Stereochemical rules for connecting disjoint protein fragments

Andrzej Kudlicki*, Małgorzata Rowicka* and Zbyszek Otwinowski*

**Dept. of Biochemistry, UT Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, TX 75390-9038, USA*

Abstract. In the process of assembling a protein model, the electron density is fitted with fragments consisting of several peptide units. In low-quality regions of electron density maps, one runs into problems arising from either too many substantially different fragments, or a missing peptide bond. It is then crucial to reduce the set of hypotheses by disqualifying inconsistent ones.

To this end, we characterize the local shape of a main-chain segment by angles between vectors connecting four consecutive C-alpha atoms. The local conformation is described by three quasi-conformation angles: two flat and one dihedral angle. We investigate the probability distribution of these angles and find that the conformational space of the angles is highly restricted. This allows one to connect disjoint fragments more efficiently. Since the procedure of matching hypotheses is computationally expensive, one needs a simple description of the quasi-conformation angle space. We present such a convenient description and compute its parameters using data from the PDB. A complete set of parameters are provided.

BACKGROUND: PROTEIN STRUCTURE DETERMINATION

A very important part in many areas of biochemistry, biophysics and pharmacology is the knowledge of the molecular shapes (structures) of proteins. The structure of a protein may be characterized on several levels. Especially important are the general tertiary and quaternary structure of a protein (*the protein fold*), and the specific arrangement of atoms in the protein's *active site*, the part of the molecule responsible for its main biochemical activity.

Several experimental techniques are used by researchers to determine shapes of biological macromolecules; most prevalently employed is one based on crystallizing them, and analyzing their X-ray diffraction patterns. Protein crystallography is the technique reported for about 82% of the total of over 26,000 structures deposited in the *Protein Data Bank* (PDB) [2, 3], the main public repository of structures of biological macromolecules.

The process of determining the structure of a macromolecule using X-ray crystallography consists of the following steps:

1. Obtaining protein crystals.
2. Collecting diffraction data.
3. Solving the inverse problem in order to determine the electron density of the molecule (*phasing*).

FIGURE 1. A region in an experimental density map with several fitted fragments. The highlighted area shows the interface between two *disjoint* fragments.

4. Building the *protein model*, i.e. placing atoms and chemical bonds in the density map.
5. Refinement – applying corrections to the density map and atomic coordinates.

In this paper, we concentrate on the last steps: building and refining protein models, especially on building the main-chain, or *backbone* of the protein molecule. In our approach¹ to assembling a protein model, the electron density is first fitted with typical protein main-chain fragments. The main-chain fragments in our library consist of several peptide units, and correspond to clusters of small secondary structure backbone motives from known structures, deposited in the PDB. In the next step, we create and evaluate *main-chain hypotheses* as sets of compatible fragments. This step is achieved by means of a genetic algorithm. Finally, we attach side-chain fragments to the previously built main chain.

Assembling protein fragments into hypotheses requires designing a scoring function to assess whether two fragments are compatible or not. In the case of partly overlapping fragments, this task has been solved based on an assessment of how exact the overlap, and a direct calculation of Van der Waals forces. It is often also important to determine, whether two *disjoint* fragments, separated by a small distance, can both belong to the same protein main-chain. A typical case is presented in Fig. 1. Proper distinguishing between pairs of compatible and incompatible disjoint fragments is most important in preventing building main-chain fragments into the protein side-chain. Nearby disjoint fragments may be compatible (this means that they can be included in one main-chain hypothesis), if the distance between them can accomodate one peptide unit. We find, that this condition is not sufficient, but additional requirements have to be satisfied, concerning the orientation of these fragments.

¹ Contrary to most current methods, e.g. the popular ARP/WARP program [5] whose algorithm is based on first locating individual atoms in the density field, and then assembling a model from these atoms. This approach requires high resolution data.

FIGURE 2. Definition of the conformation angles ξ, χ and τ . Spheres represent C_α atoms, and lines the virtual bonds.

We describe these requirements in terms of a probability distribution function (PDF), serving as an additional statistical *prior* used in protein model building. This prior will also be used in refining a main-chain model. The PDF is constructed using data from structures deposited in the PDB.

PARAMETRIZATION

To describe the local conformation of a protein main chain, we start from relative coordinates of 4 subsequent C_α atoms. The conformation of these atoms can be described by three angles, one dihedral and two flat angles. Let us denote the coordinates of subsequent C_α atoms by

$$\mathbf{r}_1 = [x_1, y_1, z_1], \mathbf{r}_2 = [x_2, y_2, z_2], \text{etc.},$$

and their relative positions by

$$\mathbf{d}_1 = \mathbf{r}_2 - \mathbf{r}_1, \mathbf{d}_2 = \mathbf{r}_3 - \mathbf{r}_2, \mathbf{d}_3 = \mathbf{r}_4 - \mathbf{r}_3.$$

The vectors $\mathbf{d}_i$ correspond to so-called *virtual bonds* between the C_α atoms. We describe their relative orientation by three angles χ, ξ and τ (see Fig. 2). The angle χ between $\mathbf{d}_1$ and $\mathbf{d}_2$ is defined by

$$\cos \chi = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{|\mathbf{d}_1| |\mathbf{d}_2|}.$$

The value of χ is between 0 and π . Analogously, we define the angle ξ between $\mathbf{d}_3$ and $\mathbf{d}_4$. The dihedral angle τ is between the planes defined by $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$, and $\mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4$, and has the full range of $0 - 2\pi$. Rather than using the three angles themselves, we parametrize the local backbone conformation by $w = \cos \chi, x = \cos \xi$ and τ . In this set of variables, each volume element

$$dw dx d\tau = d(\cos \chi) d(\cos \xi) d\tau = \sin \chi d\chi \sin \xi d\xi d\tau$$

corresponds to the same surface element in the real space. The variables w and x have values between -1 and 1.

FIGURE 3. An isosurface representation of the probability distribution function for (w, x, τ) , calculated from structures deposited in the PDB.

DATA

To determine the probability distribution function for w , x and τ , we analyzed a subset of the PDB, consisting of structures obtained from X-ray diffraction at resolutions better than 2.5 Å. After removing close homologs, the dataset consisted of 4170 chains, containing over 10^6 amino acids. From the data we created a 3-D histogram, equally spaced in the three variables. This histogram can be used as a prior probability lookup table when assessing compatibility of disjoint fragments.

The data is shown in Fig. 3. An important feature of the plot is that it is not symmetrical with respect to swapping w , and x . This means, that the PDF for χ , ξ , τ contains information that cannot be retrieved from the distribution of Ramachandran angles ϕ and ψ , see e.g. [4].

EXPONENTIAL MODEL

The histogram representation of the PDF suffers from several drawbacks. First, it is exactly zero in bins, where no data point fell in, which makes the prior too prohibitive for practical use. Second, it is nondifferentiable and as such inconvenient for refinement procedures, which usually involve derivatives. Also the large number of parameters, corresponding to the number of bins is not practical, as using it requires distributing a large data array with the modeling software. Hence, it arises the need for an analytical

description of the probability distribution.

To find a functional form of the PDF we will use the Maximum Entropy Principle. To represent the uninformed prior, we first construct a set of orthonormal base functions on $[-1 \dots 1] \times [-1 \dots 1] \times [0 \dots 2\pi]$. We choose them as

$$C_{ijk}(w, x, \tau) = \mathcal{L}_i(w) \mathcal{L}_j(x) \cos(k\tau)$$

and

$$S_{ijk}(w, x, \tau) = \mathcal{L}_i(w) \mathcal{L}_j(x) \sin(k\tau),$$

where i, j, k are integers, and $\mathcal{L}_i(\cdot)$ denotes the Legendre polynomial of the i -th order. The PDB data can be reduced to moments in this basis:

$$\begin{aligned} M_{ijk}^{(c)} &= \frac{1}{N} \sum_{n=1}^N \mathcal{L}_i(w_n) \mathcal{L}_j(x_n) \cos(k\tau_n), \\ M_{ijk}^{(s)} &= \frac{1}{N} \sum_{n=1}^N \mathcal{L}_i(w_n) \mathcal{L}_j(x_n) \sin(k\tau_n). \end{aligned}$$

As the functional form of constraints for maximum-entropy modeling, we choose the values of the same moments, calculated from the investigated PDF with respect to the above base functions:

$$\begin{aligned} \mu_{l,n,m}^{(c)} &= \int_{-1}^1 dw \int_{-1}^1 dx \int_0^{2\pi} d\tau C_{ijk}(w, x, \tau) p(w, x, \tau | \text{coeff}) \\ \mu_{l,n,m}^{(s)} &= \int_{-1}^1 dw \int_{-1}^1 dx \int_0^{2\pi} d\tau S_{ijk}(w, x, \tau) p(w, x, \tau | \text{coeff}) \end{aligned}$$

Using Lagrange multipliers one can show that for such constraints the appropriate probability distribution p is an exponential function with the exponent being a linear function of the constraints:

$$p(w, x, \tau | \text{coeff}) = \exp \left(-\lambda - \sum_{i,j \leq S^l, k \leq S^r} N(a_{ijk}, b_{ijk}, w, x, \tau) \right), \quad (1)$$

where

$$\begin{aligned} N(a_{ijk}, b_{ijk}, w, x, \tau) &= \\ &= a_{ijk} \mathcal{L}_i(w) \mathcal{L}_j(x) \cos k\tau + b_{ijk} \mathcal{L}_i(w) \mathcal{L}_j(x) \sin k\tau. \end{aligned}$$

λ is a normalization constant, introduced to assure that

$$\int_{-1}^1 dw \int_{-1}^1 dx \int_0^{2\pi} d\tau p(w, x, \tau | \text{coeff}) = 1,$$

and a_{ijk} and b_{ijk} are the parameters of the distribution. S^l and S^r denote the maximum orders to which the Lagrange and trigonometric expansions are calculated.

We assume these coefficients to be *a priori* statistically independent. To keep optimizations convergent, a_{ijk} and b_{ijk} are assigned a Gaussian prior probability distribution

$$p(a_{ijk}) = \sqrt{\frac{c}{\pi}} e^{-c (a_{ijk})^2},$$

with $c = 0.25$. To find values of the coefficients, we use Bayes' theorem:

$$\begin{aligned} p(\{a_{ijk}, b_{ijk}\}_{i,j \leq S^l, k \leq S^t} | w, x, \tau) &= \\ &= \frac{p(\{a_{ijk}, b_{ijk}\}_{i,j \leq S^l, k \leq S^t}) \prod_{n=1}^N p(w_n, x_n, \tau_n | \{a_{ijk}, b_{ijk}\}_{i,j \leq S^l, k \leq S^t})}{\prod_{n=1}^N p(w_n, x_n, \tau_n)}. \end{aligned}$$

This task boils down to maximizing the following expression:

$$\lambda + \sum_{i,j \leq S^l, k \leq S^t} \{a_{ijk} M_{ijk}^{(c)} + b_{ijk} M_{ijk}^{(s)}\} - c \sum_{i,j \leq S^l, k \leq S^t} \{(a_{ijk})^2 + (b_{ijk})^2\},$$

where

$$\lambda = \log \left\{ \int_{-1}^1 dw \int_{-1}^1 dx \int_0^{2\pi} d\tau \exp \left(- \sum_{i,j \leq S^l, k \leq S^t} N(a_{ijk}, b_{ijk}, w, x, \tau) \right) \right\}.$$

The integral for λ is calculated numerically, and the whole optimization is performed using a modification of the conjugate gradient routine from the Numerical Recipes[1]. The required gradients were computed analytically, requiring only one simple numerical integration for evaluating $\partial \lambda / \partial a_{ijk}$.

RESULTS AND CONCLUSIONS

After performing the minimization for different values of S^l and S^t , we find, that the probability distribution described by Equation (1) reproduces the locations and magnitudes of the major peaks of the probability distribution when four orders are computed in both Legendre polynomials and trigonometric functions. To get a better representation, we compute the orders 0...6 in Legendre, and 0...3 in trigonometric expansion. The thus constructed PDF provides a better approximation (the model is plotted in Fig. 3). However it still misses some details, e.g. narrow peaks corresponding to hairpins. In a next stage of the project, we plan to use a combination of uninformed and informed prior to accomodate these features. The values of a_{ijk} and b_{ijk} for four orders of expansion are given in Table 1.

TABLE 1. Values of the parameters a_{ijk} and b_{ijk} calculated for four orders of both trigonometric and Legendre expansion.

i	0	0	1	1	2	2	3	3
	a_{ijk}	b_{ijk}	a_{ijk}	b_{ijk}	a_{ijk}	b_{ijk}	a_{ijk}	b_{ijk}
j=0	0.0000	0.0000	0.4795	0.0000	-0.6496	0.0000	-0.3305	0.0000
	-0.0185	0.3406	-0.1722	-0.0246	-0.1468	-0.1848	0.1011	0.0330
	-0.0228	0.9266	0.0878	0.1250	0.0996	-0.3866	-0.0259	-0.1253
	-0.5055	0.1766	-0.0412	-0.0285	0.2165	-0.1214	0.0173	0.0132
j=1	0.4795	0.0000	0.1884	0.0000	-0.0819	0.0000	-0.1212	0.0000
	-0.1988	-0.1541	-0.1031	-0.0548	0.0086	0.0394	0.0597	0.0528
	0.0640	0.1783	0.0451	0.0573	0.0144	-0.0477	-0.0138	-0.0509
	-0.0123	-0.0416	-0.0194	-0.0269	-0.0172	-0.0006	-0.0002	0.0216
j=2	-0.6491	0.0000	-0.0809	0.0000	0.2085	0.0000	0.0634	0.0000
	-0.1660	-0.3318	-0.0034	-0.0448	0.0806	0.1330	-0.0009	0.0360
	0.0526	-0.3212	-0.0086	-0.0110	-0.0332	0.1600	0.0041	0.0187
	0.2524	-0.1300	0.0048	-0.0103	-0.1272	0.0617	-0.0095	0.0123
j=3	-0.3303	0.0000	-0.1184	0.0000	0.0668	0.0000	0.0785	0.0000
	0.1224	0.0491	0.0632	0.0175	-0.0073	-0.0145	-0.0405	-0.0216
	-0.0651	-0.1240	-0.0356	-0.0315	-0.0017	0.0410	0.0149	0.0315
	0.0231	0.0213	0.0153	0.0149	0.0054	0.0006	-0.0026	-0.0138

ACKNOWLEDGMENTS

This work was supported by NIH grant No. 53163. The authors thank Mirek Gilski, Jan Zelinka and a referee for helpful discussions and suggestions. We are grateful to E.T. Jaynes Center for generous travel support.

REFERENCES

1. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, 1992.
2. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.
3. Protein Data Bank. <http://www.rcsb.org/pdb>.
4. M. Rowicka and Z. Otwinowski, Maximum-Entropy modeling of torsion angle probability distribution in proteins, *The 23rd annual conference on Bayesian methods and maximum entropy in science and engineering*, Jackson Hole, Wyoming, 2004.
5. Lamzin, V.S., Perrakis, A. & Wilson, K.S. (2001) *The ARP/WARP suite for automated construction and refinement of protein models*. In Int. Tables for Crystallography. Vol. F: Crystallography of biological macromolecules (Rossmann, M.G. & Arnold, E. eds.), Dordrecht, Kluwer Academic Publishers, The Netherlands, pp. 720-722