II. STRUCTURE

1. Conformational Properties of Amino Acids. Implications for Protein Structures

a. Proteins are heteropolymers made of amino acids

The building blocks of proteins are *amino acids.* An amino acid is so-named because it contains an *amino* group ($-NH_2$) at one end, and a carboxylic *acid* (–COOH) at the other end. At physiological pH both the amino and carboxylic groups are completely ionized. The amino acid can thus act as either an acid or a base. The amino and carboxylic groups are joined by a tetrafunctional carbon atom, called the C^{α}-atom, or α-carbon, hence the name α-amino acid. Combined, this unit is called the *backbone* of the amino acid. Attached to the C^{α}-atom are the amino acid *sidechain,* often denoted by the generic symbol "R", and a hydrogen atom.

Proteins are *linear* polymer molecules – chains of monomer units covalently strung together like beads on a necklace. The monomer in every protein – each bead, in the necklace analogy – is an amino acid; its generic formula is –[NH (C^{α}HR) (CO)]–. Each amino acid thus contributes two *polar* groups, -N-H and –C=O, to the protein backbone, after polymerizing into a linear chain. A polymer is called a *homopolymer* if all of its monomers are identical. In proteins, each monomer can be one of twenty different types of *natural* amino acids, depending on the identity of its side group R. See *Figure II.1.1* and *Table II.1.1*. It is also possible to have other amino acids – compounds with the same backbone, but with different sidechains. Such *non-natural* amino acids are occasionally found in biological proteins. Since they are made of more than one monomer type, proteins are *heteropolymers*. Typically, a protein may have 30 to 1,000 monomers.(*)

Although all proteins are synthesized in cells as linear polymer chains, the covalent connectivities of some proteins can be more complex because chemical changes sometimes take place after a protein's initial synthesis. These are called *posttranslational modifications*. Some proteins can also develop intrachain disulfide and other cross-links. These cross-links provide extra stability to native structures.

(*) Note that the term *monomer* is also used in structural biology for designating substructures formed by a single chain molecule, i.e., subunits, when the structure comprises more than one molecule. *Dimer*, *trimer* or *multimer* then refer to structures composed of two, three or multiple macromolecular chains. Likewise, *homodimer* and *heterodimer* describe proteins composed of two identical or two different chains.

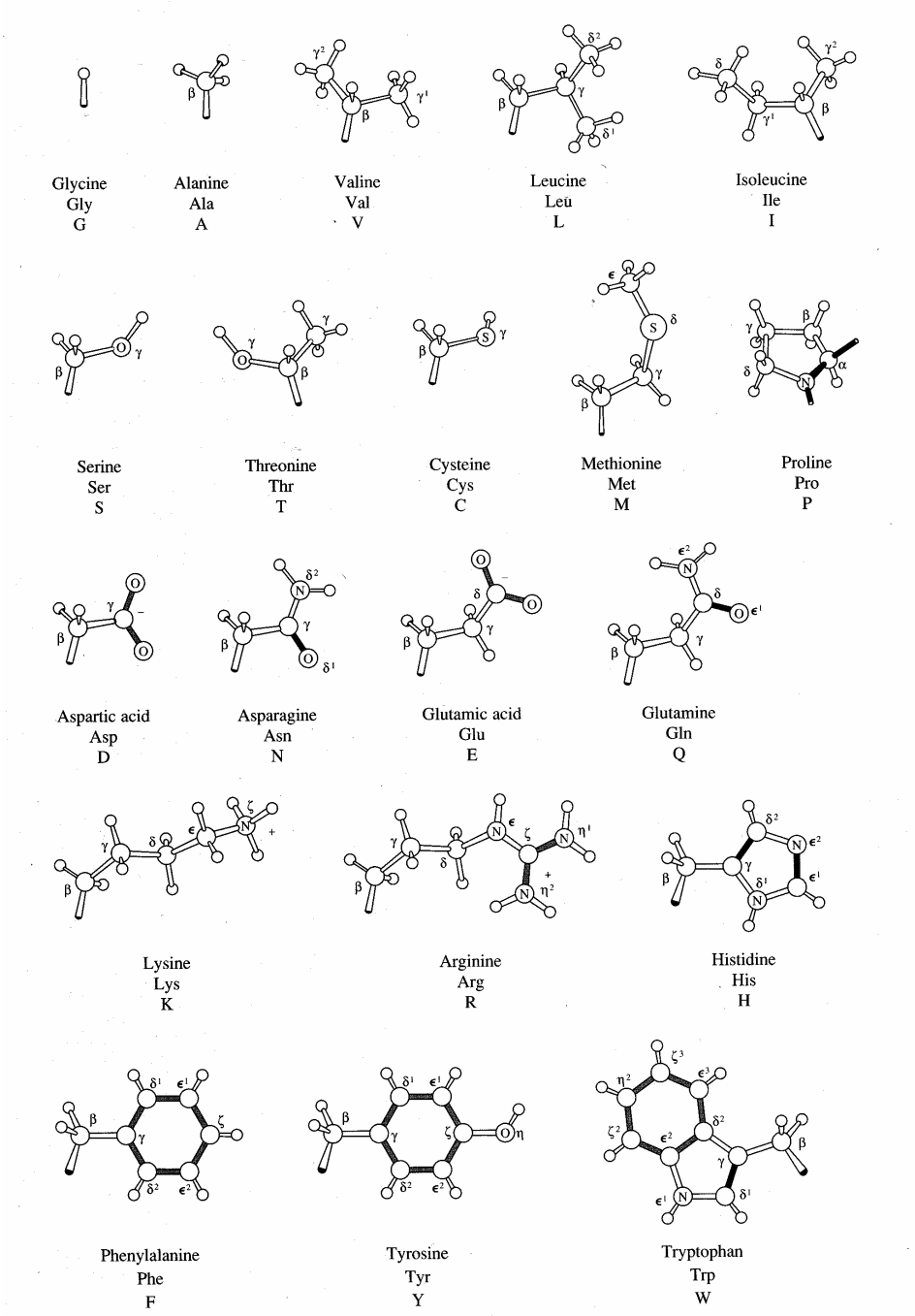


Figure II.1.1. *Side groups of the twenty different types of amino acids naturally occurring in proteins. Unlabelled spheres are hydrogen (small) and carbon atoms (large). Other atoms (O, N and S) are labelled. Double bonds and partial double bonds are shown in black. Backbone peptide bonds are not shown, except for those in Pro, which are indicated in black. The superscripts β, γ, etc. distinguish the atoms of a given type along the side chain; C^{β} is, for example, the carbon atom bonded to the backbone C^{α} atom. (Figure 1.1 of {Creighton 1993 ID: 280}).*

Table II.1.1 Properties of Amino Acids [a]

Residue Type	Mass[b] (daltons)	Volume[c] (Å^3)	Accessible surface area[d] (Å^3)
Ala	71.1	91.5	113
Arg	156.2	202.1	241
Asn	114.1	135.2	158
Asp	115.1	124.5	151
Cys	103.2	111.7	140
Gln	128.1	161.1	189
Glu	129.1	155.1	183
Gly	57.1	66.4	85
His	137.1	167.3	194
Ile	113.2	168.8	182
Leu	113.2	167.9	180
Lys	128.2	171.3	211
Met	131.2	170.8	204
Phe	147.2	203.4	218
Pro	97.1	129.3	143
Ser	87.1	99.1	122
Thr	101.1	122.1	146
Trp	186.2	237.6	259
Tyr	163.2	203.6	229
Val	99.1	141.7	160

[a] Adapted from Table 1.1 of Ref. {Creighton 1993 ID: 280} and Table 1.1 of Ref. {Darby & Creighton 1993 ID: 359}.

[b] Molecular weight of unionized amino acid minus that of water.

[C] The volumes V, except for Arg, have been taken from Table 2 of{Chothia 1975 ID: 48}, and V for Arg from Table 4 of {Chothia 1975 ID: 48}

[d] Measured in a Gly-X-Gly tripeptide in an extended conformation, where X is the given amino acid residue {Miller, Janin, et al. 1987 ID: 85}.

Amino acids have *chirality*, or *handedness*: they are not superimposable on their mirror image, in the same way as a left hand is not superimposable on a right hand. This is a typical property of molecules containing a *chiral* carbon atom, – a tetrahedral carbon having four different substituents, such as the C^{α}-atoms of amino acids (except glycine). *Figure II.1.2* shows the two possible *optical isomers* called the *L-isomer* – also known as the left-handed (*levo*) form, or the *D-isomer* – the right-handed (*dextro*) form of amino acids. These rotate the plane of plane-polarized light in opposite directions. For some reason, biology evolved using almost exclusively L-amino acids. Note that these two isomers are fundamentally of different character from the rotational isomers introduced in the Appendix § II.A1, in that there is no possibility of interconversion between the D- and L- isomers, except by breaking chemical bonds.

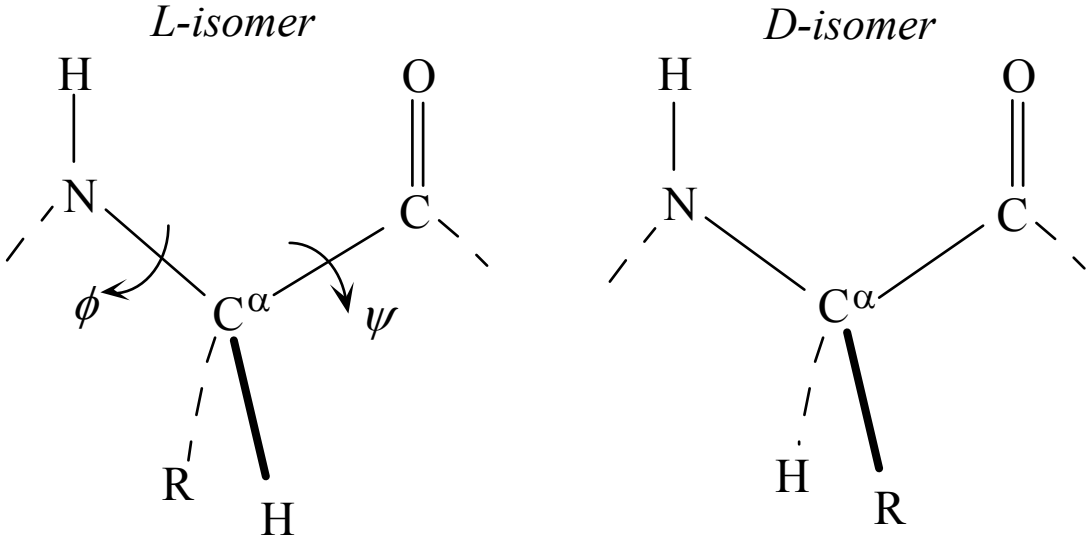


Figure II.1.2. *Two optical isomers, L- and D-, for α-carbons of amino acids. Note the difference in the relative position of the sidechains R. Dashed and boldface lines designates the bonds pointing behind and in front of the plane spanned by the backbone bonds N-C^{α} and C^{α}-C. The side groups N-H and C=O lie in the plane of the backbone bonds. The L-isomer is the form naturally occurring in proteins. The two backbone bonds flanking the α-carbon are rotatable. Their rotational angles φ and ψ are indicated (on the L-isomer).*

b. The different structures of amino acids lead to different properties

Table I.1.1 shows more than a 3-fold range in the volumes of amino acids, and nearly a 7-fold range in their frequencies of occurrence in proteins. Depending on the chemical properties of their sidechains, amino acids can be grouped into classes, such as *charged*,

polar (P*)*, and *nonpolar*. Lysine and arginine are positively charged; aspartic acid and glutamic acid are negatively charged. Alanine, valine, leucine, isoleucine, and phenylalanine sidechains are *nonpolar,* since they are comprised of hydrocarbons only, $–CH_2–$ (methylene) and $–CH_3$ (methyl) groups, which are not readily polarized by electrical fields. Glycine may also be classified in this group, having only a hydrogen atom, the smallest possible 'sidechain', affixed to its C^{α}-atom. Nonpolar groups have an aversion to water. They tend to cluster together and form contacts among themselves to avoid water. Amino acids having nonpolar sidechains are therefore *hydrophobic* (H)*,* except for glycine and alanine, the two smallest amino acids in which the backbone polar groups dominate the behavior. On the other hand, cysteine and methionine, despite their polar sulfur atoms, tend to be buried in the interior of proteins, similarly to hydrophobic residues. Serine and threonine are *polar* because of their –OH (hydroxyl) terminal groups; and asparagine and glutamine are polar because of their $(C=O)-NH_2$ groups at the sidechain termini. These have an affinity for water.

These groupings are neither precisely defined nor fully accurate. Several amino acids have multiple personalities, having both nonpolar and polar or ionic character. Most of these groups have ionization states determined by their intrinsic pKa values together with the local pH. For example, tyrosine has a bulky hydrophobic part, its phenyl moiety, and a polar group, –OH at the sidechain terminus. Likewise, tryptophan, with its indole group containing two rings, is closer in character to the group of hydrophobic amino acids than the polar ones. The histidine sidechain, a heterocyclic imidazole, can ionize at physiological pH, and thus could be assigned either to the group of charged amino acids or to the group of uncharged ones, depending on the local pH. Lysine is classified as a charged residue because its terminal amino group is ionized under most physiological conditions, but its sidechain also contains a hydrophobic segment of four methylene groups. Likewise, the arginine sidechain contains three methylene groups.

Size, shape and flexibility are important properties, almost as important as hydrophobicity and polarity, in determining the behavior of amino acids. The charged sidechains Lys and Arg having bulky and highly flexible sidechains, for example, are fundamentally different from the charged residue Asp. Lys and Arg are almost exclusively solvent-exposed, wobbling in the surrounding solution thus increasing the solubility of the protein and usually being targets for enzyme/DNA recognition; whereas Asp can accommodate inner positions, the entropy cost of restricting its conformational freedom being much lower than that of Lys or Arg. Phe, a hydrophobic sidechain, has much in common with the other aromatic sidechains Tyr and Trp. Even the residues having aliphatic sidechains, differ in their flexibility, the position of the branching of the sidechain being an important determinant of flexibility. Val and Ile are branched at their β-carbon, which confers some stiffening in the main chain; whereas no main chain hindrance occur with Leu because of its branching at the more distant γ-carbon. Gly is unique. Its small size aids in accommodating tightly packed regions, while maintaining an intrinsic flexibility due to the rotational freedom of the corresponding (ϕ,ψ) angles. Not surprisingly, Gly is comparatively well conserved during evolution.

The Venn diagram in *Figure II.1.3* summarizes some of these chemical and physical properties. It is clear that amino acid substitutions depend closely on the extent of the physical and chemical properties that they share, and this Venn diagram can thus be correlated with amino acid conservation/mutation expectancies.

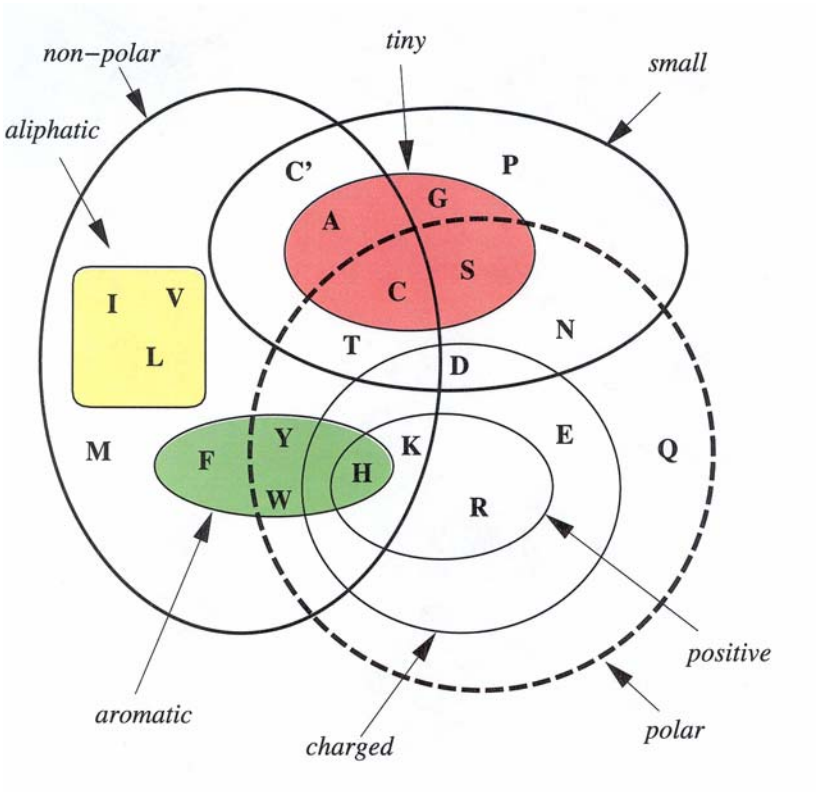


Figure II.1.3. *Venn diagram illustrating the physical and chemical similarities and differences of amino acids A similar diagram was proposed by Taylor {Taylor 1986 ID: 516}.*

c. In proteins, amino acids are connected by peptide bonds

In proteins, an amino acid is covalently connected to its neighbor by a *peptide bond,* hence proteins are sometimes called *polypeptides*. The peptide bond is a covalent link from the carbonyl end of one repeat unit –[NH –C$^{\alpha}$HR–CO]–, to the amino at the beginning of the next. The chemical process of forming a peptide bond is a *condensation* reaction, because it involves the release of one water molecule. The formation of a polypeptide of n repeat units (or *residues*), on the other hand, releases n-1 water molecules. The corresponding reaction is

n(*H*NH–C$^{\alpha}$HR–CO*OH)* —> *H*–[NH –C$^{\alpha}$HR–CO]$_n$ *OH* + (n-1)H_2O

In conformity with the above head-to-tail connection of repeat units, the protein ends are referred to as the N- and C-terminal ends, and amino acids along the chain are assigned serial indices 1, 2, …., n starting from the N-terminus.

Figure II.1.4 displays a segment of two amino acids i and i+1 in the *trans* conformation of the backbone. For simplicity, only the C^{β} atoms are displayed. Three types of backbone bonds are distinguished, contributed by each residue: N – C^{α}, C^{α} – C and the peptide bond C – N. The peptide bond is quite rigid, with its directly attached carbonyl oxygen, amide hydrogen and two successive C^{α} atoms essentially on the same plane (*Figure II.1.5*). Its rigidity originates in its having partial double bond nature. The two backbone bonds flanking the C^{α} carbons, on the other hand, are free to rotate about their own axis. The backbone may thus be viewed as a succession of planar groups (peptide units) that are relatively free to pivot around the C^{α} carbons. This distinctive structural feature permits us to express the backbone configuration in terms of *virtual bonds* that join successive α-carbons. See *Figure II.1.6.*

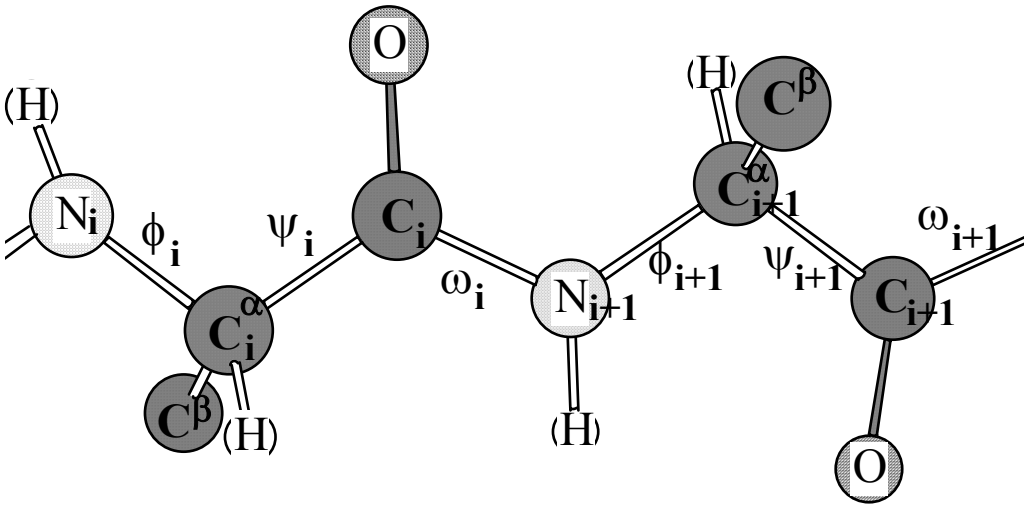


Figure II.1.4. *Segment of a polypeptide chain. Residues i and i+1 are displayed. Backbone atoms are indexed by the residue they belong to. For clarity only the C^{β} atoms of the sidechains are shown. Backbone torsions associated with the bonds $N_i - C_i^{\alpha}$, $C_i^{\alpha} - C_i$ and $C_i - N_{i+1}$ are denoted as ϕ_i, ψ_i and ω_i, respectively. The ith peptide bond connects the atoms C_i and N_{i+1}. Usually ω is trans ($\omega = 180^0$) but occasionally it can be in the cis form ($\omega = 0$).*

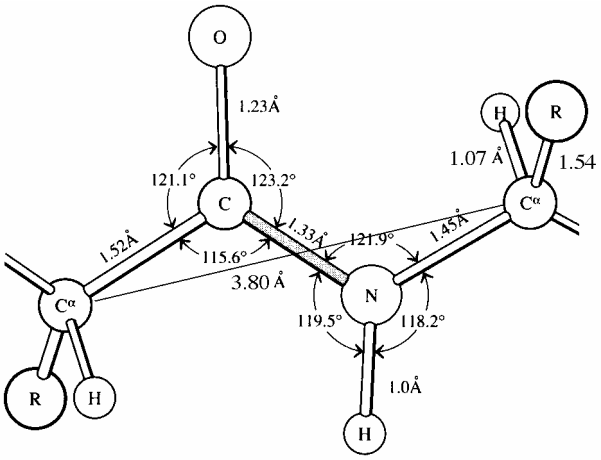


Figure II.1.5. *Geometry of the protein backbone with a peptide bond in the trans conformation, showing two adjacent C^{α} atoms (adapted from Figure 1.2 in {Creighton 1993 ID: 495}. Original ref: G.N.Ramachandran et al. Biochim. Biophys. Acta 359, 298-302, 1974)).*

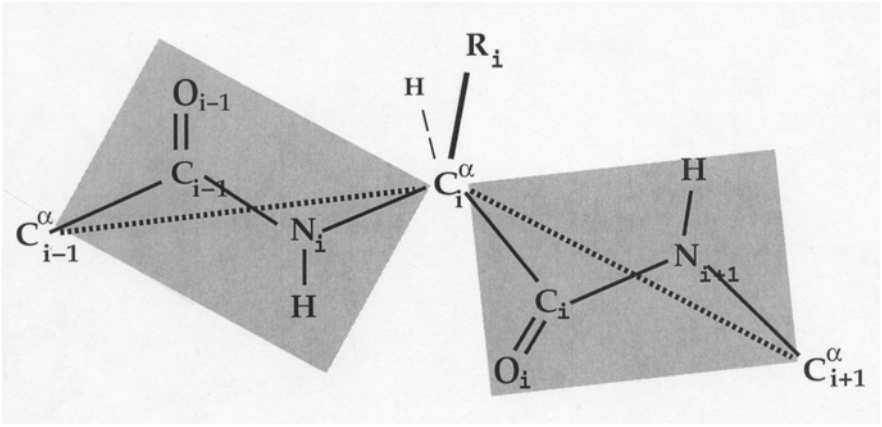


Figure II.1.6. *Virtual bond representation of the protein backbone. Dotted lines are the virtual bonds connecting successive α-carbons. This representation takes advantage of the planarity of the three successive backbone bonds, C^{α}_{i-1}-C_i, C_i-N_i and N_i-C^{α}_i corresponding to each amino acid. The lengths of the virtual bonds are fixed to the extent that the bond lengths and bond angles are fixed and the amide bond adheres to the trans conformation.*

The rotational freedom around the backbone is described by two dihedral angles, ϕ and ψ. The rotation around the N–C^{α} bond defines the torsional angle ϕ, and that around C^{α}– C

defines ψ. The peptide bond rotational angle is designated as ω. It usually assumes the *trans* planar form of the peptide unit, - except for the peptide bond preceding a proline which is more likely to adopt the *cis* state because of the constraints imposed by its neighboring proline sidechain, a five-membered pyrrolidine ring.

d. Ramachandran maps indicate the accessible ranges of ϕ and ψ angles

Some values of ϕ and ψ are more favorable than others. Also, different amino acids have different ϕ ψ angle preferences. Ramachandran and coworkers were the first to show that the main determinant of the different preferences of different amino acids is the degree to which the particular amino acid's sidechain interferes with the backbone by steric collision {Saisekharan 1962 ID: 504}{Ramachandran, Ramakrishnan, et al. 1963 ID: 501}{Ramakrishnan & Ramachandran 1965 ID: 503}{Ramachandran 1966 ID: 500}{Ramachandran & .& Saisekharan 1968 ID: 502}. In their pioneering work, the permissible values of ϕ and ψ were determined for Ala and Gly, using fixed bond lengths and bond angles (*Figure II.1.5*), and atoms interacting via *hard-sphere potentials* (*Table II.1.2*). The resulting so-called *Ramachandran maps* are shown *Figure II.1.7*. These are two-dimensional plots of accessible pairs of ϕ and ψ angles for the ith residue, considering atoms from C^{α}_{i-1} to C^{α}_{i+1}, inclusive, the peptide bonds being constrained to their planar *trans* conformations.

TABLE II.1.2. Interatomic Distances for Hard-sphere Potentials[a]

Atom Pair	Minimum Allowable Separation(Å)	Outer Limit(Å)
C·····C	3.2	3.0
C·····O	2.8	2.7
C·····N	2.9	2.8
C·····H	2.4	2.2
O·····O	2.8	2.7
O·····N	2.7	2.6
O·····H	2.4	2.2
N·····N	2.7	2.6
N·····H	2.4	2.2
H·····H	2.0	1.9

[a]used in Ramachandran plots {Ramachandran, Ramakrishnan, et al. 1963 ID: 501}. See *Figure II.1.7*.

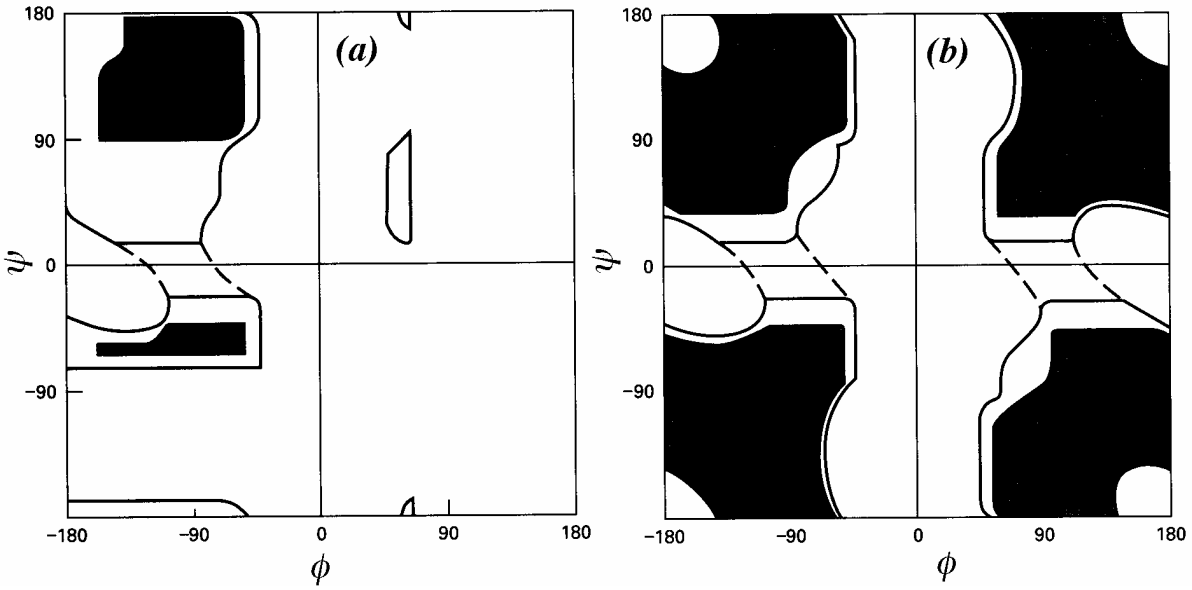


Figure II.1.7. *Ramachandran plots of the adjacent dihedral angles ϕ and ψ in (a) alanine, and (b) glycine. Black regions represent the allowed regions on the basis of the fixed bond lengths and bond angles, and hard-sphere potentials with distance parameters listed in the second column of Table II.1.1. Regions enclosed by the solid lines comply with the outer limiting separations listed in the third column of Table II.1.1. The dashed curves refer to boundaries that may be reached by slight distortions in bond angles and bond lengths. The presence of chiral C^{α} atoms in Ala (and in all other amino acids) is responsible for the asymmetric distribution of dihedral angles in part (a), and the presence of C^{β} excludes the portions that are accessible in Gly. (adapted from Page 175 of {Creighton 1993 ID: 495}; original ref {Ramachandran & .& Saisekharan 1968 ID: 502}).*

More detailed calculations using soft potentials (with attractions) confirm the original findings of Ramachandran and coworkers. *Figure II.1.8* displays the $\phi\psi$ energy maps from the molecular mechanics calculations of Flory and coworkers for alanine and glycine {Brant & Flory 1965 ID: 505}{Brant, Miller, et al. 1967 ID: 506}. London attractive dispersion forces and electronic repulsions are included in the calculations, along with electrostatic interactions and intrinsic torsional potentials such that for a given pair of dihedral angles the conformational energy becomes

$$E(\phi_i, \psi_i) = (E_{\phi}^{o}/2)(1 + \cos 3\phi_i) + (E_{\psi}^{o}/2)(1 + \cos 3\psi_i) +$$

$$\sum_{k,l} \left[\frac{a_{kl}}{r_{kl}^{m}} - \frac{c_{kl}}{r_{kl}^{6}} + \frac{332\, q_k\, q_l}{\varepsilon\, r_{kl}^{2}} \right] \qquad \text{(II.1.1)}$$

The first two terms represent the intrinsic torsional potentials of the respective bonds N-C^{α} and C^{α}-C. They are threefold symmetric with minima at *trans*, *gauche*$^{+}$ and *gauche*$^{-}$ states. See the dashed curve in *Figure II.1.5*. The energy barriers were taken as E_{ϕ}^{o} = 1.5 kcal/mol and E_{ψ}^{o} = 1.0 kcal/mol. The contributions of these terms to $E(\phi_i, \psi_i)$ are relatively small. The dominant term is the summation over non-bonded interactions. This term includes all the pairs (k,l) whose distances $r_{kl} = |\mathbf{r}_l - \mathbf{r}_k|$ vary with ϕ_i and ψ_i, in a segment comprised of the investigated residue at the center, and planar peptide units on both sides. a_{kl} and c_{kl} are empirical energy parameters specific to atoms k and l; they equate to the Lennard-Jones potential parameters if the exponent m =12. The last term in the summation describes the electrostatic interaction. The polar groups are approximated therein by assigning partial charges (q_k) of ± 0.281*e* to amide N and H atoms, and ± 0 .394 *e* to carbonyl C and O, where *e* represents the electronic charge. The dielectric permittivity ε is taken as 3.5, and the conversion factor 332 yields the Coulombic interaction in units of kcal/mol, when r_{kl} is in Å and *e* is in units of electron charge. *Table II.1.3* lists the parameters used {Brant & Flory 1965 ID: 505}{Brant, Miller, et al. 1967 ID: 506} in obtaining *Figure II.1.8*. In the same table, the set of parameters proposed by Lifson and coworkers {Warshell, Levitt, et al. 1970 ID: 517}is also presented.

TABLE II.1.3. Non-bonded interaction parameters for conformational energy calculations[a]

	Brant-Miller-Flory[b]			*Warshel-Levitt-Lifson* [c]		
Atom or group	r_k^0(Å)	α_k(Å^3) [d]	N_k [d]	pair[e]	e_{kk}	σ_{kk}
H	1.20	0.42	0.9	H…H	-0.01	2.94
C (carbonyl)	1.70	1.30	5	C…C	-0.19	4.23
Cα	1.70	0.93	5	O…O	-0.23	3.00
N	1.55	1.15	6	N…N	-0.19	3.60
O	1.50	0.84	7			
CH_2	1.85	1.77	7			

[a] Note that the 6-12 Lennard Jones potential can be alternatively written as $E = e_{kl}[-(r_{kl}/\sigma_{kl})^{12} + 2(r_{kl}/\sigma_{kl})^6]$ where e_{kl} and σ_{kl} are the energy and length parameters given by $e_{kl} = -b_{kl}^2/(4a_{kl})$ and $\sigma_{kl} = (2a_{kl}/b_{kl})^{1/6}$.

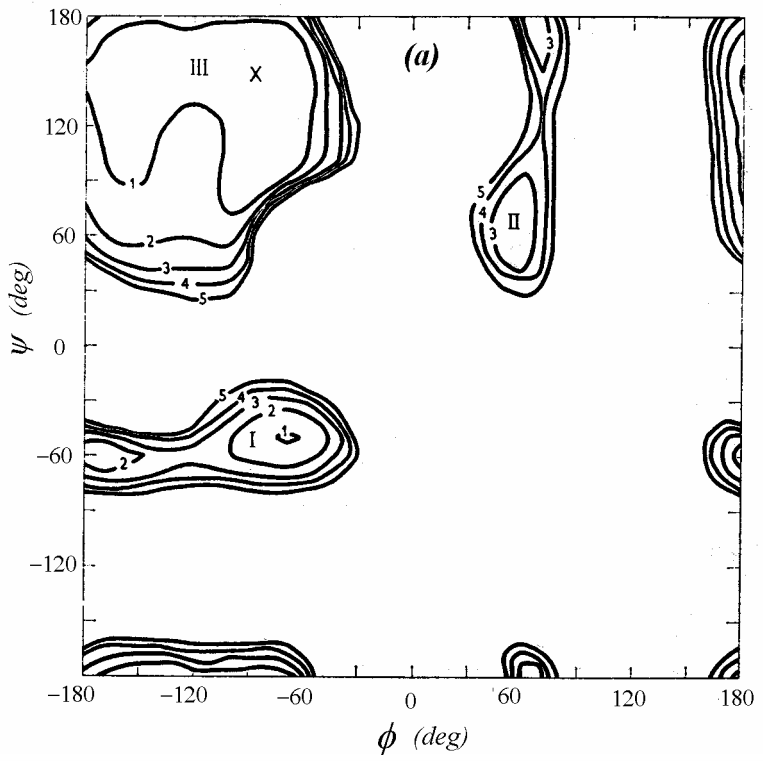
[b] from {Brant, Miller, et al. 1967 ID: 506}.

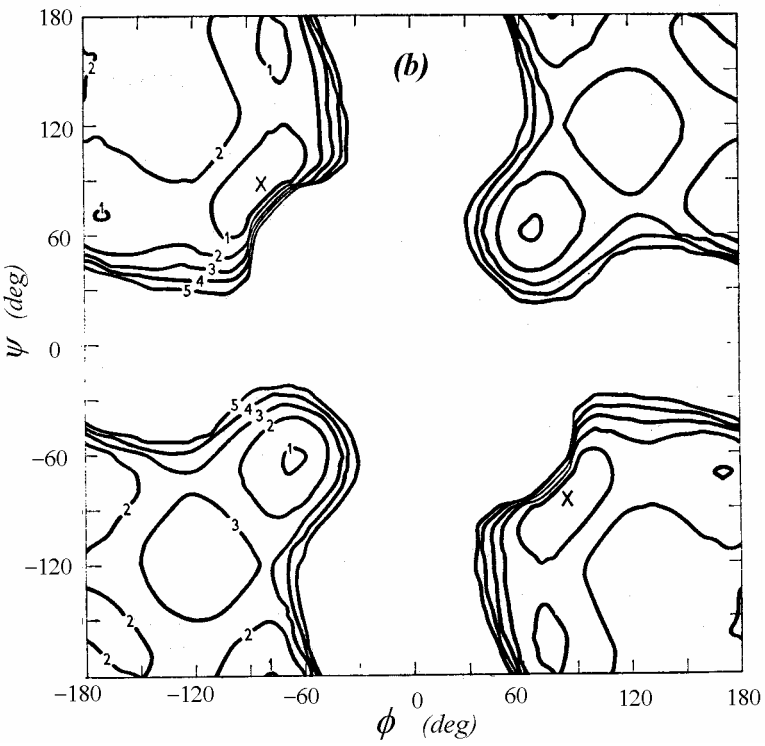
[c] from {Warshell, Levitt, et al. 1970 ID: 517}

[d] α_k is the atomic or group polarizability, and N_k is the effective number of valence shell electrons. c_{kl} and a_{kl} (eq II.1.1) are evaluated using the Slater-Kirkwood equation {Pitzer 1957 ID: 519} and requiring the 6-12 potential to be minimized at $r = r_k^0 + r_l^0$, using m = 12

[e] for contacts between nonidentical types of atoms, use the mixing rules $e_{kl} = (e_{kk}\, e_{ll})^{1/2}$, and $\sigma_{kl} = (\sigma_{kk} + \sigma_{ll})/2$

Figure II.1.8. *(next page) Conformational energy maps for (a) alanine, and (b) glycine calculated by Brant et al. {Brant & Flory 1965 ID: 505} using eq II.1.1 with the parameters listed on the left half of Table II.1.3. Contours are drawn at 1.0 kcal/mol intervals. Three minima, I, II and III are observed in alanine. The lowest energy state is marked as (x). This region is characteristic of β- sheets, whereas the minima I and II correspond to right-handed and left-handed α-helices, respectively. (modified from Figures 7 and 5 on pages 263 and 260, of {Flory 1969 ID: 460})*


(a)
III
II
I
180
120
60
0
−60
−120
−180
−120
−60
0
60
120
180
ψ (deg)
ϕ (deg)


(b)
180
120
60
0
−60
−120
−180
−120
−60
0
60
120
180
ψ (deg)
ϕ (deg)

The energy map for alanine (*Figure II.1.8,* part *(a)*) confirms the general features of the hard core plots from *Figure II.1.7* and shows three favorable regions of torsional angles: one is called the *alpha* region (region I) which includes the α-helical conformations (see below), another is the *beta* region (region III), which includes parallel and anti-parallel β-sheets and the collagen triple helix conformation, and the third is the left-handed α-helix region (region II). ϕ_i values in the range $\phi_i < 0°$ bring the $(CH_3)_i$ and $(N–H)_{i+1}$ groups into proximity without overlap of their van der Waals radii, while $\phi_i > 0°$ give rise to an overlap between $(CH_3)_i$ and the larger polar group $(C= O)_{i-1}$. The lower right quadrant is almost entirely precluded due to the steric clash between $(CH_3)_i$, $(C= O)_{i-1}$ and $(N–H)_{i+1}$.

It is interesting to note that the preference of *L*-polypeptides for right-handed α-helices (region I) over left-handed ones (region II) can be rationalized with regard to the larger size of the carbonyl O compared with the amide H, that restricts the access to region II. However, in proteins, there is a preference for right-handed α-helices even stronger than that indicated on these Ramachandran maps. As will be shown in § II.3.a, α-helices essentially owe their stability to the hydrogen bonds formed between the polar groups $(C = O)_i$ and $(N-H)_{i+4}$, an interaction that is not included in the Ramachandran maps. Likewise, region III is favored by the hydrogen bonds formed between strands in β-sheets. Thus, Ramachandran maps are useful for providing information on the *intrinsic* preferences of amino acids in polypeptides, or on the ranges of *accessible or allowed* dihedral angles; however, the precise *distribution* of dihedral angles in proteins can, and does, depend on non-local interactions (see *Figure I.4.6*).

The above theoretical results are confirmed by experimental data. *Figure II.1.9* shows the (ϕ, ψ) populations observed in proteins averaged over all amino acids except glycine and proline. *Figure II.1.10* shows the populations for some individual amino acids. Glycine and proline are unusual. Because the sidechain of glycine is only a hydrogen atom, glycine is unusually flexible, in conformity with the maps shown here. Glycine populates the left-handed helical region much more than other amino acids. Proline, on the other hand, is unusually rigid because the backbone is locked down by a small hydrocarbon ring that constitutes the side chain.

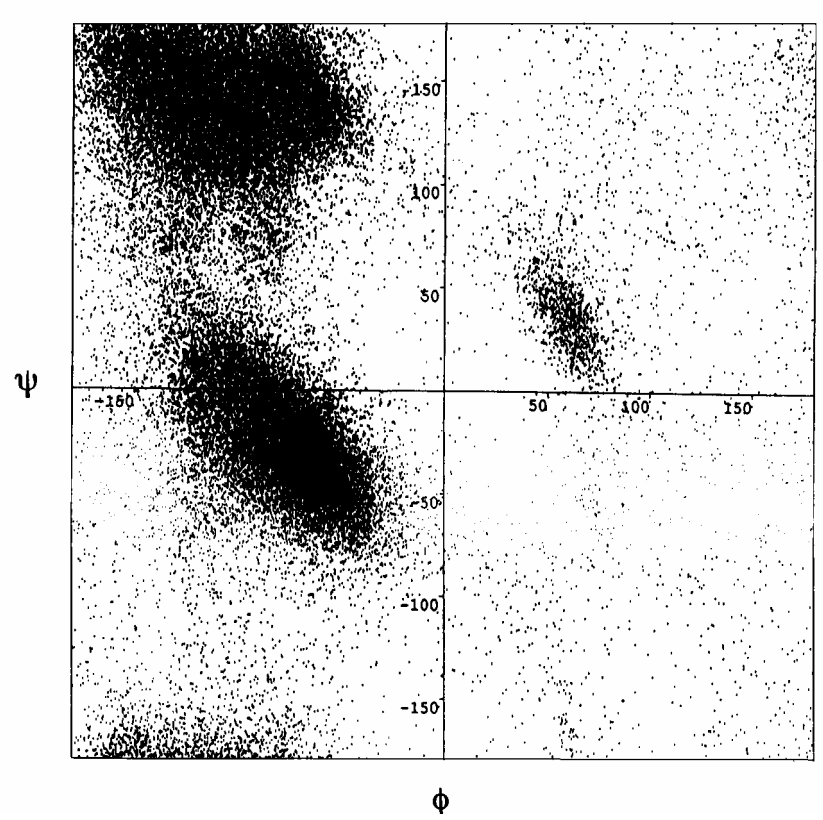


Figure II.1.9. *Ramachandran map of known protein native structures showing the (ϕ, ψ) distribution of all residues, except glycine and proline. Dots represent the observed (ϕ, ψ) pairs in 310 protein structures in the Brookhaven Protein Databank (adapted from {Thornton 1992 ID: 496})*

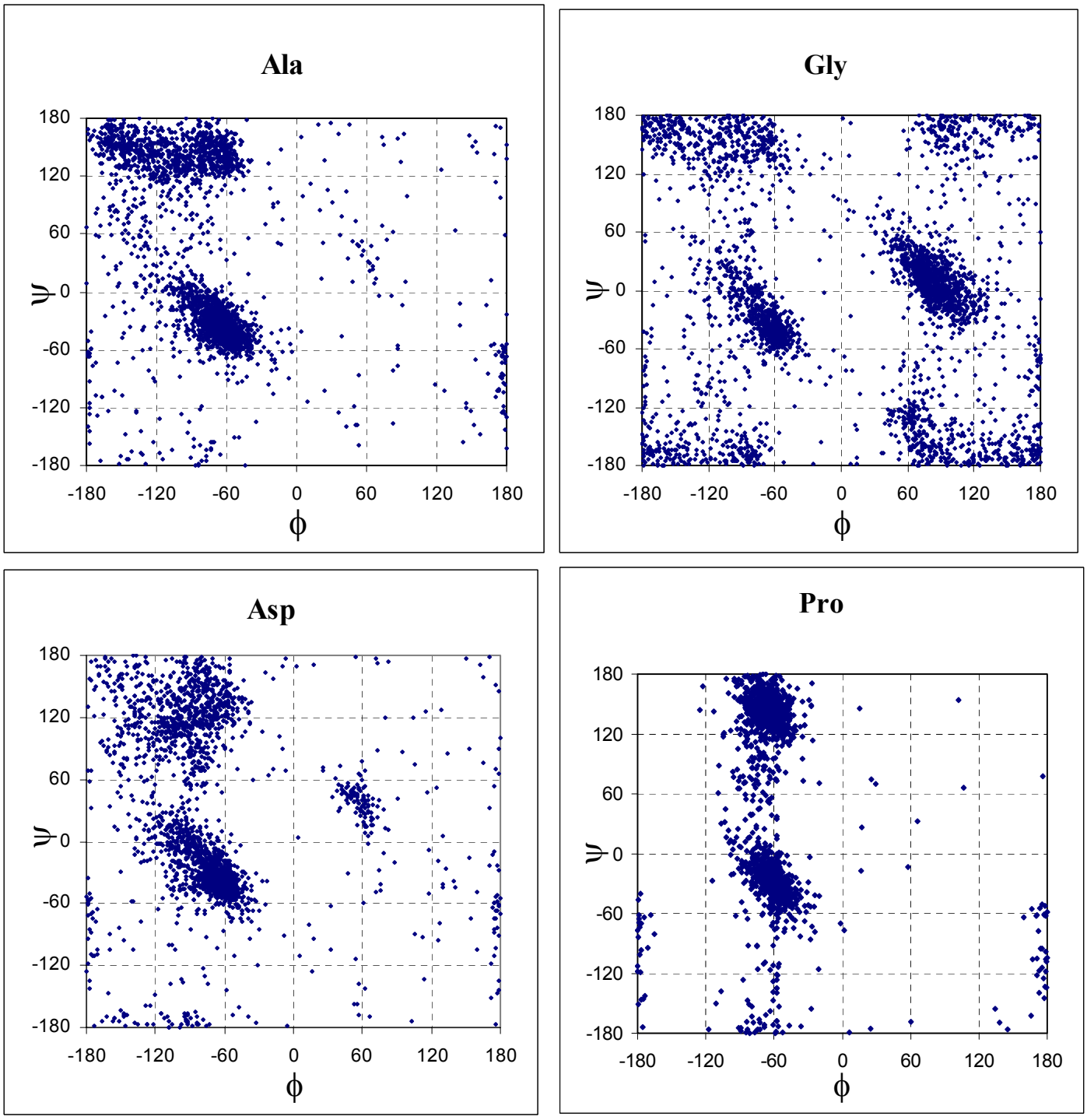


Figure II.1.10. *Ramachandran map of main chain conformations for four residue types, showing (ϕ, ψ) angles for alanine (top left), aspartic acid (bottom left), glycine (top right) and proline (bottom right) in high resolution protein X-ray structures. See the highly expanded area around glycine and its essentially symmetric distribution about ϕ = 0 and ψ = 0. The distribution for Pro is the most highly confined, with ϕ restricted to be near -60° because of the ring.*