PROTEINS: Structure, Function, and Bioinformatics 60:679–689 (2005)

# Main-Chain Conformational Tendencies of Amino Acids

**Robert J. Anderson,[1,2] Zhiping Weng,[2] Robert K. Campbell,[1] and Xuliang Jiang[1]***
[1]*Serono Reproductive Biology Institute, One Technology Place, Rockland, Massachusetts*
[2]*Department of Biomedical Engineering, 44 Cummington Street, Boston University, Boston, Massachusetts*

***ABSTRACT*** **A Ramachandran plot is a visual representation of the main-chain conformational tendencies of an amino acid. Despite forty years of research, the shape of Ramachandran plots is still a matter of debate. The issue in making a Ramachandran plot based on experimental data is deciding whether sparse data represent genuine conformations. We present here a simple solution to settle the ambiguities of the sparse data, and explain how we verified the accuracies of our plots using an independent dataset. To obtain our results, we then measured the pair-wise distances of main-chain conformational tendencies among amino acids, and showed that the conformational relationships of amino acids are well preserved in a two-dimensional map, leading to the conclusion that the conformational diversity space of amino acids is largely two dimensional. We further noticed that amino acids in early and late evolutionary stages are located in different zones in the two-dimensional map. In addition to these conclusions, we here present an amino acid substitution table derived from experimental data. Proteins 2005;60:679–689.** © **2005 Wiley-Liss, Inc.**

**Key words: Ramachandran plot; folding tendency; two-dimensional map; diversity space; substitution table**

## INTRODUCTION

Since Ramachandran and Ramakrishnan published their work on the allowed conformations of the protein backbone in 1963,[1] the $\phi$–$\psi$ diagram, commonly called Ramachandran plot, has been one of the most powerful diagnostic tools used in protein structure analysis. Despite four decades of research in this area, there are discrepancies regarding the exact shape of the allowed regions. The Ramachandran plot can be generated in two ways, based on theoretical calculations or based on experimental observations. The original Ramachandran plot was constructed from steric analysis of hard-sphere amino acid models. More sophisticated calculations have been introduced over the years, but none has satisfactorily reflected the experimental observations of all twenty amino acid types.[2,3] Similarly, there has been limited success in the construction of accurate Ramachandran plots using experimental data. As the number of available protein structures has increased rapidly over recent years, many groups have attempted to produce the Ramachandran plot using the wealth of information in the Protein Data Bank (PDB).[4,5] The most widely used plot of this type is from the PROCHECK program.[6,7] This plot, however, is more than a decade old. Another plot, as produced by the program WHAT_CHECK or WHAT_IF,[8,9] is also popular, but it is dramatically different from the PROCHECK plot. Lovell et al.[10] recently generated yet another different type of plot, emphasizing the unsettled debate on this subject.

While the amino-acid conformations in the dense data area of the Ramachandran plot represent their common conformations in protein structures, it is difficult to know whether the conformations in sparse areas are rare but genuine conformations or whether they are errors in structural determination. Ideally, the question would be settled by inspection of the experimental electron density to reveal whether a certain unprecedented conformation could be a genuine feature of a structure or is more likely to be an error in the model. In practice, however, the electron density might be too weak or ambiguous to make a good judgment, and a more "normal" conformation could explain the experimental data equally well. The conundrum is how to define the normal conformations. We decided to investigate the issue with a large number of data to establish the statistical significance of the result and to cross-validate its reliability with an independent dataset.

The comparison of Ramachandran plots is another outstanding issue in this field. The shapes of $\phi$–$\psi$ angle distributions in Ramachandran plots are compared by eye to determine the similarities of main-chain conformational preferences of amino acid residues.[10] The comparison by eye, however, is subjective and lacks the rigor of quantitative analyses. Another complicating factor besides the eyeball comparison of the shapes of $\phi$–$\psi$ angle distributions is the varying height of observations over a large number of pixels. When both factors are considered, comparison of the plots becomes very complicated, as there are 190 permutable pairs of 20 amino acids.

The initial motivation for this work was to develop a method to prioritize candidate residues for mutations in protein engineering projects. Protein engineers mutate residues seeking to change desired side-chain properties with the assumption that the backbone structure will not

---

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

*Corresponding author: Xuliang Jiang, Serono Reproductive Biology Institute, One Technology Place, Rockland, MA 02370. E-mail: xuliang.jiang@serono.com

Received 26 September 2004; Revised 3 March 2005; Accepted 3 March 2005

Published online 14 July 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20530

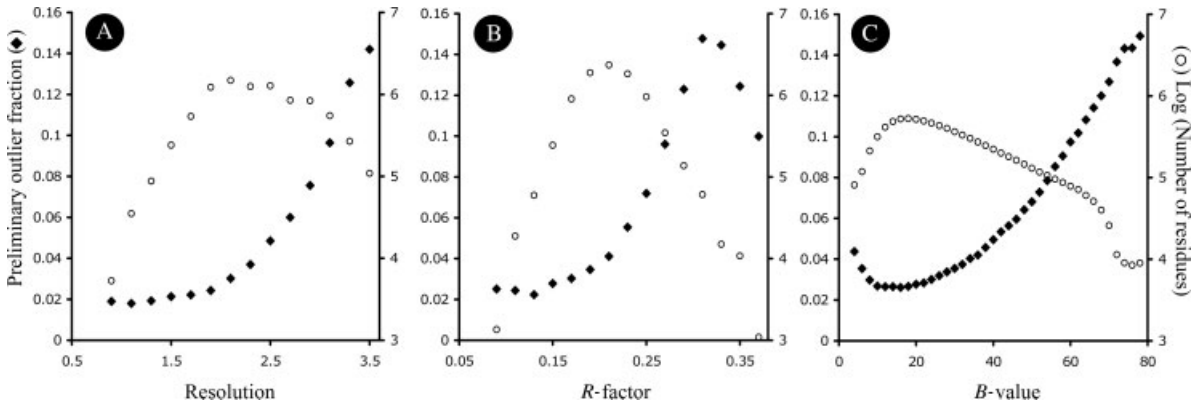


Fig. 1. Parameter selection: (**A**) resolution, (**B**) *R*-factor, and (**C**) average backbone temperature factor *B*.

change. However, a single mutation may have adverse effects on the backbone structure, which in turn may affect its activity, folding pathways, and expression level. We believe that it is as important for the new residue type to maintain the same or similar main-chain conformation of the wild-type residue as to adopt the desired property of the side chain. With this goal in mind, we set out to measure the similarity of main-chain conformations quantitatively, since they are defined by $\phi$ and $\psi$ dihedral angles of each amino acid residue.

## MATERIALS AND METHODS

### Database Creation

A collection of 21,085 protein structures was downloaded from the May 2003 release of the PDB. To facilitate the efficiency of information usage, we constructed a relational database using Oracle 8.1.6. Data derived directly from but not included in PDB files (dihedral angles, distances between $\alpha$-carbons, etc.) are also included and indexed. The SCOP 1.63 classification scheme is included in the relational database,[11] as are Astral-derived lists of SCOP domains with varying degrees of sequence identity.[12] PDB data were divided into three levels of organization: atomic information; residue-specific information, and structure-specific information. Within each residue record, the residue *B*-value was defined as the average temperature factor of the backbone atoms: $C_\alpha$, C, O, and N.

### Crystallographic Parameter Selection

For the preliminary stage, we mimicked Klewegt and Jones' approach[13] by selecting the residues that met all of the following criteria:

1. They are listed in the Astral-derived SCOP95 domains.
2. They are non-glycine.
3. The have resolution better than 2.0 Å.
4. They have chain length more than 20 amino acids.
5. They have residue *B*-value greater than 1.0 and less than two standard deviations above the mean *B*-value of the whole amino acid chain.
6. The standard deviation of *B*-values among the backbone atoms $C_\alpha$, C, O, and N is larger than 0.

The $\phi$–$\psi$ angle data from the selected residues were binned into 1296 $10 \times 10°$ pixels over the range from $-180$ to 180°. The pixels were then classified into two categories: main body and preliminary outlier. Based on the criteria of a previous study,[13] we considered the most populous pixels contributing 98% of the $\phi$–$\psi$ data to be the main body and those contributing the remaining 2% to be preliminary outliers.

For our analysis of the crystallographic parameter selection, we selected the $\phi$–$\psi$ angle data from residues of the crystal structures in the PDB that meet all of the following criteria:

1. They are non-glycine.
2. They have non-powder diffraction data.
3. They have resolutions ranging between 0.5 and 3.5 Å.
4. They have *R*-factors ranging between 0.03 and 0.43.
5. They have residue *B*-values larger than 1.0.
6. The standard deviation of *B*-values among the backbone atoms is larger than 0.

We then categorized each residue as an outlier or part of the main body according to the $\phi$–$\psi$ angle map generated by the Kleywegt-mimic set.

For each parameter (resolution, *R*-factor, and residue *B*-value), the $\phi$–$\psi$ data were binned into bin sizes of 0.2 Å, 0.02, and 2 Å$^2$, respectively. The first bin, however, contained all acceptable data up to 0.9 Å resolution, 0.09 *R*-factor, and 4 Å$^2$ *B*-value. The outlier fraction for a particular parameter value was defined as the ratio of preliminary outlier to main-body data, and was then plotted against each of the three parameters. A cutoff value for a given parameter was selected at the preliminary outlier fraction of approximately 0.05 (Fig. 1).

### Data Selection and Ramachandran Plot Generation

We applied three types of filters in the data selection process. The first type of filter selected for crystal structure quality, and the parameters for selection were determined by the abovementioned method, in addition to the requirement of *B*-values larger than 1.0 and standard deviation of *B*-values among the backbone atoms larger than 0. The

second type of filter was geometric. We made distinctions between *cis* and *trans* amino acid orientations using ω angles (0 ± 60° for *cis*, 180 ± 60° for trans) and $C_\alpha$–$C_\alpha$ distances (2.0-3.6 Å for *cis*, 3.0-4.6 Å for *trans*). We also selected residues with proper chiral volumes (in the range of 0.22 and 0.62 Å$^3$; chiral volume is defined as the vector product of three bonds centered at the $C_\alpha$ atom: $C_\alpha$—N, $C_\alpha$—C, and $C_\alpha$—$C_\beta$. L-amino acids have positive values while D-amino acids have negative ones). The third type of parameter was used to remove redundancies and over-represented structural families. Only residues found within 4383 peptide chains in the Astral-derived SCOP40 dataset were included in this study.

The ϕ–ψ angles of the residues that met all of the above criteria using Boolean AND operators via Structured Query Language (SQL) statements were then grouped according to their associated residue types in order to generate the corresponding Ramachandran plots. Each plot was divided into 10 × 10° pixels and contours were drawn based on δ values using Microsoft Excel 2000 software. The $\delta_i$ value at pixel $i$ is defined as $\ln(N_i/\mu)$, where $N_i$ is the data count at pixel $i$ and μ is the expected value, which is the ratio of the total number of data points to the total number of pixels. The term $\delta_i$ is related to the free energy of the conformational state at pixel $i$.

We used δ values as the criterion to accept data at each pixel. If the δ value was less than −4, the number of observations (less than 0.001% of the total number of observations) was very small, and the data in that pixel were not accepted. Pixels with δ values equal to or greater than zero were counted as significant for the conformation. Pixels with δ values in between were subject to cross-examination with neighboring pixels. If more than one neighboring pixel had δ values equal to or greater than −4, the pixel in question was considered to be significant. However, if all eight neighbors of a pixel had δ values greater than −4, the pixel was considered significant, regardless of its population.

To verify the accuracy and robustness of our plots, we created an independent set of sequences to mimic the published Astral-derived list of SCOP40 sequences[12] and generated Ramachandran plots from the independent set. We removed the Astral SCOP40 sequences from the PDB and selected a set from the remaining PDB sequences such that no two sequences had more than 40% identity. We filtered the coordinates of the residues in the independent set using the previously mentioned crystallographic quality and geometric constraints, generated its Ramachandran plot, and compared the plot to that from the original SCOP40-filtered dataset to detect any sequence bias. We also generated additional Ramachandran plots from these sets without the constraints of the crystallographic and geometric parameters to determine the effects of data quality on the plots.

### Distance and Similarity Measurement

To perform a direct measurement of similarities and dissimilarities of different Ramachandran plots, we employed a city-block distance method[14] to calculate the difference between the conformational preferences of two amino acid residues. For those pixels with no data points, we assigned their δ values to −8 (corresponding to the data occurrence less than 0.0002% of the total number of data points). We also repeated the distance calculation with values other than −8 in order to resolve the dependence of the distance calculation on this value. The formula to calculate the city-block metric distance score between residue $X$ and residue $Z$ is,

$$D_{xz} = \sum_i |\delta_{i,x} - \delta_{i,z}| \quad (1)$$

where $\delta_{i,x}$ is the δ value of residue $X$ at pixel $i$, and $\delta_{i,z}$ is that of residue $Z$ in the corresponding pixel. Since −δ is proportional to free energy and free energy is directly related to folding tendencies, $D_{x,z}$ is the gross difference of folding propensity of amino acids $X$ and $Z$.

### Clustering and Two-Dimensional Mapping of Conformational Relationships Between Amino Acids

To visualize the main-chain conformational relationships of all amino acids, we first calculated the $D_{X,Z}$ for all pairs of residues and built a 20 × 20 distance matrix. The matrix was used to cluster the residues using the complete linkage algorithm.[14]

To illustrate the structural relationships among all amino acids in the form of a two-dimensional map, we employed a simplex method with simulated annealing to reduce the 20 × 20 distance matrix into a two-dimensional map.[15] The target function to be minimized was the sum of the absolute difference between the original distance matrix and the derived distance matrix from the two-dimensional map.

## RESULTS AND DISCUSSION

### Determination of Cutoff Values of Crystallographic Parameters

In the interest of selecting quality data for our analysis, we used three crystallographic parameters–resolution, *R*-factor, and *B*-value–as the filters to determine the quality of a protein structure or the reliability of the coordinates of each residue. In order to determine whether a sparse area represents a low occurrence of a normal conformation or warrants further investigation of a bizarre conformation, we sought to explore the limit of the parameters to include all high-quality data so that the total amount of data would be large enough to make a statistically significant distinction. Our survey into various studies on this subject showed that different investigators applied quite different cutoff values in the data selection process. We decided to investigate the relationships between the cutoffs and outlier fractions.

We first mimicked Kleywegt and Jones' work[13] by treating the most populated ϕ–ψ pixels comprising 98% of selected residues (see Materials and Methods for details) as the main body of the plot and those contributing the remaining 2% as the preliminary outliers. We then plotted the fraction of outliers of the whole PDB dataset (see

Materials and Methods) against each of the three crystallographic parameters (Fig. 1). As we expected, the preliminary outlier fraction generally increased, as resolution became worse or *R*-factor and *B*-value increased. At high resolutions, low *R*-factors, or *B*-values, the outlier fraction approached the plateau at the expected limit of 0.02. However, the outlier fraction increased for data points with *B* less than 8. Our investigation into this abnormal upward trend indicated a higher fraction of residues determined at a poor resolution (>2.5 Å) at low *B*-factor range ($B < 8$) than that at the intermediate *B*-factor range ($B \approx 8$–22), and the fraction of poor resolution increases again at higher *B*-factor range ($B > 22$). In other words, the trend of the preliminary outlier fraction in the *B*-factor figure reflects the percentage of residues determined at poor resolution. We also noticed that the outlier fraction decreased as the *R*-factor increased after 0.30. This observation is consistent with the idea that as data quality deteriorates, the influence of the input model outweighs that of the crystallographic data. We settled a cutoff value for a given parameter at the preliminary outlier fraction of approximately 0.05 where a significant increase of the outlier fraction was noticed for the parameter beyond the cutoff (Fig. 1). The cutoffs for resolution, *R*-factor and *B*-value were measured as 2.5 Å, 0.23 and 38, respectively.

### Data Selection

Resolution, *R*-factor and *B*-value alone do not sufficiently reflect various aspects of protein structure quality. In addition to these three crystallographic filters, we included a second set of filters in the data selection process to assess data quality in terms of the geometry of chemical bonds related to a residue. The geometric filters are chiral volume, $C_\alpha$–$C_\alpha$ distance and ω angles. Unlike the three aforementioned crystallographic parameters that are direct experimental measures of structure quality, the geometric parameters measure the quality of crystallographic coordinates indirectly. It is difficult to judge whether coordinates with unusual geometry are of high quality. To avoid a biased selection of the structures of average geometry, we only excluded residues whose geometric values were obviously outside the acceptable range–more than six standard deviations from the mean value. We also only selected L-amino acid residues for this study because there were not enough D-amino acids for a valid analysis for our purposes. The acceptable geometries are chiral volume between 0.22 and 0.62 $Å^3$, $C_\alpha$–$C_\alpha$ distance between 3.0 and 4.6 Å for *trans*-peptide bonds or between 2.0 to 3.6 Å for *cis*-peptide bonds, and ω torsional angle in the range of 180 ± 60° for *trans*-peptides and 0 ± 60° for *cis*-peptides.

To eliminate repetitious data due to the over-representation of some structure types in the PDB, we applied a third type of filter in the data selection process. Most previous studies applied percentage of sequence identity as this type of filter. The cutoff values vary dramatically from as low as 25% to as high as 95%.[16,17] The danger of a low percentage cutoff is that it may exclude some rare conformations that present only in one type of structure family. They are, nevertheless, genuine if they have been determined independently. The problem of a high percentage cutoff, on the other hand, is that the conformations of overly represented structures might overshadow those of rare structures, and these conformations in the PDB might be excluded as the outliers due to their rare occurrence. As it is difficult to discern from the information in the PDB what method was used to phase the structure, we decided to use a percentage identity cutoff that would ensure the structures in our selected set are independently phased from each other. We resolved to only include residues in the Astral-derived SCOP40 dataset, since it is representative of almost all protein structure families.[11] In addition, the 40% identity cutoff is compatible with our concerns about including independently phased structures.

We selected 377,428 amino acid residues that passed all three types of filters to constitute the fully filtered set, which was used to populate our Ramachandran plots. Due to the distinct geometries, we treated Ramachandran plots of residues with *trans*-peptide neighbors separately from those with *cis*-configuration. The numbers of residues with *trans*-peptide bonds on both sides are: Ala, 32,409; Arg, 18,369; Asn, 16,243; Asp, 21,346; Cys, 5,597 (thionyl: 3,974; disulfide: 1,623); Gln, 13,643; Glu, 23,947; Gly, 28,303; His, 8,853; Ile, 22,707; Leu, 35,124; Lys, 20,579; Met, 8,023; Pro, 16,400; Phe, 15,913; Ser, 21,703; Thr, 21,241; Trp, 5,676; Tyr, 14,023; Val, 28,329; pre-Pro (all residues preceding proline in the sequence, excluding Pro and Gly), 14,909; non-Gly and non-Pro residues, 332,722. There is only one case (PDB code 1F0I) of two consecutive residues with *cis*-peptide bonds in the selected set, and both residues are proline. The number of *cis*-Pro residue is 894. Including proline, the number of residues in *cis*-configuration in our dataset is 1005.

### Ramachandran Plots

We separated normal conformations from bizarre ones using a statistical approach (as detailed in Materials and Methods) with a δ value, which is the natural logarithm of odds of an observation. We also considered the minimum required number of data points to be considered as statistically acceptable for each plot. For a total of 1296 10 × 10° pixels per plot, the minimum amount of data required for statistical significance is the point where we can begin to distinguish whether a single observation is acceptable. Since the single observation at a pixel has to be below the expected value to subject the sparse data to a cross-confirmation test, the expected value has to be larger than one so that the test will be meaningful. Therefore, the minimum number is 1297 for any residues except proline, although more data would naturally allow us to render a higher quality plot. For proline, a positive ϕ angle is not possible; hence the minimum number for proline is 649, half of the minimum number of other residues. All residues with *trans*-configurations as well as *cis*-proline met the considered minimum requirement. Other *cis*-residues were not abundant enough to meet the requirement and were not considered in this study.

The distribution of the data in each plot (Fig. 2) was contoured and shadowed based on the δ values using the

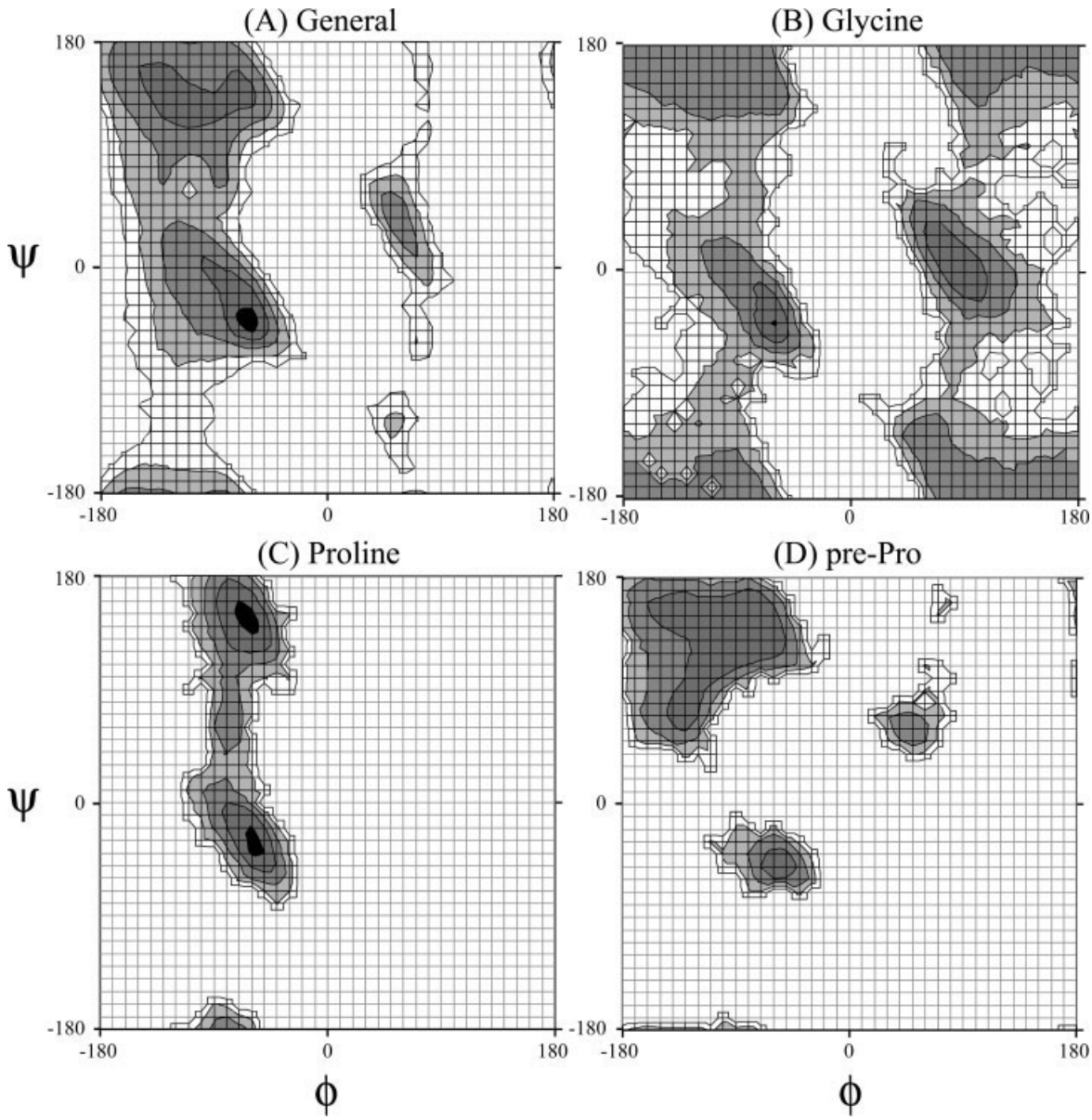


Fig. 2. Ramachandran plots of L-amino acids in *trans*-configuration from the filtered dataset. All observed dihedral angles were grouped into 10 × 10° pixels. Each pixel is gray-coded according to its δ-value: $<-4$ (white, gray grid); $\geq -4$ (white, black grid); $\geq -2$ (light gray); $\geq 0$ (gray); $\geq 2$ (darker gray); and $\geq 4$ (black). The free energy difference between neighboring contours is 1.2 kcal/mol at room temperature. (**A**) Non-Gly and non-Pro residues. (**B**) Glycine residues. (**C**) Proline residues. (**D**) Pre-Pro residues (excluding Pro and Gly).

standard Microsoft Excel 2000 software. The data-occupied areas are gray-shadowed gradually from black to white, representing the δ value being equal to or greater than 4, 2, 0, −2 and −4, respectively, and the δ value for the unoccupied-data area is less than −4. Figure 2 shows the Ramachandran plots for four distinct L-amino acid residue types of a general non-Pro and non-Gly, Pro, Gly, and pre-Pro residues in *trans*-configuration. For the non-Pro, non-Gly residues in *trans*-configurations, these contour levels include 24.0, 66.1, 94.8, 99.1, and 99.9% of the data cumulatively, and correspond to 0.2, 3.6, 12.7, 23.1, and 37.7% of the entire plot area, respectively. A complete set of Ramachandran plots can be found in the supplemental materials, and are also available at http://zlab.bu.edu/rama. Users can submit their PDB-formatted files to the web page and generate their own Ramachandran plots.

Our Ramachandran plots differ from others in two aspects: shape and definition of zones of classification (Fig. 3). Our map with δ value of −4 or greater covers only 38% of the entire plot and contains 99.9% of the total data. In contrast, the conformationally "allowed" region in the PROCHECK plot covers a much larger area (70% of the entire plot) but contains a smaller percentage of data. The "allowed" region in WHAT_CHECK plots is so small (11% of entire plot) that a significant number of high-quality data have been excluded. The contour shapes of our plots

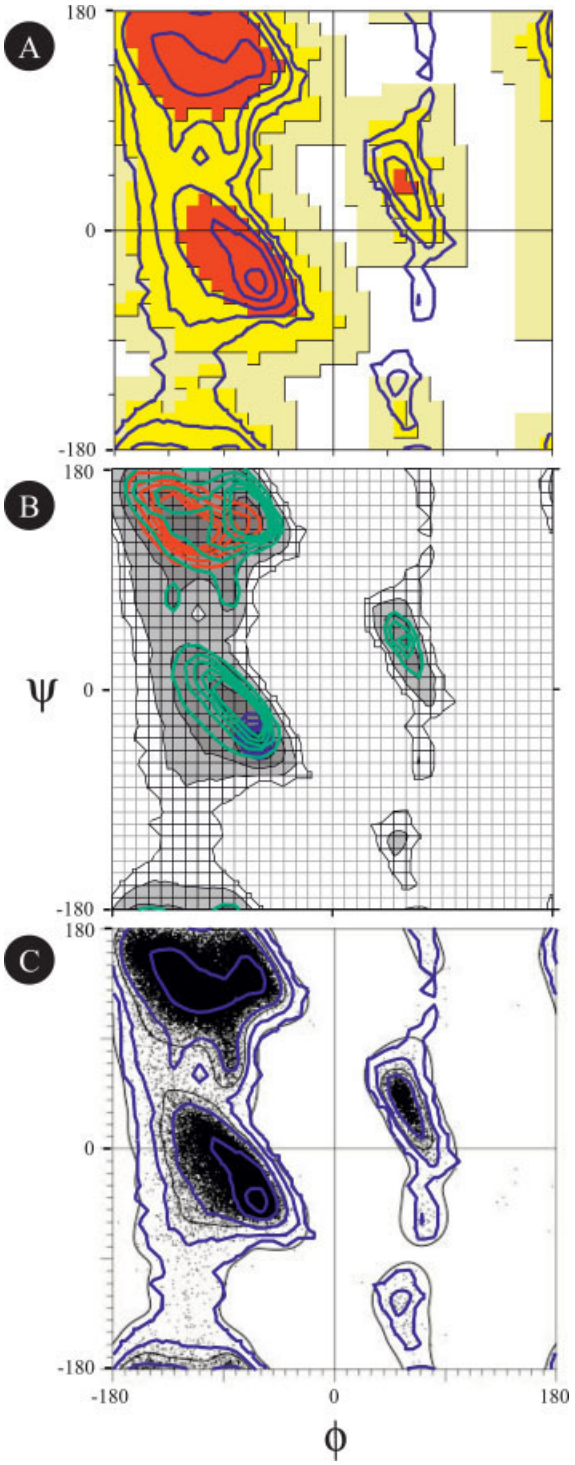


Fig. 3. Comparisons of Ramachandran plots. These figures depict our results from the general case superimposed with results from previous work. (**A**) PROCHECK Ramachandran with our δ contours superimposed in blue. (**B**) Our grayscale Ramachandran plot with WHAT_CHECK contours superimposed, shown in red for β-strand, blue for α-helix, and green for others. (**C**) Lovell and coworkers' smoothed density contours and their selected data points shown in black with our δ contours superimposed in blue.

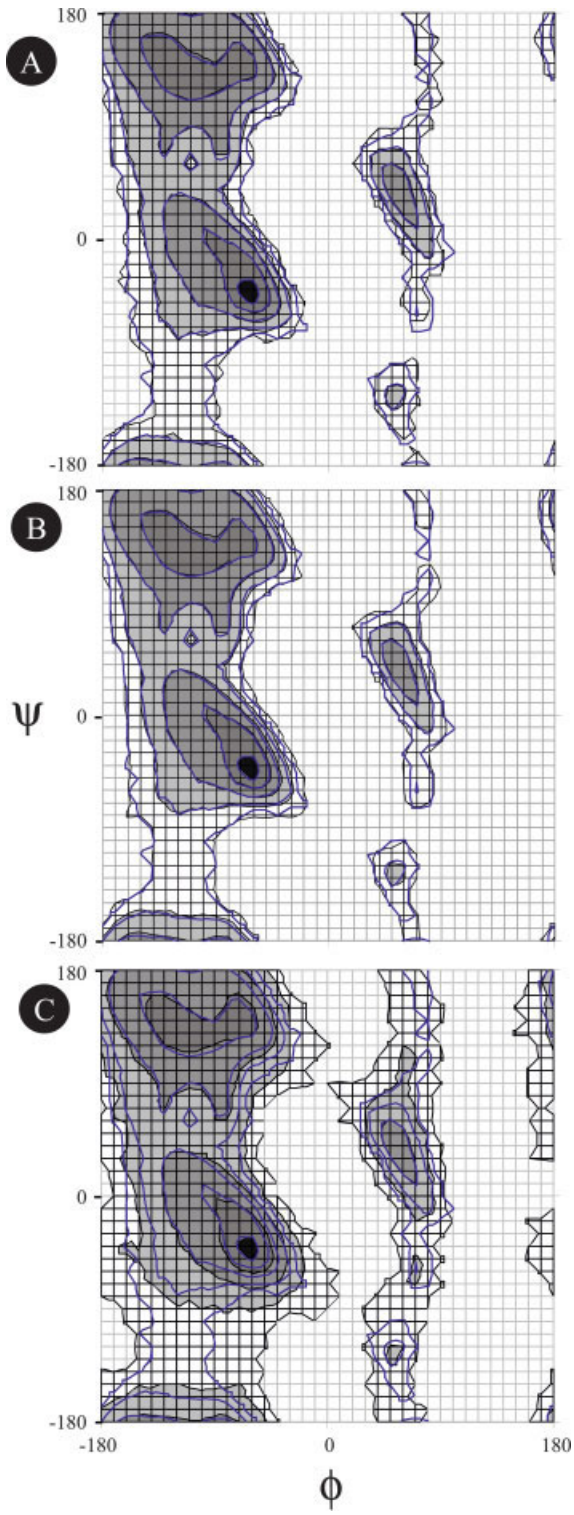


Fig. 4. Internal validation. In all of these figures, the fully filtered results are shown as blue lines superimposed on other selection sets. (**A**) The independent set was selected using crystallographic, geometric, and sequence identity filters. (**B**) The remaining set was selected using crystallographic and geometric filters, and no sequence filter. (**C**) The unfiltered set used no crystallographic or geometric filters but the SCOP40 sequence filter.

agree very well with Lovell and coworkers' plots.[10] The discrepancy in Figure 3(c) in contours and data near (−130, 80) between this work and Lovell and coworkers'

work[10] is because the latter omitted pre-Pro from the general-case plot. The precision of our plots is expected to be 10°, and the Lovell contours fall within this range. An exception is at the region around (75, 105) which is not within their "allowed regions" despite the obvious population of data in their selection. Unlike the somewhat complicated density masks used in their paper, contour shapes in our plots can be simply produced with pixel data counts and a standard Microsoft Excel package. Our plots, which are derived from a much larger amount of data, enable us to generate quality Ramachandran plots for individual residue types carrying statistical merit at each pixel.

We do not explicitly label any area in the plot as conformationally allowed or disallowed. Some high-energy conformations, while rare, may be genuine, and they may be explained by additional energy compensations (e.g., hydrogen bonds) in the local environment. Our δ values reflect the energy term of the conformation. Nevertheless, any conformations located in the areas of δ value less than −4 and at least 10° away from the −4 line are very suspicious, unless the conformations are confirmed by additional local atomic interactions or by strong electron density map.

The accuracy and robustness of our Ramachandran plots were verified internally and externally. The external validation came from a comparison with the plot by Lovell et al.[10] Although our Ramachandran plot is very similar to their plot, we felt more validations needed to be performed, given the very different nature of PROCHECK and WHAT_CHECK plots. We performed an internal validation by comparison to Ramachandran plots generated from an independent set. A set of 369,962 residues was selected from the PDB and filtered with the filters for quality, geometry, and 40% sequence identity. No residues in this set overlap with any in the above-mentioned SCOP40 set. As shown in Figure 4(a), the Ramachandran plot for the independent set is remarkably similar to the plot generated from the Astral-derived SCOP40 set, and the differences in the shape of the contours are about 10°, the size of a single grid cell. This high level of agreement demonstrates the lack of bias in the plots and credits their accuracy.

To assess the effect of the sequence identity filter on the shape of the plot, we applied the crystallographic and geometric filters to all structures in the PDB, and called the resulting list the remaining set. The remaining set contains over $3.7 \times 10^6$ residues, excluding the residues in SCOP40. The general shape of the plot of the free set is highly similar to that from the fully filtered dataset [Fig. 4(b)]. The high level of agreement between the unrestrained plot of the remaining set and our more conservative plot of the filtered set implies that the sequence identity filter plays little role in improving the plot quality. The number of structures in the PDB is so large that over-sampling likely carries little weight in the analysis of protein structures; therefore, we might be overly concerned about the potential issue. The relative insensitivity of the shape of the Ramachandran plots to the sequence identity cutoff indicates that using a higher cutoff value to include more structures may be more desirable in order to increase the number of relevant data, adding to the statistical significance of observations in sparse areas and the overall robustness of the plots. Nevertheless, we took a conservative approach and used only the plots from the fully filtered dataset for our later analysis.

To test the effect of crystallographic and geometric constraints, we also generated plots from the independent set without these constraints but including the sequence identity filter and compared these to the plot containing data that passed through the original three previously described filtering steps. Overlay of these plots shows that a notable portion of the unfiltered data contains conformations that lie more than 10° outside our lowest δ value contours [Fig. 4(c)]. Because of the nearly identical contours in the plots from the non-overlapping sets of high-quality data, the conformations in low-quality structures lying more than 10° outside of the −4 contours from high-quality structures might disappear if a higher-resolution structure were available.

Examination of the plots of individual residue types revealed some interesting features of their conformational tendencies. Generally, our individual Ramachandran plots (see supplemental materials) are similar to those from Chakrabarti and Pal's study.[5] Common to all residue types, the conformation at the pixel near (−110, 60) is less favorable (δ = −2.7 for the general case) than at surrounding pixels, which is due to the sum of the two staggered torsional angles: H-$C_\alpha$-C-O and $C_\beta$-$C_\alpha$-N-C. The conformations near ϕ = 0 basically do not exist, which is consistent with several previous studies (see, for example, refs. 10 and 13). In the other aspect, each Ramachandran plot is uniquely shaped, particularly notable in the positive ϕ area of each plot. The allowed status of ϕ–ψ angles near (70, −60) demonstrated in Lovell and coworkers' study[10] is validated by this work. Contributing to those ϕ–ψ angles mainly are β-branched residues (Ile, Val, and Thr), positively charged residues (Arg and Lys), and a few other polar residues (His, Tyr, and Ser). We also show the allowed status (although low preference) of ϕ–ψ angles around (75, 105). We observed the preference of Asp and Glu for the ϕ–ψ angles based on the selections described earlier. Further investigation of this region in the high-quality structures of the SCOP95 dataset revealed that this conformation is also preferred for Asn and Gln. Many Asn and Asp residues for these ϕ–ψ angles adopt a Type II′ β-turn-like structural motif. In the classical Type II′ β-turn, the backbone oxygen of residue $i - 1$ forms a hydrogen bond with the backbone nitrogen of residue $i + 2$. In this new motif, the δ-oxygen on the Asp or Asn side chain (residue $i$) replaces the backbone oxygen of residue $i - 1$, and forms the hydrogen bond to backbone nitrogen of residue $i + 2$. Examples of these residues are 1mda: D8:L, 1ia8: D148:A, 1ju3: D45:A, 1jz8: D164:A, 1k0m: D76:A, 3lkf: N91:A, 1eys: N194:M and 1qov: N195:M, where 1mda: D8:L means Asp in the 8$^{th}$ position of chain L in the PDB file 1mda. Our investigation also revealed that Glu and Gln residues adopting the ϕ–ψ angles are almost

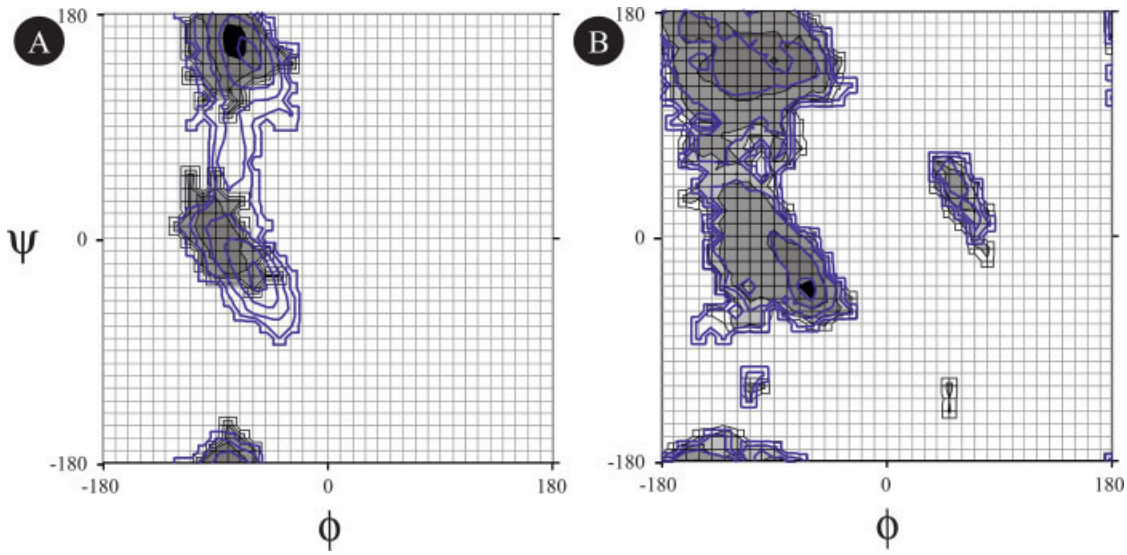


Fig. 5. Ramachandran plots of special amino acids. (**A**) depicts the distribution of *cis*-proline residues in grayscale with *trans*-proline contours superimposed in blue. (**B**) depicts the distribution of both types of cysteine residues. The disulfide is shown in grayscale, and the thionyl contours are superimposed in blue.

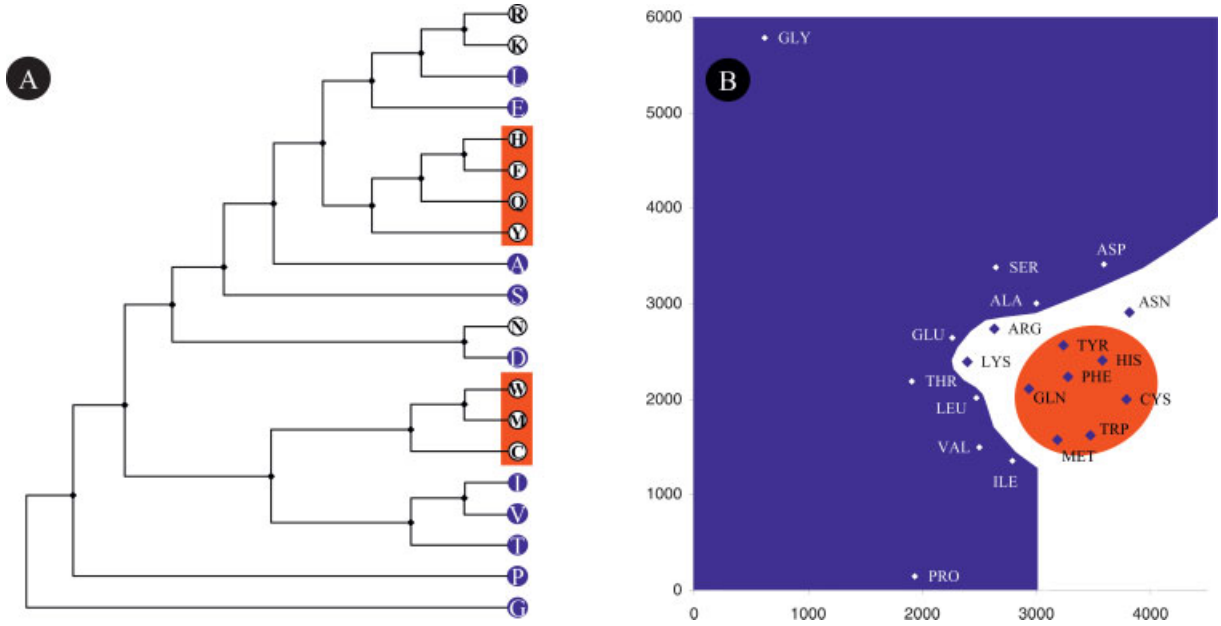


Fig. 6. Graphical representations of the distance matrix. The amino acids in blue are produced in Miller's experiment, and those in red boxes are Trifonov's late amino acids. (**A**) A clustering tree of amino acids based on complete-linkage cluster analysis. (**B**) Map of main-chain conformational tendencies of amino acids. The unit of each axis is the δ-value.

exclusively found in the active site of Glutathione S-transferase or related enzymes. Examples are 1a0f: E65:A, 1gnw: E66:A, 1axd: E66:D, 1f3a: Q66:A, 1glq: Q64:A, 1k3y: Q67:A, 1dug: Q66:A, and 2gst: Q71:A. Finally, the conformations near $\phi$–$\psi$ angles of $(50, -120)$ were singled out by Jones et al.[16] to illustrate the point that earlier classified "outliers" in a Ramachandran plot as not necessarily errors, because the active-site nucleophile of the α/β hydrolases always has this conformation. This "disallowed" region was studied previously, and it was concluded that it contains genuine conformations.[10,18,19] Our study confirms that these conformations are not only genuine ones, but also common to all residues except Pro, Met, Trp, Ile, and Val. These residues have bulky side chains and a lack of polar side chains, or have a restricted rotable bond.

We also studied the Ramachandran plots of special residues, proline and cysteine. While there are insufficient data to draw conclusions for other residues with a peptide bond in the *cis*-configuration, comparison of the proline Ramachandran plots in both configurations showed that they do noticeably differ from each other. In addition to having a smaller accessible area as compared to *trans*-Pro,

**TABLE I. Distance Matrix of Main-Chain Conformational Tendencies of L-Amino Acids in *Trans*-Configuration**

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 0 | 1038 | 1247 | 1276 | 1217 | 1084 | 1035 | 3524 | 1112 | 1364 | 1118 | 1127 | 1237 | 1121 | 2367 | 1135 | 1248 | 1282 | 1135 | 1390 |
| Arg | | 0 | 1195 | 1258 | 1062 | 913 | 937 | 3860 | 910 | 1195 | 940 | 896 | 1088 | 928 | 2304 | 1166 | 1157 | 1028 | 999 | 1172 |
| Asn | | | 0 | 1025 | 1233 | 1179 | 1262 | 3776 | 1015 | 1534 | 1263 | 1272 | 1304 | 1090 | 2634 | 1157 | 1352 | 1327 | 1111 | 1548 |
| Asp | | | | 0 | 1437 | 1350 | 1334 | 3628 | 1244 | 1773 | 1411 | 1359 | 1537 | 1292 | 2760 | 1312 | 1547 | 1571 | 1380 | 1717 |
| Cys | | | | | 0 | 955 | 1151 | 4092 | 988 | 1211 | 1149 | 1189 | 962 | 1041 | 2300 | 1387 | 1262 | 937 | 1012 | 1154 |
| Gln | | | | | | 0 | 920 | 3946 | 906 | 1074 | 961 | 950 | 962 | 835 | 2215 | 1239 | 1173 | 995 | 967 | 1065 |
| Glu | | | | | | | 0 | 3768 | 1076 | 1167 | 989 | 951 | 1158 | 1034 | 2348 | 1205 | 1139 | 1090 | 1047 | 1150 |
| Gly | | | | | | | | 0 | 3912 | 4359 | 3898 | 3791 | 4148 | 3861 | 5067 | 3561 | 3824 | 4163 | 3937 | 4215 |
| His | | | | | | | | | 0 | 1194 | 1011 | 991 | 1034 | 792 | 2369 | 1201 | 1181 | 1092 | 937 | 1201 |
| Ile | | | | | | | | | | 0 | 1018 | 1180 | 995 | 1117 | 1898 | 1651 | 1180 | 1002 | 1171 | 701 |
| Leu | | | | | | | | | | | 0 | 919 | 1026 | 948 | 2101 | 1287 | 1136 | 1087 | 1023 | 1109 |
| Lys | | | | | | | | | | | | 0 | 1136 | 974 | 2357 | 1289 | 1165 | 1182 | 976 | 1198 |
| Met | | | | | | | | | | | | | 0 | 1022 | 2020 | 1459 | 1235 | 883 | 1113 | 924 |
| Phe | | | | | | | | | | | | | | 0 | 2376 | 1219 | 1189 | 1007 | 904 | 1078 |
| Pro | | | | | | | | | | | | | | | 0 | 2738 | 2320 | 2075 | 2468 | 1984 |
| Ser | | | | | | | | | | | | | | | | 0 | 1266 | 1484 | 1212 | 1544 |
| Thr | | | | | | | | | | | | | | | | | 0 | 1245 | 1136 | 1125 |
| Trp | | | | | | | | | | | | | | | | | | 0 | 1067 | 1043 |
| Tyr | | | | | | | | | | | | | | | | | | | 0 | 1180 |
| Val | | | | | | | | | | | | | | | | | | | | 0 |

*cis*-Pro conformations shift slightly left and up, and observations are absent around $\psi$ +70. Furthermore, *cis*-Pro conformations prefer the region around ($\phi$, $\psi$) values of (−70, 150) to that of (−70, −20), while the preferences of *trans*-Pro are relatively equal in these two regions [Fig. 5(a)]. The Ramachandran plots of the two forms of cysteine, thionyl and disulfide, are quite similar to each other [Fig. 5(b)].

### Quantitative Relationship of Folding Properties of Amino Acids

Since the distribution of main-chain dihedral angles describes the folding tendency of amino acids, we attempted to quantify the differences among them using the information contained in our Ramachandran plots. We employed a city-block distance method to calculate these pairwise distances (Table I). Because it is the logarithm of the odds of observing an amino acid in a particular conformation, δ-value is related to the free energy of the conformation. Therefore, the difference between the δ-values of two residues at pixel $i$ measures the relative tendency of one residue over the other to adopt this conformation. For example, the tendency of Asn to adopt the conformation at (60, 30) is higher than Val because Asn has a δ-value of +2 and Val has one of −4, a difference of six units. When we considered all of the pixels, we concluded that the absolute difference of δ-values between two residues measures their overall divergence of folding tendencies. We defined the $D$-function as the summation of the absolute difference of δ-values as the distance of folding tendencies between two residues. Table I shows the distance matrix of all 20 L-amino acid types in the *trans*-configuration. Although the values in Table I might change depending on the assignment of the δ-values to zero-data pixels, tests of different assignments from −6 to −10 showed that they had little impact on the relative distances in the table. To test whether the dataset size of each residue is the major determinant of the distance value in the table, we also re-calculated the distance matrix based on the Ramachandran plots of residues with similar data points (≈5K). The linear correlation coefficients of the new distance matrices and that in Table I is 0.98 (0.96 if Pro and Gly are excluded). The high correlation demonstrates that the distance matrix shown in Table I represents the underlying relationship of folding tendencies of amino acids.

Table I, however, lacks the visual power to summarize the quantitative relationship of amino acids in terms of their main-chain conformations. To combine them into groups, we performed a cluster analysis on the distance matrix using a complete-linkage method,[14] resulting in Figure 6(a). While the groups clustered in Figure 6(a) roughly correspond to expectations of similar side chains (large residues such as His, Gln, Tyr, and Phe are in the same group, as are the β-branched amino acids Ile, Thr, and Val), we noticed some residues (such as Cys and His) are close in distance in Table I but appear far apart in Figure 6(a). We concluded that this dendrogram does not present a full picture of the distance relationship. To overcome the intrinsic loss of information inherent in the clustering process, we attempted to reduce the high-dimensional matrix to a two-dimensional distance map. Using a multidimensional simplex minimization method coupled with simulated annealing,[15] we reduced the distance matrix to a two-dimensional map [Fig. 6(b)]. The final error between the two-dimensional map and the original matrix is 17.6% (or 13.2% when Gly and Pro are excluded). Compared to Figure 6(a), Figure 6(b) contains distance information of all amino acids in Table I, indicating that a reduced two-dimensional map, if constructed with small errors, is a better portrait of a distance matrix than a cluster dendrogram. The fact

**TABLE II. Substitutional Priority of Amino Acids**

| Original | High similarity | | | | | | | | Medium similarity | | | | | | Low similarity | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | Glu | Arg | Gln | His | Leu | Phe | Lys | Tyr | Ser | Cys | Met | Asn | Thr | Asp | Trp | Ile | Val | Pro | Gly |
| Arg | Lys | His | Gln | Phe | Glu | Leu | Tyr | Trp | Ala | Cys | Met | Thr | Ser | Val | Ile | Asn | Asp | Pro | Gly |
| Asn | His | Asp | Phe | Tyr | Ser | Gln | Arg | Cys | Ala | Glu | Leu | Lys | Met | Trp | Thr | Ile | Val | Pro | Gly |
| Asp | Asn | His | Arg | Ala | Phe | Ser | Glu | Gln | Lys | Tyr | Leu | Cys | Met | Thr | Trp | Val | Ile | Pro | Gly |
| Cys | Trp | Gln | Met | His | Tyr | Phe | Arg | Leu | Glu | Val | Lys | Ile | Ala | Asn | Thr | Ser | Asp | Pro | Gly |
| Gln | Phe | His | Arg | Glu | Lys | Cys | Leu | Met | Tyr | Trp | Val | Ile | Ala | Thr | Asn | Ser | Asp | Pro | Gly |
| Glu | Gln | Arg | Lys | Leu | Phe | Ala | Tyr | His | Trp | Thr | Val | Cys | Met | Ile | Ser | Asn | Asp | Pro | Gly |
| Gly | Ala | Ser | Asp | Glu | Asn | Lys | Thr | Arg | Phe | Leu | His | Tyr | Gln | Cys | Met | Trp | Val | Ile | Pro |
| His | Phe | Gln | Arg | Tyr | Cys | Lys | Leu | Asn | Met | Glu | Trp | Ala | Thr | Ile | Val | Ser | Asp | Pro | Gly |
| Ile | Val | Met | Trp | Leu | Gln | Phe | Glu | Tyr | Lys | Thr | His | Arg | Cys | Ala | Asn | Ser | Asp | Pro | Gly |
| Leu | Lys | Arg | Phe | Gln | Glu | His | Ile | Tyr | Met | Trp | Val | Ala | Thr | Cys | Asn | Ser | Asp | Pro | Gly |
| Lys | Arg | Leu | Gln | Glu | Phe | Tyr | His | Ala | Met | Thr | Ile | Trp | Cys | Val | Asn | Ser | Asp | Pro | Gly |
| Met | Trp | Val | Gln | Cys | Ile | Phe | Leu | His | Arg | Tyr | Lys | Glu | Thr | Ala | Asn | Ser | Asp | Pro | Gly |
| Phe | His | Gln | Tyr | Arg | Leu | Lys | Trp | Met | Glu | Cys | Val | Asn | Ile | Ala | Thr | Ser | Asp | Pro | Gly |
| Pro | Ile | Val | Met | Trp | Leu | Gln | Cys | Arg | Thr | Glu | Lys | Ala | His | Phe | Tyr | Asn | Ser | Asp | Gly |
| Ser | Ala | Asn | Arg | His | Glu | Tyr | Phe | Gln | Thr | Leu | Lys | Asp | Cys | Met | Trp | Val | Ile | Pro | Gly |
| Thr | Val | Leu | Tyr | Glu | Arg | Lys | Gln | Ile | His | Phe | Met | Trp | Ala | Cys | Ser | Asn | Asp | Pro | Gly |
| Trp | Met | Cys | Gln | Ile | Phe | Arg | Val | Tyr | Leu | Glu | His | Lys | Thr | Ala | Asn | Ser | Asp | Pro | Gly |
| Tyr | Phe | His | Gln | Lys | Arg | Cys | Leu | Glu | Trp | Asn | Met | Ala | Thr | Ile | Val | Ser | Asp | Pro | Gly |
| Val | Ile | Met | Trp | Gln | Phe | Leu | Thr | Glu | Cys | Arg | Tyr | Lys | His | Ala | Ser | Asn | Asp | Pro | Gly |

that a two-dimensional map can quite faithfully reproduce a 20 × 20 matrix also implies that the diversity space of main-chain conformational tendencies is primarily two-dimensional in δ units, which are related to main-chain conformational energies.

We further noticed that residues that appeared at different evolutionary stages are located in different zones in Figure 6(b). The blue zone contains the early residues found in the Miller experiment,[20] and the red zone contains later residues based on analyses by Trifonov.[21,22] While the conformational space of earlier residues represents the existing conformations in the primordial soup of the early earth, nature's incorporation of later amino acids into peptide chains reflects the restrictions of structural compatibility, conformational diversity, and availability of substances during evolution. However, it remains to be seen whether the clear separation of early and later amino acids in the diversity space implies a deep role of folding tendencies in amino acid evolution, or whether it is just a coincidence of contributions from three factors: availability, structural compatibility, and conformational diversity. To test whether the clear separation is an artifact generated by our imperfect minimization process, we tried twice to generate two-dimensional maps from our random sets containing approximately the same number of data for each residue type. These experiments confirmed our earlier results. Nevertheless, we cannot firmly rule out the possibility of artifacts due to the inherent fallibility of stochastic minimization algorithms, which led us to a slightly different two-dimensional map each time. This doubt can be addressed more completely with future higher-quality Ramachandran plots when significantly more structures become available.

Finally, we attempted to generate an amino-acid substitution table for application to protein engineering projects. Substitution tables have been created to guide molecular biologists who attempt to mutate amino acid residues and maintain the main-chain conformation and folding pathway of a protein.[23,24] One approach is to mutate a residue into another with a smaller side-chain, as employed in alanine scanning technology.[25] The problem with alanine scanning, specifically, is that alanine has a strong tendency to form α-helices,[23] and therefore may change the folding pathway of the variant. Another approach is to mutate a residue to an "isosteric" one,[23,24] as has been implemented in several patent applications.[23,26] To our knowledge, none of these tables are the result of direct measurements of the similarity of amino acids in terms of conformational or folding tendencies. Since distance is inversely proportional to similarity, we can easily transform the quantitative information of folding tendencies in Table I to this type of substitution table (Table II). Upon examination of this table, we find that most of the expected close relationships are retained, such as those of Asp to Asn, Arg to Lys, Ile to Val, and Tyr to Phe.

Our table also details some unexpected features. First, the similarity between members is not necessarily commutable, i.e. the most similar residue to Tyr is Phe, but the most similar residue to Phe is His. Although this asymmetry may seem counter-intuitive, it is the nature of distance that an object's closest neighbor may have a third object even closer. Second, some residues, shown to have different preferences in previous studies, are closely related in our map. Leu and Ile, for example, have been shown to have different preferences for certain conformations[27] but are rather inter-substitutable. In fact, Ile has often appeared in Leu positions in leucine-rich-repeat proteins.[28] Third, there are some unexpected relationships between residues; for example, Met and Trp are more similar than Trp and Phe, and Glu and Asp are less similar than Arg and Lys. Although Asp and Glu are both acidic residues and are often seen to be interchangeable, their sharp difference in conformational preferences was noticed previously, whereas the difference between Arg and Lys was not

as noticeable.[29,30] In pharmaceutical industries, it is often desirable to choose another residue with very different types of side-chains to modify the physiochemical properties of a protein, such as solubility, stability, and half-life. The merit of Table II is that it suggests unobvious substitutions. However, the usefulness of Table II alone is limited for its application in protein engineering. When the conformation of a residue is known, one should consult the specific structural environment to ensure the spatial and electrostatic complementarities in addition to checking the individual plots in the supplementary materials for favorable propensity of the new residue at the specifically relevant conformation.

## ACKNOWLEDGMENTS

We thank Dr. Zhaowen Luo for technical assistance and Drs. Deanne Taylor and Steve Arkinstall for helpful discussions.

## REFERENCES

1. Ramachandran GN, Ramakrishnan C. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963;7:95–99.
2. Hu H, Elstner M, Hermans J. Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. Proteins 2003;50:451–463.
3. Ho BK, Thomas A, Brasseur R. Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. Protein Sci 2003;12:2508–2522.
4. Walther D, Cohen FE. Conformational attractors on the Ramachandran map. Acta Crystallogr D Biol Crystallogr 1999;55(2):506–517.
5. Chakrabarti P, Pal D. The interrelationships of side-chain and main-chain conformations in proteins. Prog Biophys Mol Biol 2001;76:1–102.
6. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 1993;26:283–291.
7. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. *Proteins* 1992;12:345–364.
8. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. *Nature* 1996;381:272.
9. Vriend G. WHAT IF: a molecular modeling and drug design program. *J Mol Graph* 1990;8:29,52–56.
10. Lovell SC, Davis IW, Arendall III WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson, DC. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 2003;50:437–450.
11. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
12. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000;28:254–256.
13. Kleywegt GJ, Jones TA. Phi/psi-chology: Ramachandran revisited. Structure 1996;4:1395–1400.
14. Sharma S. Applied multivariate techniques. New York: Wiley; 1996. p 185–233.
15. Press WH, Flannery BP, Teukolsky SA, Vetterling WT. Numerical recipes in C, 2nd ed. Cambridge: Cambridge University Press; 1992. p 408–455.
16. Jones TA, Kleywegt GJ, Brunger AT. Storing diffraction data. Nature 1996;383:18-19.
17. Pal D, Chakrabarti P. Cis peptide bonds in proteins: residues involved, their conformations, interactions and locations. J Mol Biol 1999;294:271–288.
18. Gunasekaran K, Ramakrishnan C, Balaram P. Disallowed Ramachandran conformations of amino acid residues in protein structures. J Mol Biol 1996;264:191–198.
19. Pal D, Chakrabarti P. On residues in the disallowed region of the Ramachandran map. Biopolymers 2002;63:195–206.
20. Miller SL. Which organic compounds could have occurred on the prebiotic earth? Cold Spring Harb Symp Quant Biol 1987;52:17–27.
21. Trifonov EN. Glycine clock: eubacteria first, archaea next, protoctista, fungi, planta and animalia at last. Gene Ther Mol Biol 1999;4:313–323.
22. Trifonov EN. Consensus temporal order of amino acids and evolution of the triplet code. Gene 2000;261:139–151.
23. Cunningham BC, Lowman HB, Wells JA et al, inventors. Genetech I, assignee. Human Growth Hormone Variants. USA patent WO 97/11178. Mar 1997.
24. Lehninger AL. Biochemistry, 2nd ed. New York: Worth; 1975.
25. Cunningham BC, Wells JA. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science 1989;244:1081–1085.
26. Filikov A, inventor. Xencor I, assignee. Novel nucleic acids and proteins with growth hormone activity. USA patent WO 00/68385. Nov 2000.
27. Swindells MB, MacArthur MW, Thornton JM. Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. Nat Struct Biol 1995;2:596–603.
28. Kobe B, Kajava AV. The leucine-rich repeat as a protein recognition motif. Curr Opin Struct Biol 2001;11:725–732.
29. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 1974;13:211–222.
30. George RA, Heringa J. An analysis of protein domain linkers: their classification and role in protein folding. Protein Eng 2002;15:871–879.