Self-Consistent Assignment of Asparagine and Glutamine Amide Rotamers in Protein Crystal Structures

Christian X. Weichenberger\(^1\) and Manfred J. Sippl\(^{1,*}\)
\(^1\) Center of Applied Molecular Engineering
University of Salzburg
Jakob Haringerstraße 5
5020 Salzburg
Austria

Summary

The current protein structure database contains unfavorable Asn/Gln amide rotamers in the order of 20%. Here, we derive a set of self-consistent potential functions to identify and correct unfavorable rotamers. Potentials of mean force for all heavy atoms are compiled from a database of high-resolution protein crystal structures. Starting from erroneous data, a refinement-correction cycle quickly converges to a self-consistent set of potentials. The refinement is entirely driven by the deposited structure data and does not involve any assumptions on molecular interactions or any artificial constraints. The refined potentials obtained in this way identify unfavorable rotamers with high confidence. Since the state of Asn/Gln rotamers is largely determined by hydrogen bond interactions, the features of the respective potentials are of interest in terms of molecular interactions, protein structure refinement, and prediction. The Asn/Gln rotamer assignment is available as a public web service intended to support protein structure refinement and modeling.

Introduction

The amide groups in the side chains of asparagine (Asn) and glutamine (Gln) are specific examples of functional groups that act simultaneously as hydrogen bond donors and acceptors. These amide groups frequently participate in hydrogen bond networks (Yoder et al., 1993), substrate binding, ligand docking (Faham et al., 1996; Evdokimov et al., 2001; Schafer et al., 2004), protein-protein interactions, and catalysis (Battiste et al., 1996; Coleman et al., 1994; Vernet et al., 1995; Raymond et al., 2004), and segments of Gln and Asn residues have been shown to form the cross-β spine of amyloid-like fibrils (Nelson et al., 2005; Sambashivan et al., 2005). The electron density near the nitrogen and oxygen atoms of Asn and Gln amide groups is compatible with two rotamers that are related by a 2-fold symmetry axis. Therefore, electron density maps obtained from X-ray diffraction experiments of protein crystals yield the positions of the oxygen and nitrogen atoms with high precision but not their identity resulting in the perpetual assignment of wrong rotamers.

The Asn and Gln amide assignment problem is the subject of repeated investigations, and as a result, the error rate is well documented (Word et al., 1999; McDonald and Thornton, 1995; Hooft et al., 1996; Word et al. (1999) provide a thorough analysis of Asn and Gln side chain amides found in 100 high-quality protein crystal structures. The authors use a variety of criteria to distinguish correct from incorrect rotamers, including the construction of moveable hydrogen atom positions required for the evaluation of interaction energies. They classify 65% of the 1557 investigated rotamers as clearly correct, 20% as clearly incorrect, and 15% as undecidable. The data set investigated by Word et al. (1999) is restricted to Asn/Gln residues whose B factors are less than 40.

Another source of confident rotamer assignments is provided by Higman et al. (2004). These authors compare Asn and Gln amide rotamers of hen egg-white lysozyme determined by residual dipolar couplings in solution with the corresponding rotamers found in 16 high-resolution protein crystal structures. The authors report confident assignments for 141 amide rotamers, 26 of which disagree with the rotamers found in the respective crystal structures.

Asn and Gln amide rotamers frequently have indistinguishable electron densities, the interactions with neighboring atoms generally yield large energy differences between the two rotamer states, where the predominating interactions are hydrogen bonds (Figure 1). Hydrogen bonds and hydrogen bond networks are ubiquitous in protein structures. Their importance in protein folding and stability is clear, but their functional form, their molecular mechanism, and their strength are debated. Hydrogen bonds form when a strongly electronegative nitrogen or oxygen atom (the hydrogen bond donor) shares a covalently attached proton with a lone pair of electrons of another oxygen atom (the hydrogen bond acceptor). In proteins, hydrogen bond donors and acceptors are always an integral part of the polypeptide backbone or the amino acid side chains, and since the respective molecular environment strongly affects the free energy, there are many individual types of hydrogen bonded interactions in proteins (Deechongkit et al., 2004). The derivation of the physical properties of these interactions from first principles is a difficult task (Finnis, 2004), but quantitative models can be obtained from the reversible work theorem of statistical thermodynamics (Chandler, 1987; Witten and Pincus, 2004) in conjunction with experimentally determined protein crystal structures (Sippl, 1990). The theorem states that the distribution of atoms observed in molecular systems is a direct consequence of the underlying molecular forces, thus promoting the quantitative analysis of atom pair interactions on the basis of radial distribution functions (Sippl, 1990, 1993). Here, we compile potentials of mean force for the interactions of Asn/Gln side chain amides with all other protein atoms, except hydrogens, and compare the two possible
rotamers in terms of the respective energy difference. We find a close correspondence between the rotamer flips suggested by the energy differences with expert annotated data.

Theory

The link between structure and energy is provided by the relationship (Chandler, 1987; Sippl, 1990):

\[ E(a, b, r) = -kT \ln |g(a, b, r)|, \]

where \( E(a, b, r) \) is the generic potential of mean force for the interaction of two particles of types \( a \) and \( b \) at separation \( r \), \( g(a, b, r) \) is the generic radial two particle distribution function, \( k \) is Boltzmann's constant, and \( T \) is the absolute temperature. The function \( E(a, b, r) \) contains annealed contributions specific for the atom pair \( (a, b) \) as well as quenched contributions, which are intrinsic to the covalent structure and compact nature of folded protein molecules. Averaging over all atom pairs \( (a, b) \) yields the quenched contributions in form of the unspecific radial distribution function \( g(r) = \langle g(a, b, r) \rangle_{a,b} \), whereas the annealed contributions are represented by the specific radial distribution functions \( g_s(a, b, r) = g(a, b, r)/g(r) \), and the associated specific potential of mean force is

\[ \varepsilon(a, b, r) = -kT \ln g(a, b, r) + kT \ln g(r) = -kT \ln \left( \frac{g_s(a, b, r)}{g(r)} \right). \]

In terms of the specific potential of mean force, the interaction at \( r \) is attractive if \( \varepsilon(a, b, r) < 0 \) and repulsive if \( \varepsilon(a, b, r) > 0 \).

The total interaction energy of the amide \( N \) and \( O \) atoms is the sum over pairwise interactions with all other atoms in their molecular environment. Denoting the rotamers found in crystal structures of proteins by \( R_i \) and the associated amide nitrogen and oxygen atoms by \( N_i \) and \( O_i \), this energy is

\[ \varepsilon(R_1) = \sum_{j=0}^2 \varepsilon(a_{N_j}, b_j, r_{N_j}) + \sum_{j=k}^3 \varepsilon(a_k, b_j, r_k), \]

where \( h = N_1 \) and \( k = O_1 \) and where the summation is over all atoms \( j \) of atom type \( b_j \) in the environment of the respective atom. In the alternative rotamer, \( R_2 \), the \( N \) and \( O \) atoms swap their positions and the interaction energy, \( \varepsilon(R_2) \), is obtained for \( h = N_2 \) and \( k = O_2 \). The energy difference, \( \Delta \varepsilon = \varepsilon(R_1) - \varepsilon(R_2) \), of the rotamers indicates whether \( R_1 \) or \( R_2 \) is the preferred rotamer.

Results and Discussion

Potentials of mean force are compiled from a database of 833 highly resolved protein crystal structures containing 7670 Asn and 6513 Gln residues, and \( \Delta \varepsilon \) is computed for the corresponding rotamers (see Experimental Procedures). The complete crystal structure is generated from the respective symmetry operations to recover the complete chemical environment of each amino acid residue. An immediate advantage of potentials of mean force is that the modeling of interactions does not require explicit hydrogen atom positions. On the other hand, it is known in advance that the database contains a substantial fraction of unfavorable rotamers, and it is clear that the potentials are affected by these rotamers. To obtain a self-consistent solution, we employ an iterative refinement protocol where all amides with \( \Delta \varepsilon > 0 \) are flipped from rotamer \( R_1 \) to \( R_2 \) and where radial distribution functions and potentials are recomputed after correction. The iteration essentially converges within the first refinement cycle (see Experimental Procedures).

Figure 2A compares the distributions of \( R_1 \) rotamers, i.e., the rotamers found in the crystal structures, as a function of \( \Delta \varepsilon \) before and after refinement. The distribution has two peaks which overlap at \( \Delta \varepsilon = 0 \). The larger peak resides in the region \( \Delta \varepsilon < 0 \), corresponding to amide groups where the \( R_1 \) rotamer is favored by interactions.
Figure 2. Distributions of Rotamer Energies in Protein Crystal Structures
(A) Distributions of 14,183 Asn and Gln side chain amides found in 833 high-resolution protein crystal structures as a function of $\Delta r$, computed from refined (solid line) and unrefined (dashed line) potentials of mean force. The bin size is $|\Delta r| = 10.0$. Refinement results in a general increase of $|\Delta r|$ values and a decrease of the number of rotamers around $\Delta r = 0$. The distribution obtained from the refined potentials has two peaks at $\Delta r = -30$ and $\Delta r = 30$.

(B) The distribution of 1557 Asn and Gln side chain rotamers investigated by Word et al. (1999) plotted as a function of $\Delta r$, where 1006 residues are classified as correct (gray line), 320 as incorrect (dashed line), and 231 as undecidable (black line). The $\Delta r$ values are computed from refined potentials. The distributions of the correct and incorrect rotamers closely match the peaks of the distributions obtained for the 14,183 rotamers in Figure 2A. The distribution of amides corresponding to undecidable cases shows two peaks, resembling the distribution of Figure 2A (solid line). Applying the threshold $|\Delta r| > 6$ to define correct and incorrect rotamers, as already discussed, and using established terminology (Baldi et al., 2000), we obtain a sensitivity of 92.7%, a specificity of 96.7%, and an overall accuracy of 95.8% relative to the classification of Word et al. (1999).

(C) Fraction of incorrect amide rotamers ($\Delta r > 6$) plotted as a function of resolution. Solid bars in dark and light gray correspond to Asn and Gln residues, respectively, where the height of each bar corresponds to the fraction of amides with $\Delta r > 6$ in the respective bin, where the bin size is 0.3 Å. The total number of amides contained in 54,189 protein chains derived from 22,574 protein crystal structures is larger than one million. The bars on the right correspond to the respective fraction of Asn and Gln rotamers found in 3,564 NMR structures.

With increasing resolution, the positions of hydrogen atoms emerge from electron density maps, and it is expected that the number of incorrect Asn and Gln rotamers decreases accordingly. In principle it should be possible to correctly assign the proper rotamer at resolutions better than 1.5 Å, but the necessary atomic occupancy refinements and comparisons of alternative models become prohibitive. The amide groups of Asn and Gln side chains on the surface of proteins frequently participate in intermolecular interactions with neighboring molecules. To assess the impact of intermolecular interactions on the preferred rotamer state, we compiled mean force potentials and the respective $\Delta r$ values from isolated protein chains completely neglecting interactions with neighboring molecules. Comparing these results with those obtained from the complete crystals, we find only 26 cases of 14,183 (0.2%) where interactions with neighboring molecules cause a clear switch in the preferred rotamer. This surprising result indicates that conformations of amide rotamers are predominantly stabilized by intramolecular interactions. We also find that distribution functions obtained from single molecules and whole crystals are very similar, so that potentials of mean force for inter- and intramolecular interactions are essentially equivalent.
a minimum where \(\langle a, b, r \rangle > 0\), indicating that the formation of this bond consumes free energy. The shallow minima near 5 Å for the repulsive interactions (C and D) originate from the covalent structure of the side chain amide. For example, whenever a hydrogen bond contact is formed between ASN-ND2 and VAL-O (A), the neighboring oxygen atom ASN-OD1 necessarily is in the distance range 5.0–6.0 Å from the VAL-O atom (C). The attractive hydrophobic interaction (F) has no barrier.

conformations are generally not contained in standard crystallographic refinement protocols. The dependency of rotamer error rate on resolution is revealed by the \(\Delta E\) values computed for more than a million Asn and Gln residues found in 54,189 protein chains contained in 22,574 protein crystal structures. We exclude Asn and Gln side chains whose amides are closer than 8 Å to nonprotein atoms since potentials of mean force for these interactions are currently not available. As expected, the number of incorrect rotamers decreases with increasing resolution, but even at resolutions better than 1 Å, the average error rate is above 10% (Figure 2C).

The amide assignment problem originates from X-ray analysis of protein crystal structures, but protein structures determined by nuclear magnetic resonance (NMR) techniques have similar error rates of 21% for Asn and 25% for Gln residues (Figure 2C). Protein crystal structures and protein NMR structures are generally derived from experimental constraints where refinement protocols are used to remove errors or to supply information that cannot be obtained from the respective experiment. These refinement protocols frequently employ scoring functions that are based on various models of molecular interactions like Lennard-Jones and Coulomb potentials. Potentials of mean force are based on the principles of statistical thermodynamics. Their functional form and variation is entirely determined by the underlying experimental data.

The specific potentials of mean force for the interactions of Asn and Gln amide nitrogen and oxygen atoms with other hydrogen bond donors and acceptors have the form of a molecular lock or shutter, characterized by an energy barrier with a maximum at \(r_m\) separating a narrow energy well at hydrogen bond contact, \(r_c\), from large distances, \(r_{\infty}\) (Figure 3). The molecular lock is a general model for spatially precise and kinetically stable atom-pair interactions where precision is determined by the tightness of the energy minimum at bond contact, \(r_c\), and stability is mediated by \(r_m = \varepsilon(r_m) - \varepsilon(r_c)\), the difference between the height of the energy barrier at \(r_m\) and the depth of the energy well at \(r_c\). A consequence of the energy barrier is that stable bonds can be formed even if work is expended in the formation of bond contacts (Figure 3B). The barrier creating the lock is a specific feature of hydrogen bonds and other highly polar interactions (Sippl, 1996). Other potentials have comparatively broad attractive basins, and they lack barriers. Typical examples are hydrophobic interactions (Chandler, 2005) between aliphatic carbon atoms (Figure 3F). Such interactions are less precise since the range of preferred
distances is considerably larger as compared to hydrogen bond interactions, and their stability is solely determined by $\epsilon(r)$.

Contrasting potentials of mean force before and after rotamer correction, we find that the major consequence of refinement is a sharpening of the minima at $r_c$ and an increase of energy barriers at $r_m$ (Figure 3). Compared to the rate of 20% of unfavorable rotamers in the original crystal structures, the effect on the potentials is rather small, but these small changes add constructively so that the refinement results in an overall doubling of the absolute values of the energy differences $|\Delta E|$ between rotamers (Figure 2). The resulting potential functions can be used to validate and refine Asn/Gln rotamers in protein crystal structures, which is available as a public web service (http://flipper.services.came.sbg.ac.at).

Experimental Procedures

Compilation of Potentials of Mean Force

Radial distribution functions are obtained from the number densities $n(a,b,r)$ found in protein crystal structures, where $a$ and $b$ are atom types and $r$ is the distance between atoms (Sippl et al., 1996; Sippl, 1996). Distances are sampled in bin sizes of 0.25 Å with sparse data corrections for low counts (Sippl, 1990). The number densities are compiled from a library of 833 protein chains, containing 7,670 asparagine and 6,513 glutamine residues, i.e., a total of 14,183 side chain rotamers (Figure 2). The resulting potential functions can be used to validate and refine Asn/Gln rotamers in protein crystal structures, which is available as a public web service (http://flipper.services.came.sbg.ac.at).

Refinement of Potentials of Mean Force

Potentials are compiled from the protein crystal structures, and the energy differences $\Delta E$ of the two rotamers of all Asn and Gln residues are computed. All rotamers with $\Delta E > 0$ are flipped to the alternative rotamer, and mean force potentials are recompiled from the altered protein crystal structures. The compilation of potentials and the consecutive flipping of high-energy rotamers $|\Delta E | > 0$ constitutes a single refinement cycle. The refinement process is repeated as long as rotamer flips are observed. The first refinement cycle flips 2,802 of the 14,183 Asn and Gln residues. This number drops to 129 in the second cycle. Very few rotamers are flipped in subsequent iterations, and the whole set of rotamers is stable after 14 refinement cycles.

Implementation of Asn/Gln Rotamer Correction

Asn and Gln rotamer correction is available as a web service, accessible at http://flipper.services.came.sbg.ac.at. The service, implemented in Python and C, has a response time in the order of seconds and accepts PDB codes or protein structure files in PDB format.

Acknowledgments

M.J.S. is indebted to Tom Blundell who pointed out that potentials of mean force involving the amides of Asn and Gln residues are necessarily incorrect due to the known error rate in protein crystal structures. The sglite program library, which was used to generate complete crystal structures, was kindly provided by Ralf Grosse-Kunstleve.

Published: June 13, 2006

Received: February 13, 2006

Revised: March 31, 2006

Accepted: April 1, 2006

References


