, \mathbf{b}) = f_{ji}(R^{\mathrm{T}}, -R^{\mathrm{T}}\mathbf{b})$. Note that integrating over position yields $f_{ij}(R) = f_{ji}(R^{\mathrm{T}})$, whereas integrating over orientation does not yield any useful relationship.

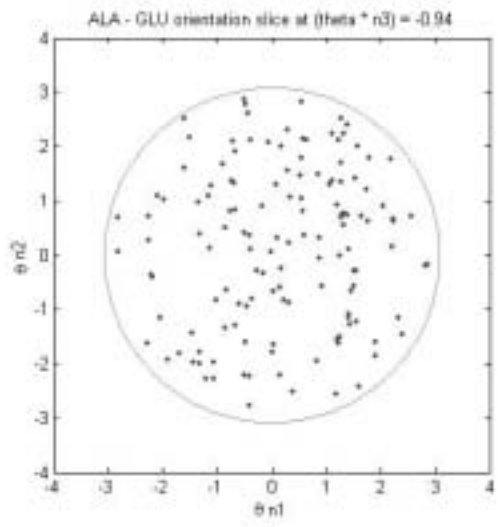
Residue pairs in secondary structures are examined in more detail. The set of plots in Figure 12 displays distributions of relative orientation data of all the residue pairs that are within the same α helix. Here θn_3 varies from 0.0 to 0.94. Note in Figure 12(d) ($\theta n_3 = 0.0$) that a symmetry exists in the clusters. In fact, we can picture a distribution plot for a negative value of θn_3 easily using the symmetry and the plot for the corresponding positive θn_3. We see from these plots that clusters move to the lower-left area and disappear as θn_3 grows.

As shown in Figure 13(a), the relative orientation data of pairs that are in different α helices are distributed widely. In other cases where pairs are either within the same 3_{10} helix or in different parallel/antiparallel strands, clusters are found. See Figures 13(b), 13(c), and 13(d). We excluded pairs that are either in different 3_{10} helices or within the same parallel/antiparallel strands because their portions are negligibly small.

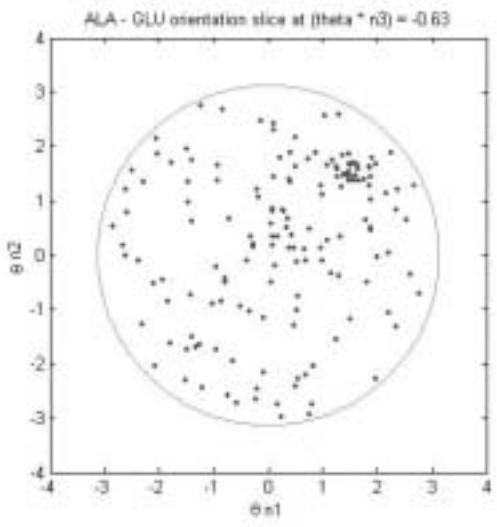
Residue pairs that are both sequentially and spatially proximal are now discussed. The sequential distance of every pair is either 1 or 2, and the spatial distance is less than 10.0 Å. The distribution of relative orientation data of all types of pairs is displayed in Figure 14. The value of θn_3 varies from 0.0 to 1.26. Several clusters appear in each plot and they can be differentiated by the sequential distance. For example, clusters labeled as a, c, e in Figure 14(a) have the sequential distance of 2, while the sequential distance of clusters b and d is 1. In Figure 14(d), the sequential distance of clusters b and c is 1 and that of cluster a is 2. We observe that concentrated areas with the sequential distance of 1 become larger as the value of θn_3 grows.

4.2. Distributions of Relative Position between Residues

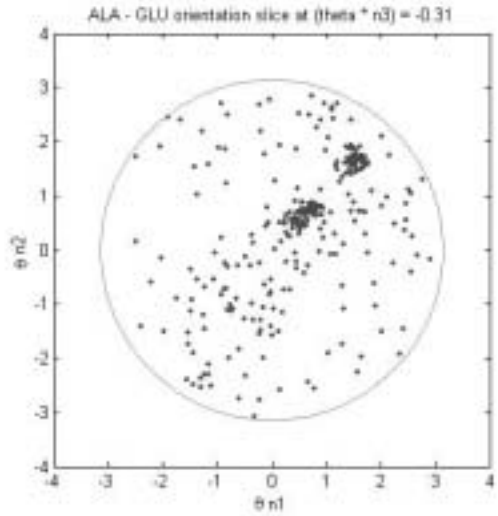
Now we begin to discuss the distribution of relative position data of residue pairs whose sequential distance is 3 or higher and whose spatial distance is less than 10.0 Å. Figures 15–20



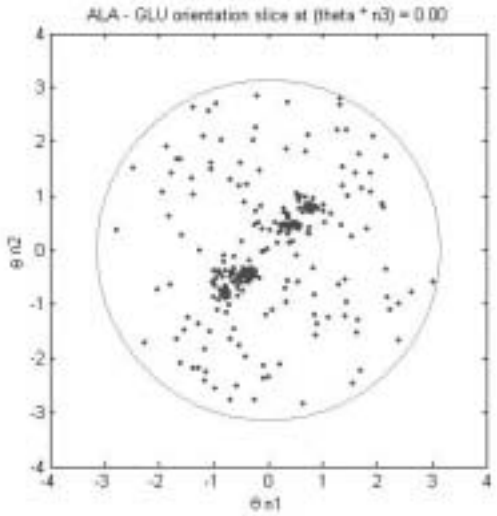
(a) θn_3 = -0.94



(b) θn_3 = -0.63

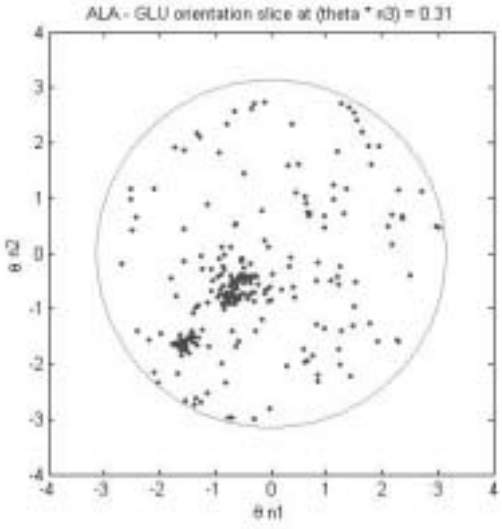


(c) θn_3 = -0.31

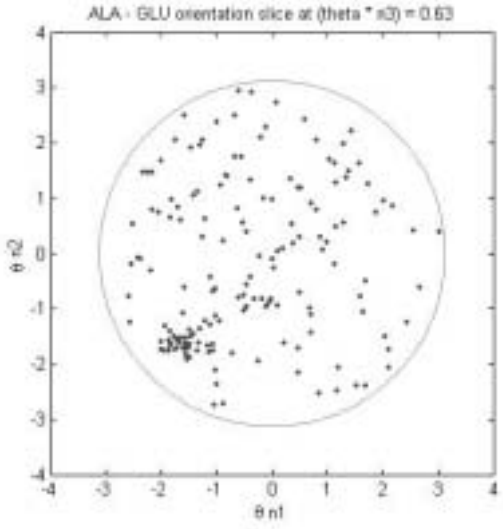


(d) $\theta n_3 = 0.00$

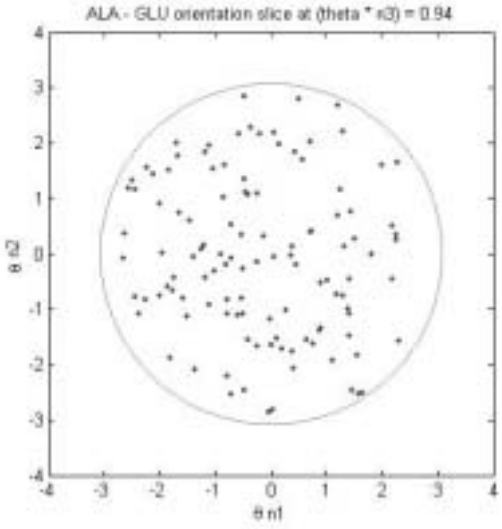
Fig. 9. Distribution of relative orientation data of the alanine–glutamic acid pair as θn_3 varies (continued on next page).



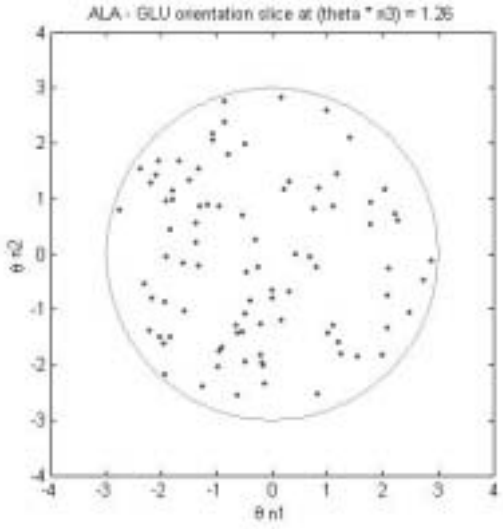
(e) $\theta n_3 = 0.31$



(f) $\theta n_3 = 0.63$

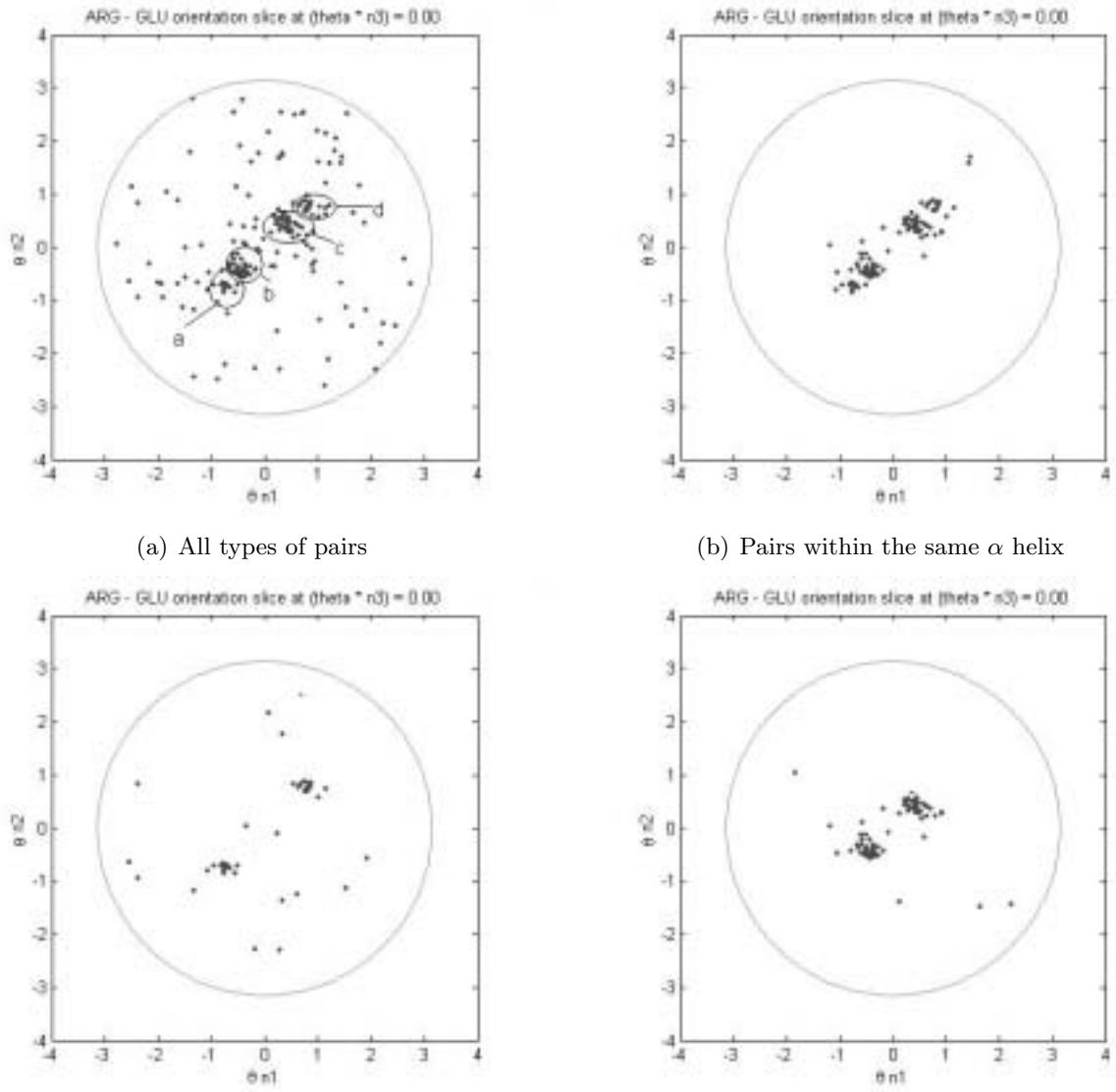


(g) $\theta n_3 = 0.94$



(h) $\theta n_3 = 1.26$

Fig. 9. (continued from previous page).



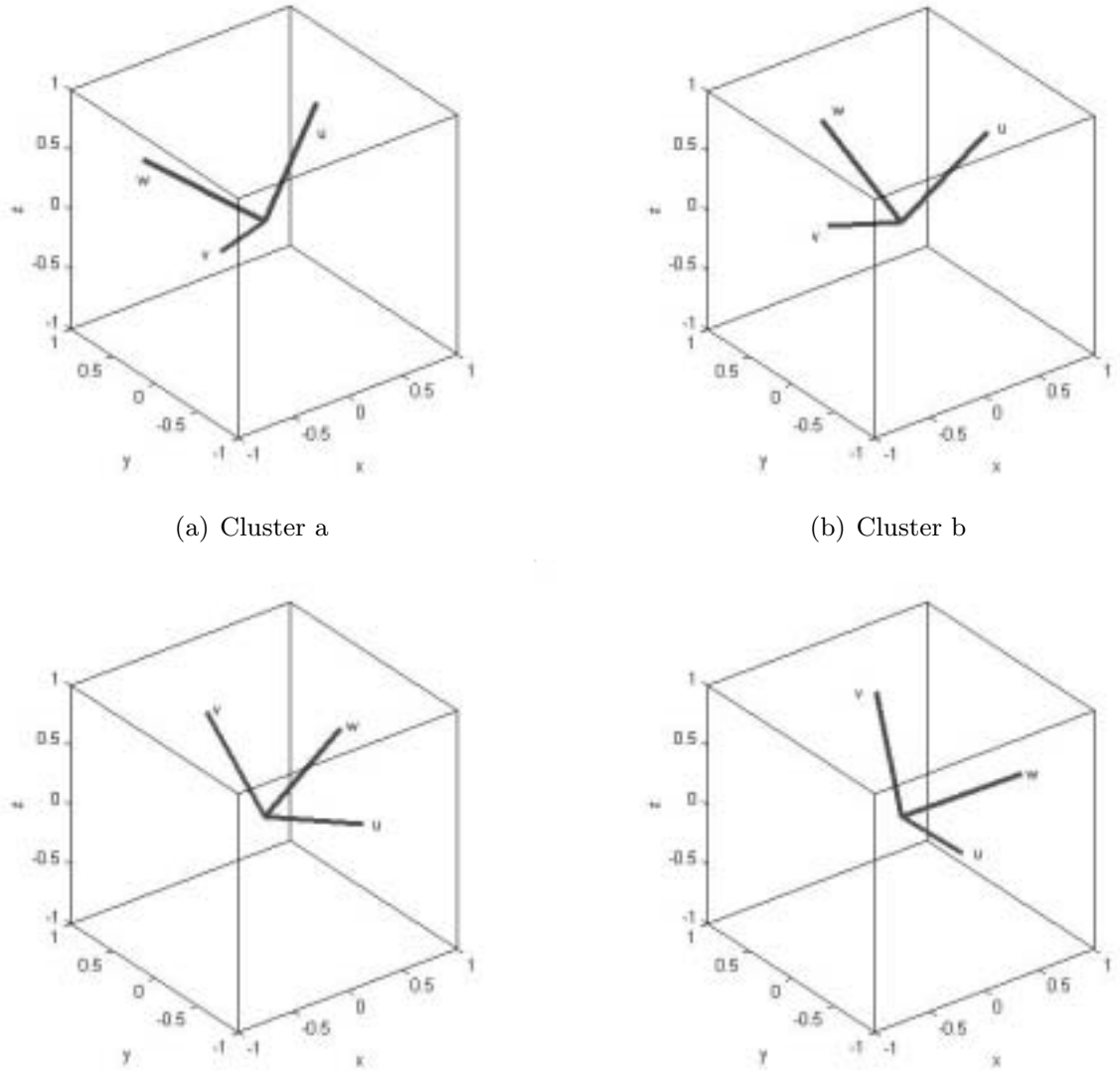
(a) All types of pairs

(b) Pairs within the same α helix

(c) Pairs with the sequential distance of 3

(d) Pairs with the sequential distance of 4

Fig. 10. Distribution of relative orientation data of the arginine–glutamic acid pair at $\theta n_3 = 0.0$.



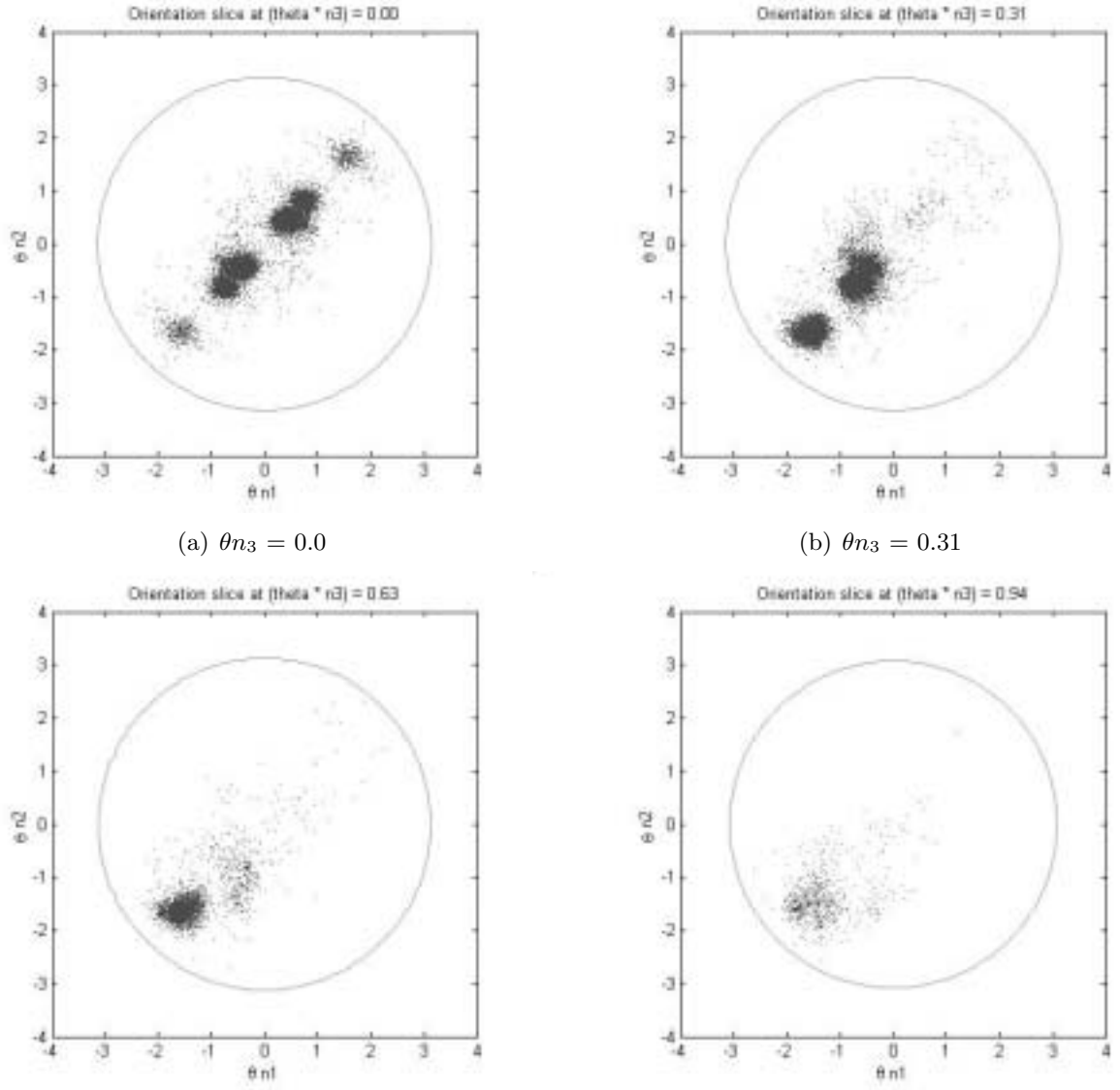
(a) Cluster a

(b) Cluster b

(c) Cluster c

(d) Cluster d

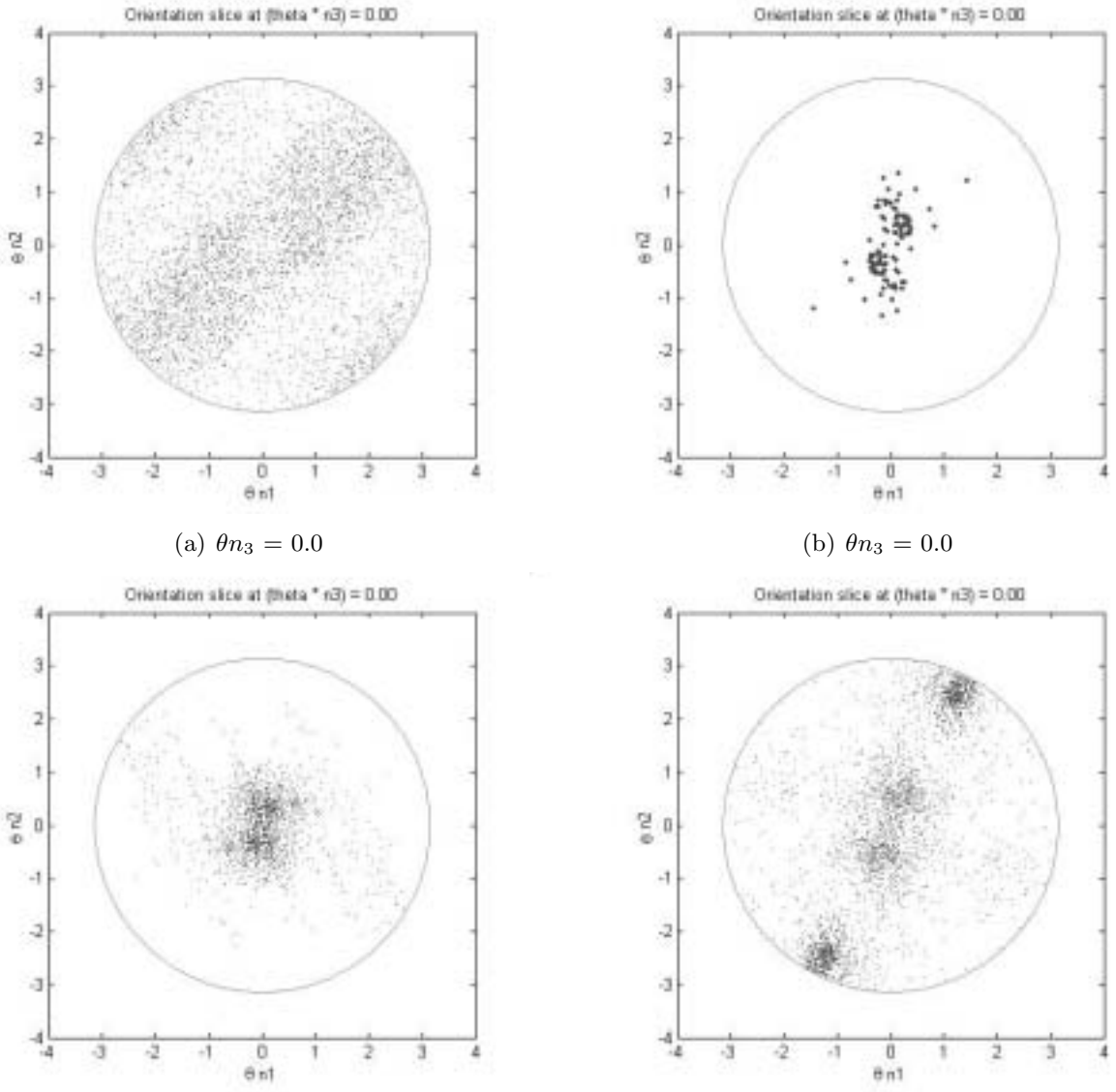
Fig. 11. Mean of orientation of clusters of the arginine–glutamic acid pair at $\theta n_3 = 0.0$.



(a) $\theta n_3 = 0.0$ (b) $\theta n_3 = 0.31$

(c) $\theta n_3 = 0.63$ (d) $\theta n_3 = 0.94$

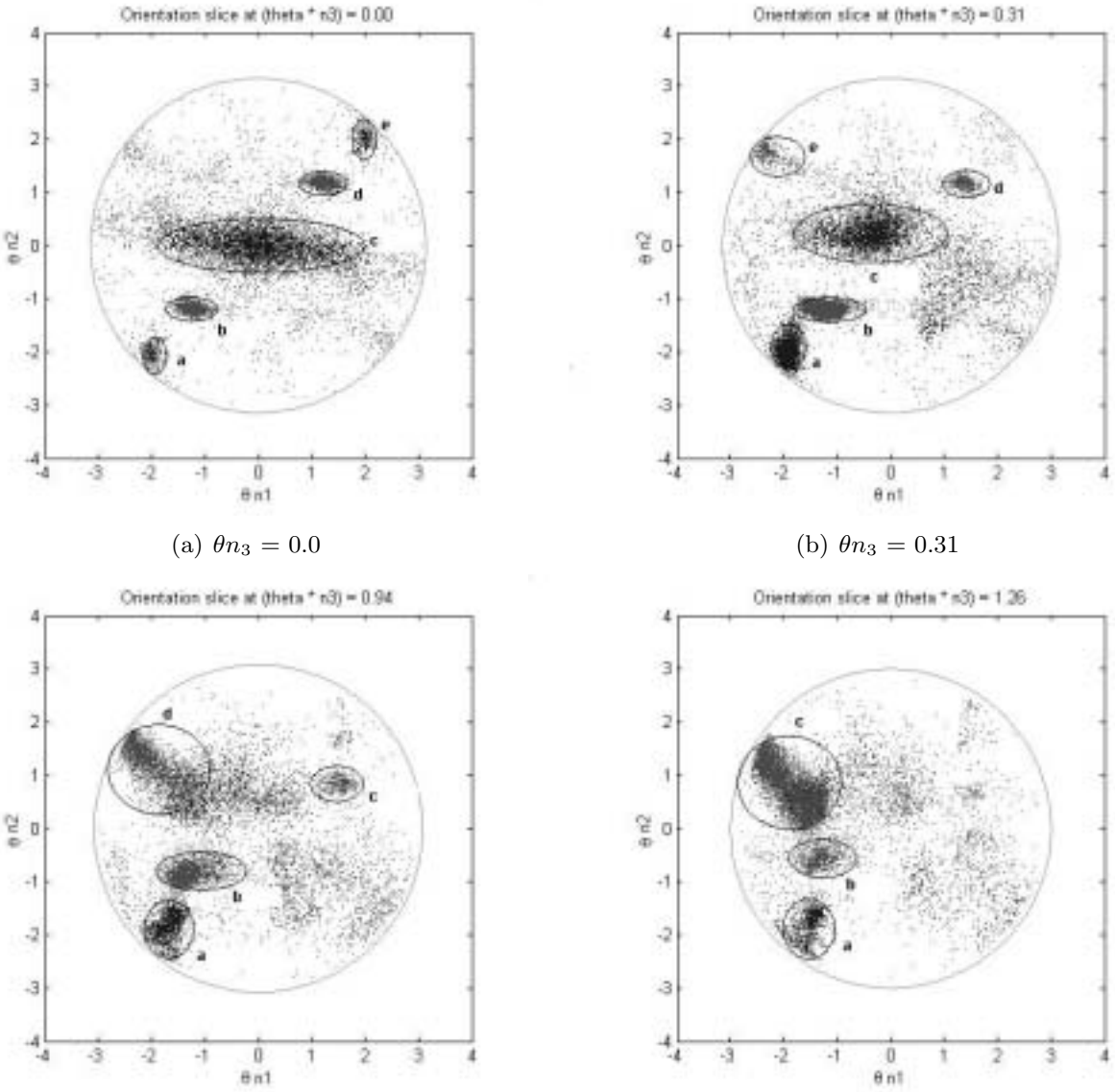
Fig. 12. Distribution of relative orientation data of all types of pairs that are within the same α helix.



(a) $\theta n_3 = 0.0$ (b) $\theta n_3 = 0.0$

(c) $\theta n_3 = 0.0$ (d) $\theta n_3 = 0.0$

Fig. 13. Distribution of relative orientation data: (a) all types of pairs in different α helices; (b) all types of pairs within the same 3_{10} helix; (c) all types of pairs in different parallel strands; (d) all types of pairs in different antiparallel strands.



(a) $\theta n_3 = 0.0$

(b) $\theta n_3 = 0.31$

(c) $\theta n_3 = 0.94$

(d) $\theta n_3 = 1.26$

Fig. 14. Distribution of relative orientation data of all types of pairs that are both sequentially and spatially proximal.

show plots of relative position data between the local coordinate frames of two amino acids. The left plot of each figure displays the relative position data in the form of points. The plots are planar slices that are cut at the origin and perpendicular to the z-axis. The thickness of each slice is 1.0 Å. Note that the overall shape of each plot appears like a ring. This is because steric effects limit amino acids from coming to close to each other.

In order to visualize the positional data with a continuous distribution function, we use the usual three-dimensional Gaussian function which is the three-dimensional version of eq. (1). Here the parameter Kt is set to a value between 0.3 and 0.9.

Most hydrophobic–hydrophobic pairs appear to have multiple clusters on the slice. Since hydrophobic residues make non-specific interactions, this result confirms the expectation. In particular, pairs associated with valine have three or four clusters along a circle whose radius is about 5.0 Å, i.e., half of the cutoff value. For example, Figure 15 displays the distribution of relative position data of the leucine–valine pair.

Every charged–charged pair is found to have two clusters along a circle with the radius of about 5.0 Å. In this case, electrostatic interactions are specific, so preferred orientations are shown. The radius of 5.0 Å is thought to be related to the distance above which salt bridges are unstable (Kumar and Nussinov 1999). Figure 16 is for the glutamic acid–lysine pair.

In polar–polar amino acid pairs, only pairs with glutamine show two clusters. The distribution for asparagine–glutamine pair is illustrated in Figure 17. This plot looks similar to the plot of glutamic acid–lysine pair. In terms of hydrophobicity scales, the hydrophobicity of glutamine and asparagine is adjacent to that of glutamic acid and lysine (Lesk 2001).

Some of the charged–hydrophobic pairs are found to have several clusters. Since the charged residues may have aliphatic side chains, they form non-specific hydrophobic interactions with the hydrophobic interaction. Figure 18 shows the distribution for the arginine–valine pair. This plot looks similar to that of the leucine–valine pair in Figure 15 but the plot of the arginine–valine pair looks sparser. This is because the number of residues for arginine is much smaller than that for leucine in the data set used in this analysis (see Table 2).

In polar–charged pairs, pairs with glutamine show two small clusters along a circle whose radius is about 5.0 Å. For instance, Figure 19 shows the distribution for the glutamine–lysine pair. This also seems to be related to the hydrophobicity of the residues.

In hydrophobic–polar pairs, strong clusters are not found. Figure 20 displays the distribution for the proline–threonine pair. Note that although proline belongs to the hydrophobic group, it is the least hydrophobic in the group.

A set of plots in Figure 21 displays how the relative position data of the glutamic acid–lysine pair are distributed as the value of z changes from -3.0 to 2.0 Å. Clusters appear to move from upper right to lower left by clockwise rotation as z increases. This is thought to be related to helix geometry because helices are the largest contributor to clusters in this residue pair, which is explained more clearly in Figure 22.

As we did for orientation data earlier, we discuss sources of clusters in distribution plots of relative position data to extract more detailed information about clusters. Again, more attention was paid to residues in secondary structures. We also examined the relationship between sequential distance and each cluster. Pairs that have the sequential distance of 3, 4, 5 were investigated thoroughly. For instance, we take the relative position data of the glutamic acid–lysine pair.

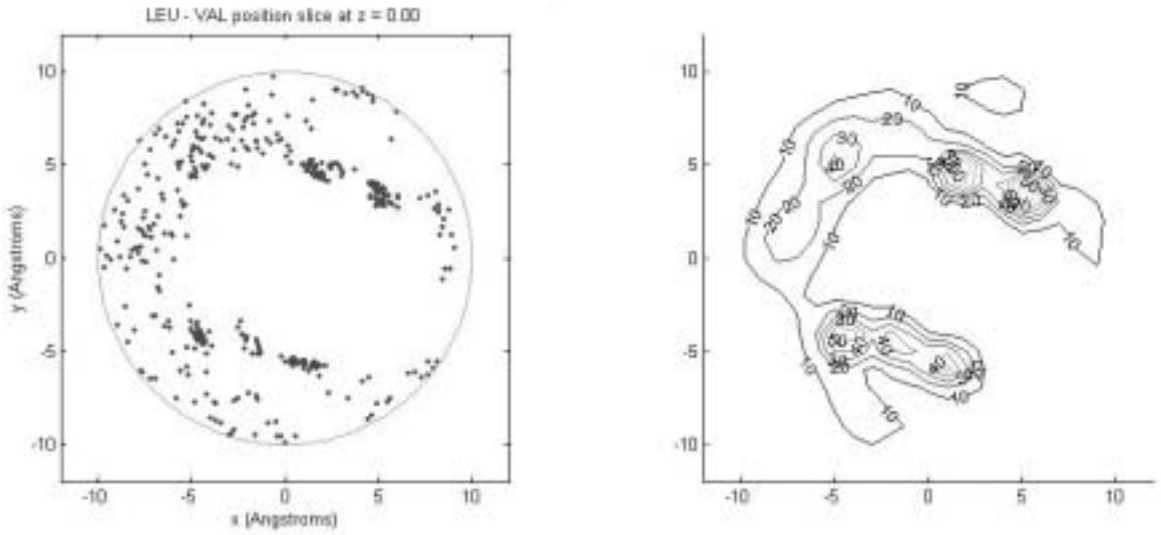
Figure 22 illustrates distribution plots of the relative position of the pair when z = 0.0. From Figure 22(a), we see that two clusters labeled as a and b exist along a circle whose radius is about 5.0 Å. Looking at Figure 22(b), we understand that the major contributors of those clusters are residue pairs within the same α helix. However, contributions of other secondary structures like 3_{10} helices or β sheets to the clusters are negligible, and thus they were not included in the plots. If we observe Figures 22(c) and 22(d), clusters a and b are mostly from pairs with the sequential distance of 4.

A set of plots in Figure 23 displays distribution plots of relative position of the pair when $z = -3.0$ Å. From Figure 23(a), we can find one bigger cluster labeled as a and one smaller cluster labeled as b. From Figure 23(b), we see that those clusters are from residue pairs within the same α helix. Looking at Figures 23(c) and 23(d), cluster a is from pairs with the sequential distance of 3 and cluster b is mostly from pairs with the sequential distance of 5.

Residue pairs in secondary structures are examined in more detail. A set of plots in Figure 24 displays distributions of relative position data of all types of pairs that are within the same α helix. Here z varies from 0.0 to 4.0 Å. We see from these plots that most clusters are concentrated in the lower left area and disappear as θn_3 grows.

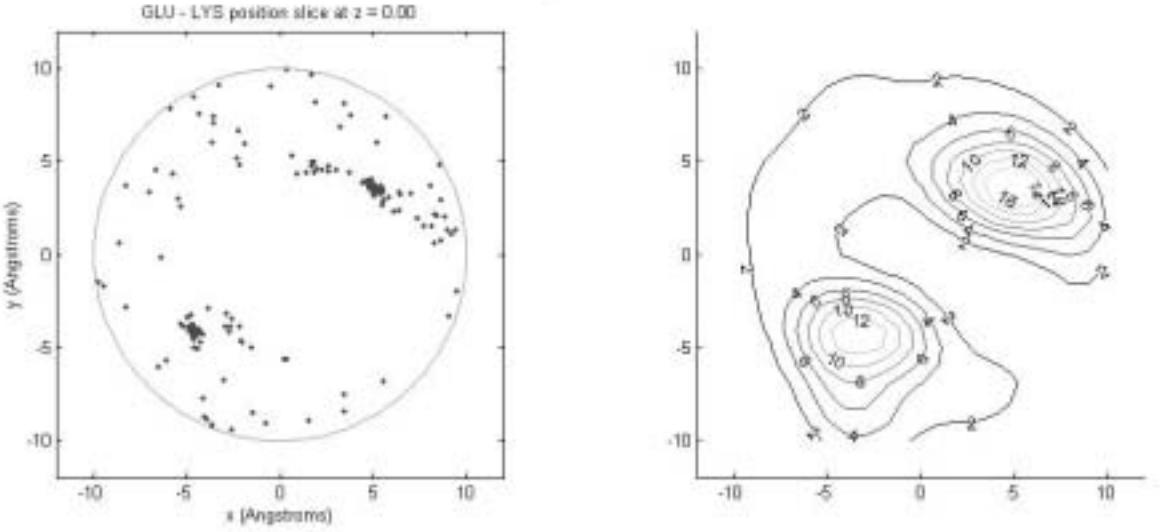
Figure 25(a) shows that the relative position data of pairs that are in different α helices is distributed widely. For other types of pairs, clusters are found in the distribution of relative position data. In particular, we observe similar patterns in the distributions of relative position of pairs in different parallel strands (Figure 25(c)) and pairs in different antiparallel strands (Figure 25(d)). We excluded pairs that are either in different 3_{10} helices or within the same parallel/antiparallel strands because their portions are negligibly small.

Residue pairs which are both sequentially and spatially proximal are now discussed. Again, the sequential distance of every pair is either 1 or 2 and the spatial distance is less than 10.0 Å. A set of plots in Figure 26 displays how the relative position data of all types of pairs are distributed as the value of z changes from 0.0 to 5.0 Å. It is notable that clusters can be separated by the sequential distance of residue pairs. In Figures 26(a), 26(b), and 26(c), the sequential distance of inner clusters is 1 and that of outer clusters is 2. Clusters with



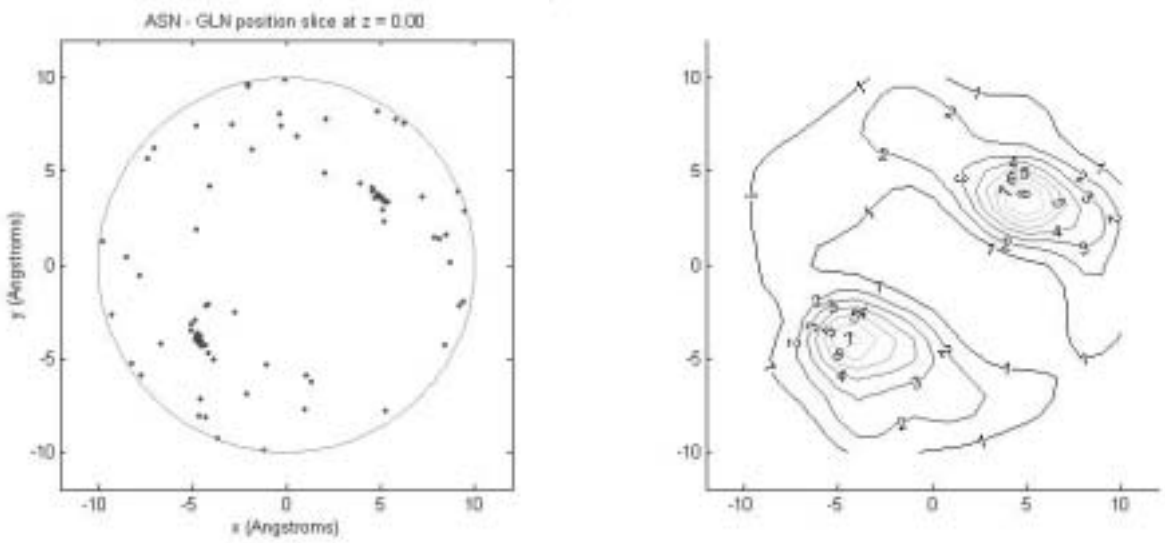
(a) (b)

Fig. 15. Distribution of relative position data of the leucine–valine pair.



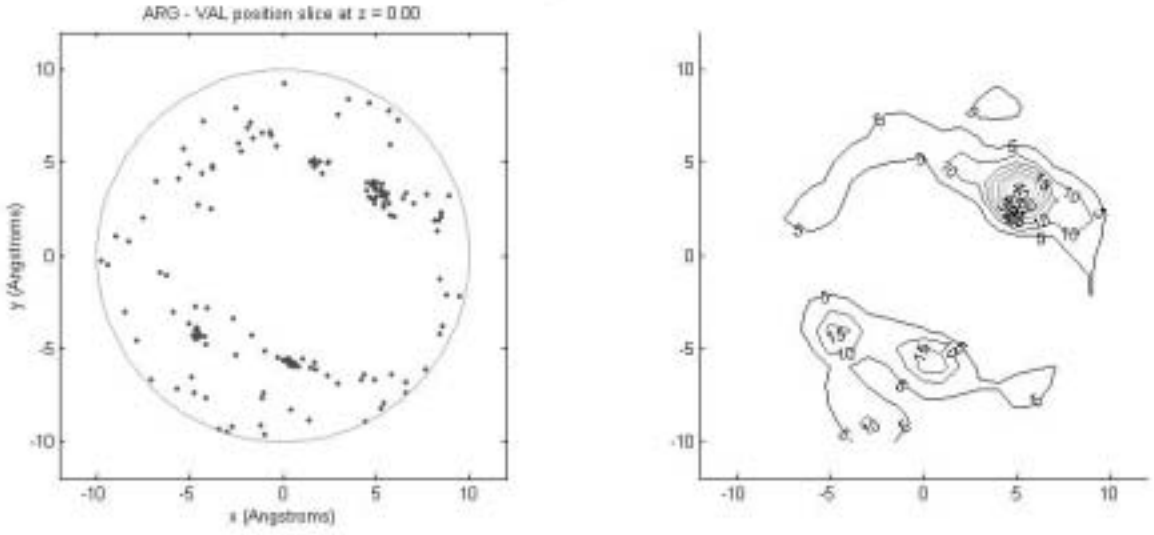
(a) (b)

Fig. 16. Distribution of relative position data of the glutamic acid–lysine pair.



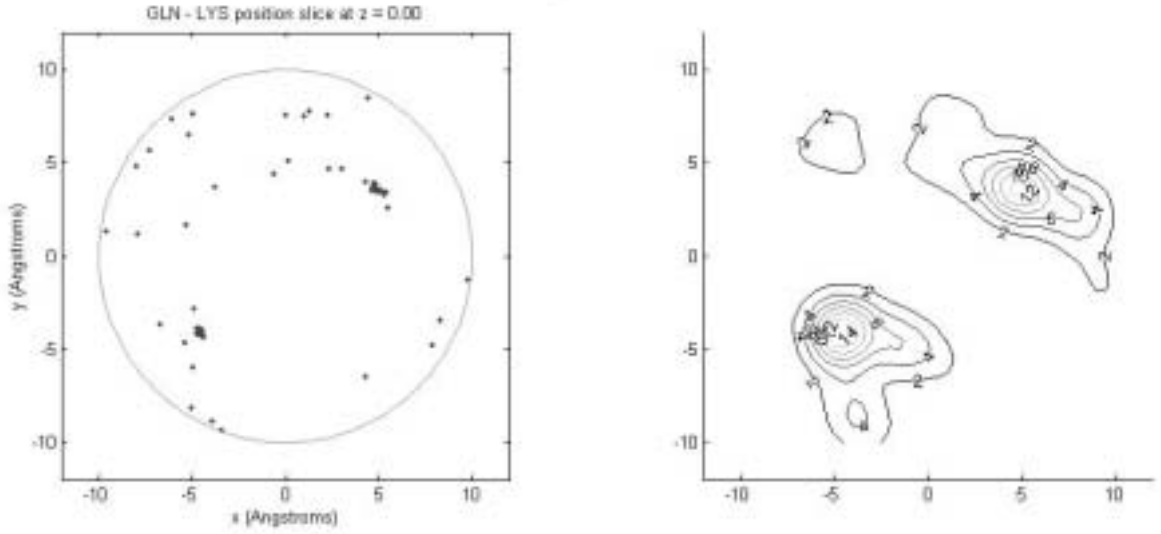
(a) (b)

Fig. 17. Distribution of relative position data of the asparagine–glutamine pair.



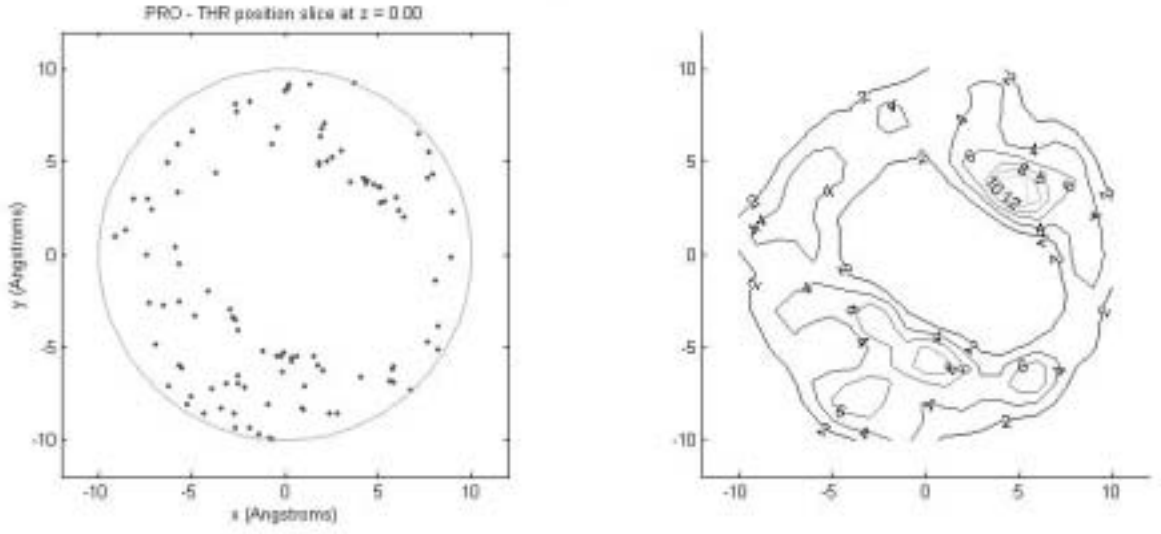
(a) (b)

Fig. 18. Distribution of relative position data of the arginine–valine pair.



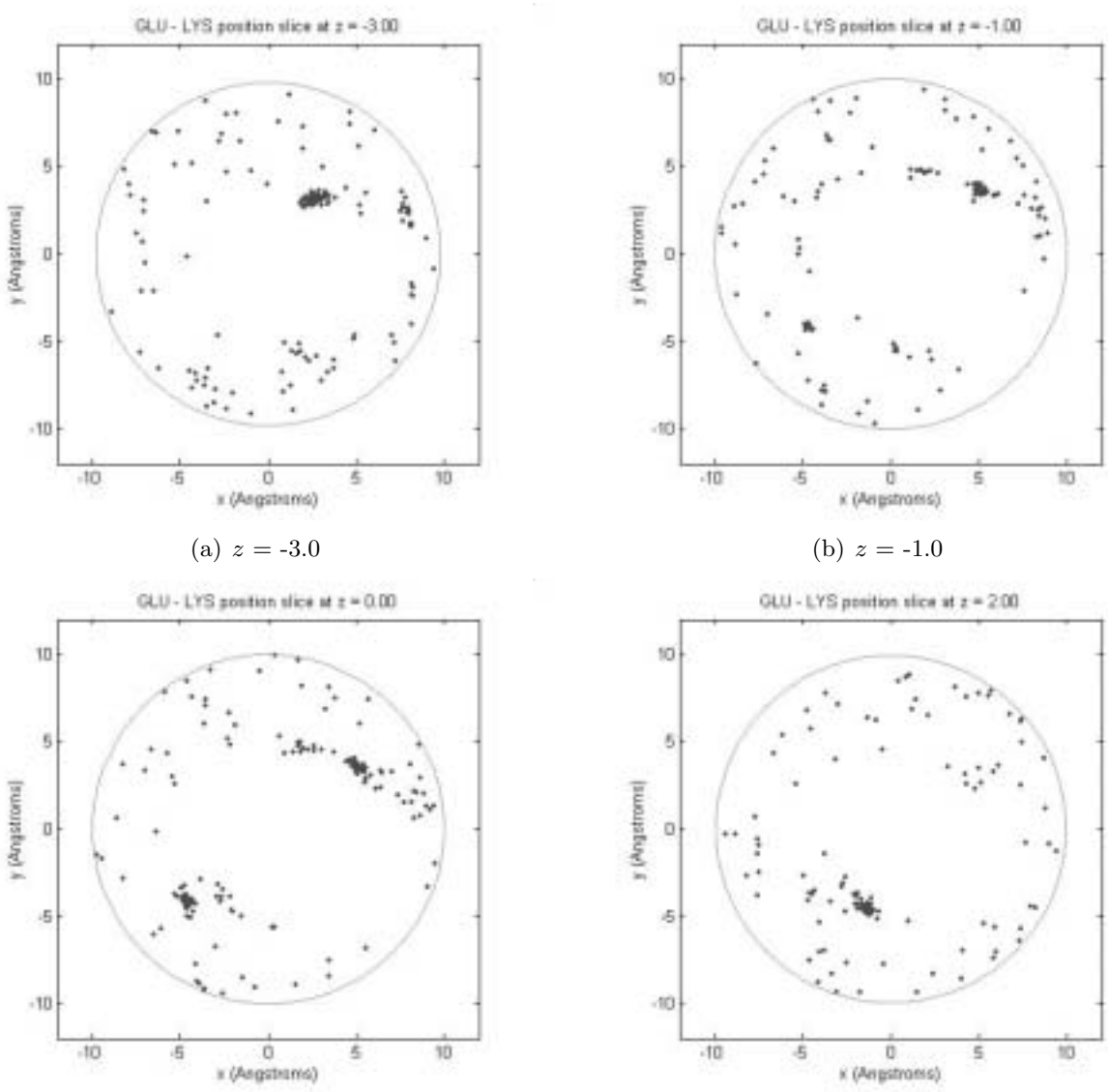
(a) (b)

Fig. 19. Distribution of relative position data of the glutamine–lysine pair.



(a) (b)

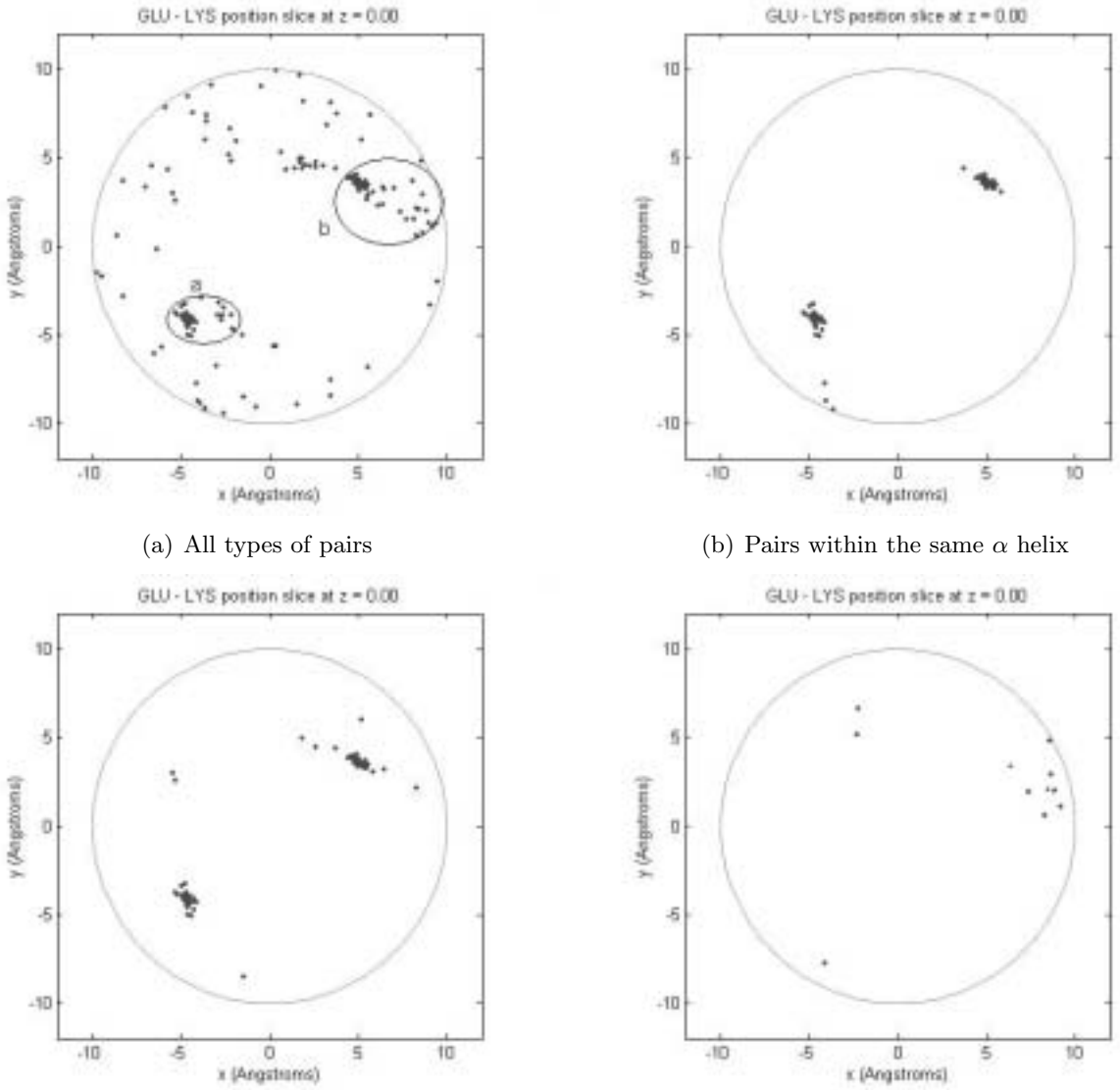
Fig. 20. Distribution of relative position data of the proline–threonine pair.



(a) $z = $ -3.0 (b) $z = $ -1.0

(c) $z = 0.0$ (d) $z = 2.0$

Fig. 21. Distribution of relative position data of the glutamic acid–lysine pair as the value of z varies.



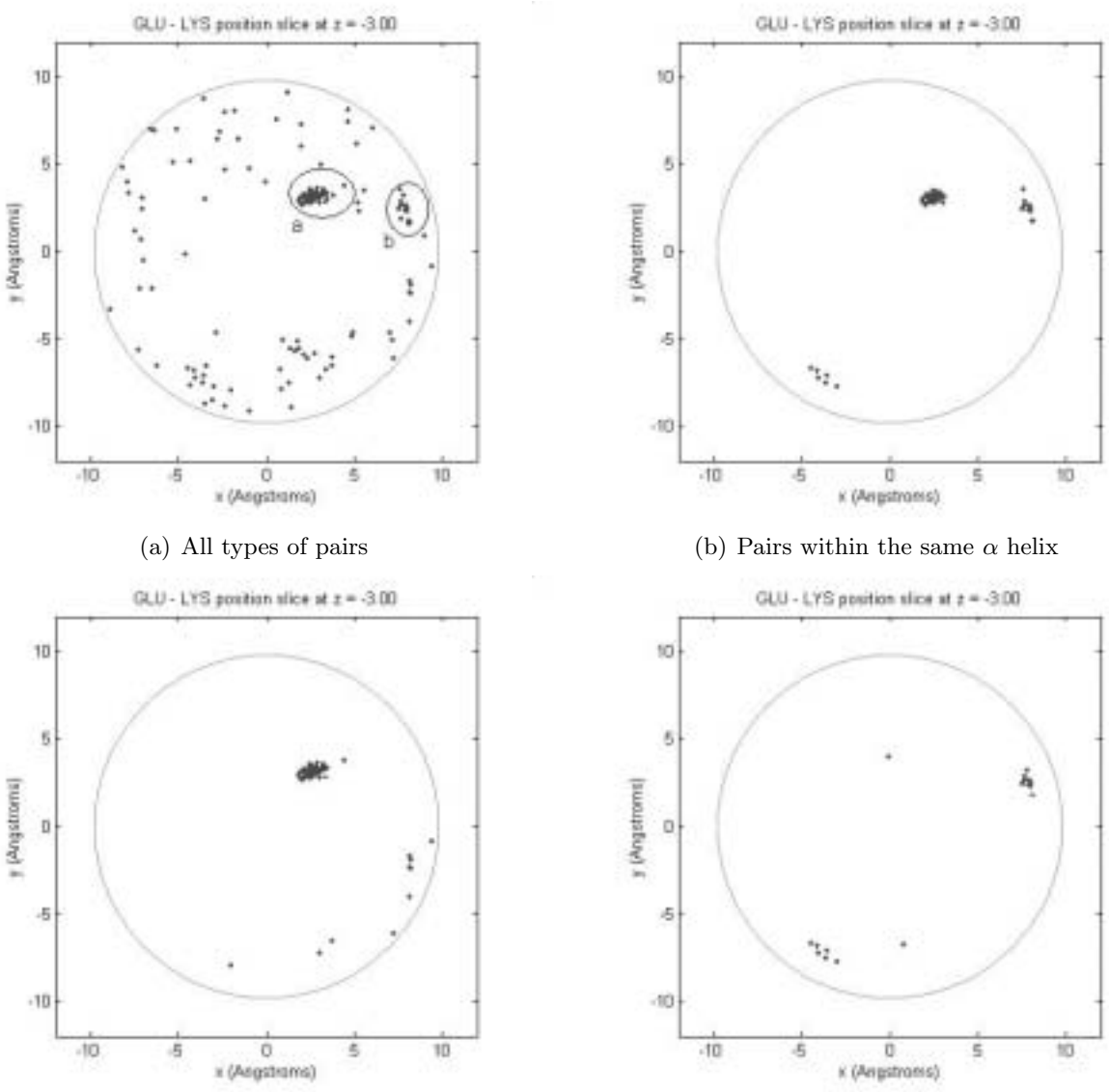
(a) All types of pairs

(b) Pairs within the same α helix

(c) Pairs with the sequential distance of 4

(d) Pairs with the sequential distance of 5

Fig. 22. Distribution of relative position data of the glutamic acid–lysine pair at $z = 0.0$.



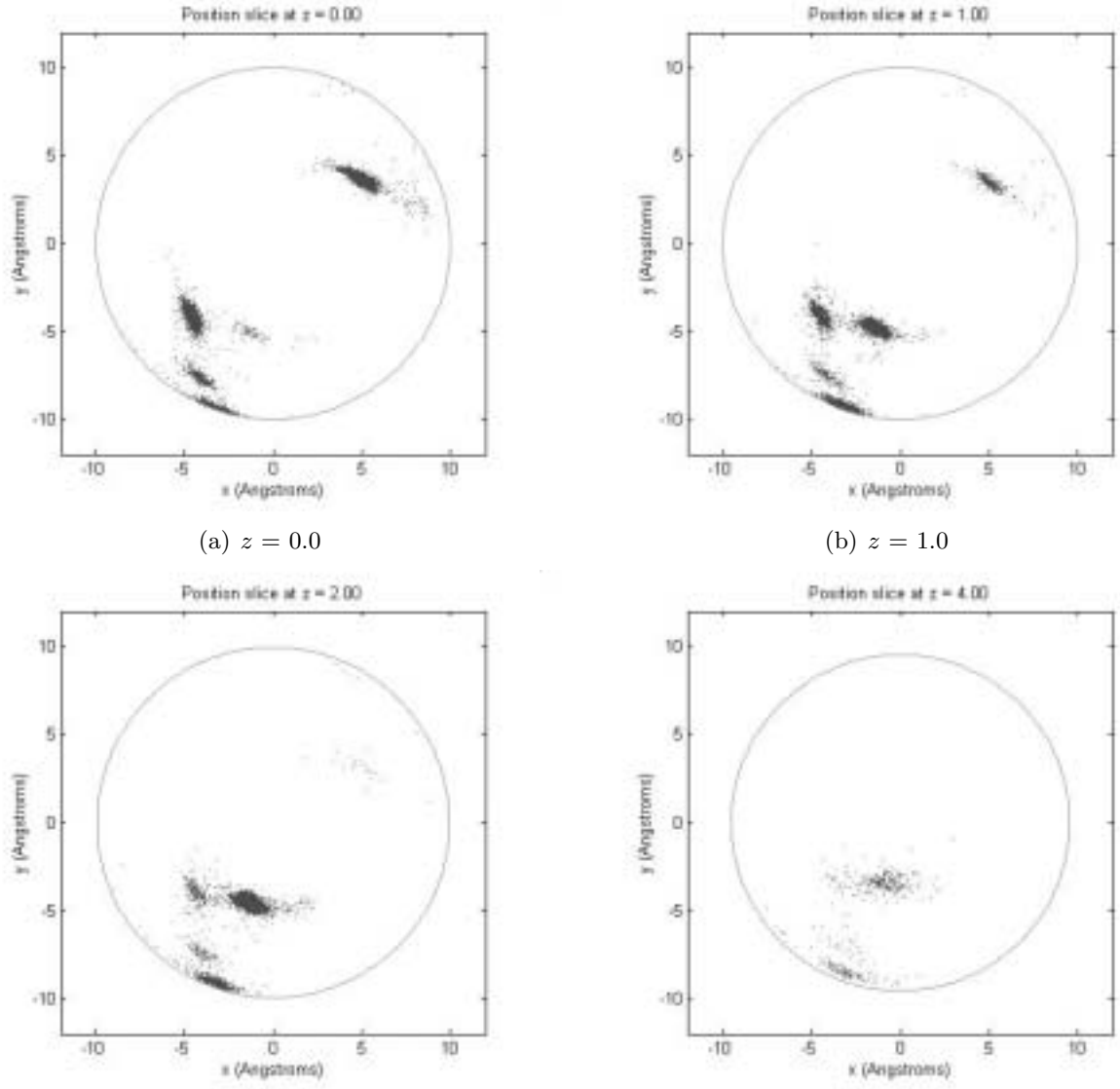
(a) All types of pairs

(b) Pairs within the same α helix

(c) Pairs with the sequential distance of 3

(d) Pairs with the sequential distance of 5

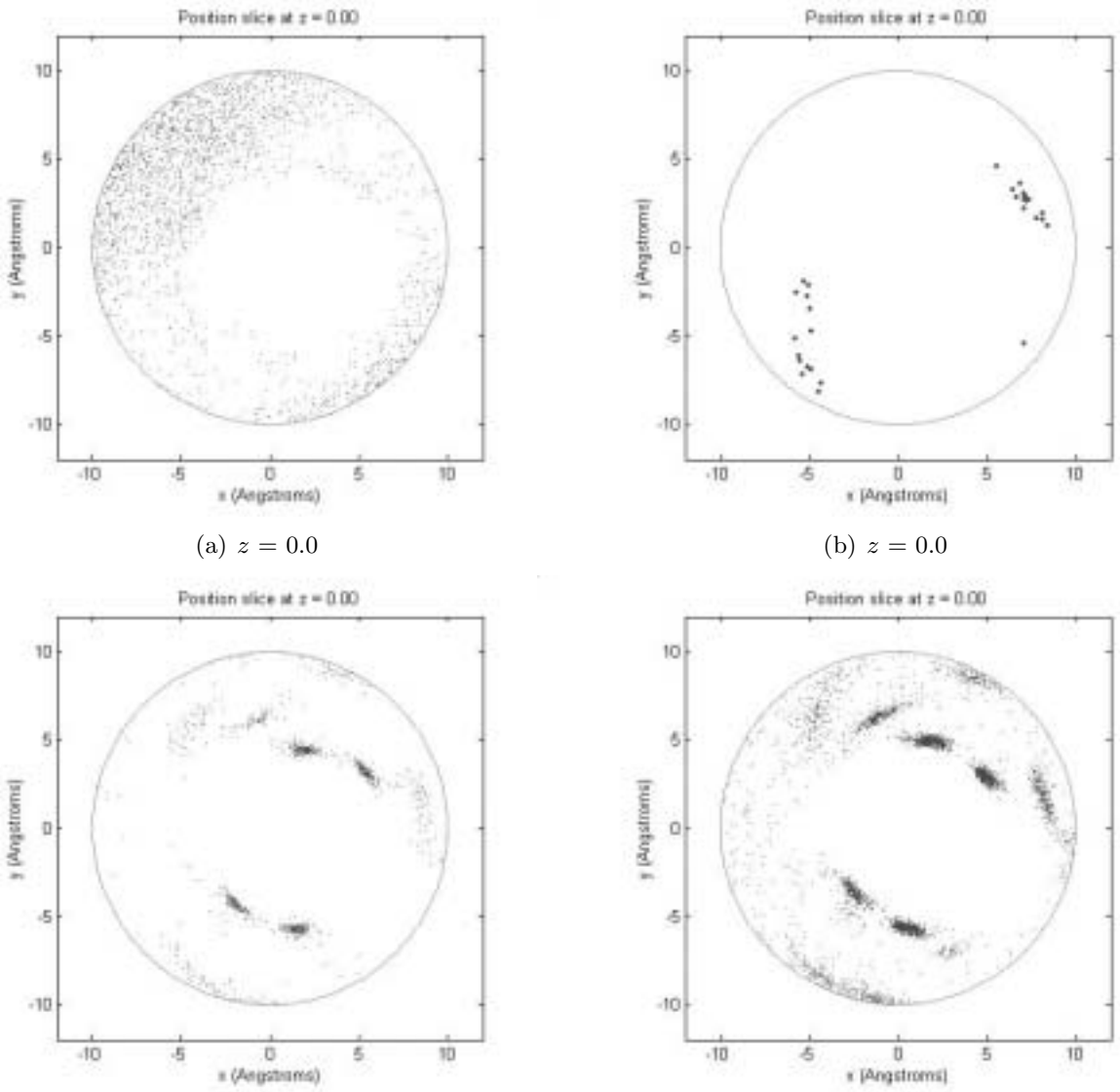
Fig. 23. Distribution of relative position data of the glutamic acid–lysine pair at $z = -3.0$ Å.



(a) $z = 0.0$ (b) $z = 1.0$

(c) $z = 2.0$ (d) $z = 4.0$

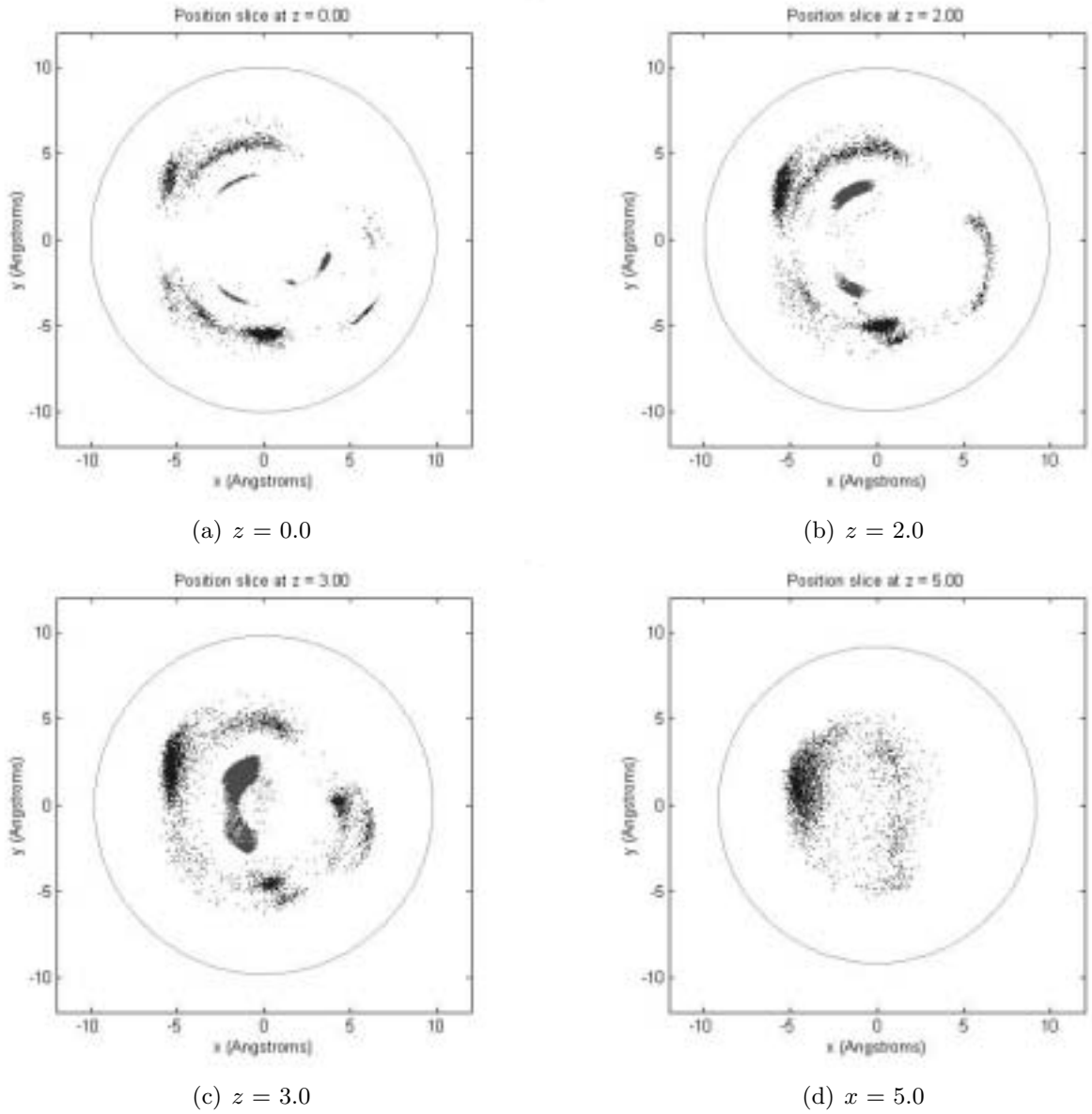
Fig. 24. Distribution of relative position data of all types of pairs that are within the same α helix.



(a) $z = 0.0$ (b) $z = 0.0$

(c) $z = 0.0$ (d) $z = 0.0$

Fig. 25. Distribution of relative position data: (a) all types of pairs in different α helices; (b) all types of pairs within the same 3_{10} helix; (c) all types of pairs in different parallel strands; (d) all types of pairs in different antiparallel strands.



(a) $z = 0.0$ (b) $z = 2.0$

(c) $z = 3.0$ (d) $x = 5.0$

Fig. 26. Distribution of relative position data of all types of pairs that are both sequentially and spatially proximal.

the sequential distance of 1 are no longer seen in the plot of $z = 5.0$ in Figure 26(d).

5. Conclusion

In this paper patterns in the relative position and orientation between alpha-carbons in proteins in the PDB were sought. Quantitative methods for characterizing such patterns may have applications in protein structure prediction and drug design. We have presented a new visualization method to describe pose distribution of amino acid pairs that are proximal in space and distal in sequence. Distribution data were visualized in the form of continuous distributions by using Gaussian distribution functions on $SO(3)$ and $\mathbb{R}^3$. Hence, we discussed how the classical Gaussian functions can be generalized to capture both positional and orientational data. The method was applied to 168 proteins in the PDB, whose resolution is 2.0 Å or better and whose R-factor is less than 20%. Two cutoff values were used so that the sequential distance of residue pairs is 3 or higher and the spatial distance of residue pairs is less than 10.0 Å.

The pose distribution for each amino acid pair type was examined, and characteristics for each group type (e.g., hydrophobic–hydrophobic) were discussed. Multiple clusters were found in many group types and sources of such

clusters in distribution plots were also discussed. In several cases, hydrophobicity and electrostatic properties of residue types are found to be important factors. For example, multiple clusters in the distribution of positional data for hydrophobic–hydrophobic pairs are due to the fact that hydrophobic residues make non-specific interactions. In the case of charged–charged pairs, preferred orientations are shown because electrostatic interactions are specific.

It was also found that residues in secondary structures, i.e., helices or sheets, made significant contributions. We examined intensively amino acid pairs with the sequential distance of 3, 4, 5. The largest parts of clusters were found to be from residue pairs within the same α helix. For comparison, residue pairs which are both sequentially and spatially proximal were investigated. Distribution plots of relative orientation data of all types of pairs were also displayed. It was found that several clusters appeared in each plot and they could be differentiated by the sequential distance. The mathematical techniques of pose analysis have been a useful tool to characterize the distribution of relative position and orientation of residue pairs.

Developing statistical potentials using our analysis of pose data in proteins will be explored in future work. This is expected to be a useful tool to develop efficient computational methods for protein fold recognition and protein structure prediction, and also for simulations of coarse-grained models of protein conformational fluctuations. The results in our paper could be helpful as a guideline for generating new models of polypeptide chains. Pose information can be extracted from new models and compared with the results in our paper. We are also interested in extending the results by applying the approach of pose analysis to side chains in proteins and also developing an effective computer software tool for three-dimensional data visualization.

Acknowledgments

The authors would like to thank the reviewers for valuable suggestions. This work was supported by the faculty research fund of Konkuk University in 2003 and was initiated while SL was a student at JHU, and while the authors were supported by the National Science Foundation under Grant No. IIS-0098382.

References

Bahar, I., and Jernigan, R. L. 1996. Coordination geometry of non-bonded residues in globular proteins. *Folding and Design* 1:357–370.

Banavar, J. R., Maritan, A., and Seno, F. 2002. Anisotropic effective interactions in a coarse-grained tube picture of proteins. *Proteins* 49:246–254.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. 2000. The protein data bank. *Nucleic Acids Research* 28:235–242.

Branden, C., and Tooze, J. 1999. *Introduction to Protein Structure*, 2nd edition, Garland Publishing, New York.

Buchete, N.-V., Straub, J. E., and Thirumalai, D. 2004a. Orienatation-dependent coarse-grained potentials by statistical analysis of molecular structural databases. *Polymer* 45:597–608.

Buchete, N.-V., Straub, J. E., and Thirumalai, D. 2004b. Orienatational potentials extracted from protein structures improve native fold recognition. *Protein Science* 13:862–874.

Buchete, N.-V., Straub, J. E., and Thirumalai, D. 2004c. Development of novel statistical potentials for protein fold recognition. *Current Opinion in Structural Biology* 14:1–8.

Chakrabarti, P., and Debnath, P. 2001. The interrelationships of side-chain and main-chain conformations in proteins. *Progress in Biophysics and Molecular Biology* 76:1–102.

Chirikjian, G. S., and Chétalet, O. 2002. Sampling and convolution on motion groups using generalized Gaussian functions. *Electronic Journal of Computational Kinematics* 1(1).

Chirikjian, G. S., and Kyatkin, A. B. 2000. *Engineering Applications of Noncommutative Harmonic Analysis*, CRC Press, Boca Raton, FL.

Kemp, J. P., and Chen, Z. Y. 1998. Formation of helical states in wormlike polymer chains. *Physics Review Letters* 81:3880–3883.

Kreyszig, E. 1999. *Advanced Engineering Mathematics*, 8th edition, Wiley, New York.

Kumar, S., and Nussinov, R. 1999. Salt bridge stability in monomeric proteins. *Journal of Molecular Biology* 293:1241–1255.

Lee, S., Fichtinger, G., and Chirikjian, G. S. 2002. Numerical algorithms for spatial registration of line fiducials from cross-sectional images. *Medical Physics* 29(8):1881–1891.

Lee, S. 2002. Pose Analysis in Image Registration and Protein Statistics. PhD Dissertation, Johns Hopkins University.

Lesk, A. M. 2001. *Introduction to Protein Architecture*, Oxford University Press, New York.

Murray, R. M., Li, Z., and Sastry, S. S. 1994. *A Mathematical Introduction to Robotic Manipulation*, CRC Press, Boca Raton, FL.

Ramachandran, G. N., and Sasisekharan, V. 1968. Conformation of polypeptides and proteins. *Advances in Protein Chemistry* 23:283–437.

Trovato, A., Ferkinghoff-Borg, J., and Jensen, M. H. 2003. Compact phases of polymers with hydrogen bonding. *Physics Review E* 67:021805.

Varshalovich, D. A., Moskalev, A. N., and Khersonskii, V. K. 1988. *Quantum Theory of Angular Momentum*, World Scientific, Singapore.