OPEN ACCESS

GAVIN PUBLISHERS

Current Research in
Bioorganic & Organic Chemistry

Short Communication

Li W. Curr Res Bioorg Org Chem: CRBOC-112.
DOI: 10.29011/CRBOC -112. 100012

NMR-Observed Atomic Bond Length Stability Supports a Dimensionality Shift in Protein Main Chain 3D Structure Description and Representation

Wei Li[*]

Department of Pharmacology, Shantou University Medical College, Shantou City, Guangdong Province, P. R. China

[*]**Corresponding author:** Wei Li, Department of Pharmacology, Shantou University Medical College, No. 22, Xinling Road, Shantou City, Guangdong Province, P. R. China. Tel: +8615817969015; Email: liweiqidong@stu.edu.cn ; wli23@126.com

Citation: Wei Li (2018) NMR-Observed Atomic Bond Length Stability Supports a Dimensionality Shift in Protein Main Chain 3D Structure Description and Representation. Curr Res Bioorg Org Chem: CRBOC-112. DOI: 10.29011/CRBOC -112. 100012

Received Date: 07 August, 2018; **Accepted Date:** 21 August, 2018; **Published Date:** 27 August, 2018

Abstract

To date, the Cartesian (x, y, z) coordinate system is the default system in the Protein Data Bank to specify atomic positions in protein structures. Presented here is an alternative spherical coordinate system approach for a three-dimensional lossless deconstruction of protein main chain structures experimentally determined by NMR spectroscopy. To the default Cartesian system and a previously reported global spherical approach, this alternative local spherical approach provides a geometric description of the three-dimensional structure of protein main chains, which requires only two parameters (θ and φ), instead of the default three, i.e., x, y, z. Intrinsically a simpler approach than the default and previously reported approaches, this 2018 one induces a dimensionality shift from three to two, allowing it to find its potential application in significantly increasing the efficiency of protein structure-centered researches.

Keywords: Atomic Bond Length Stability; Dimensionality Shift; Frequency Distribution; Protein Main Chain Structure; Spherical Coordinate System

Experimentally Determined Three-Dimensional Structures of Protein

Proteins are the fundamental units of all living cells. Protein Data Bank (PDB) is the main international repository for protein 3D structures [1]. While the number of experimentally determined protein structures keeps increasing, with the number of Cryo-EM structures also on the rise, X-ray crystallography and NMR spectroscopy remain two main biophysical tools, both with strengths and weaknesses, and reciprocal complementarity in structural biology researches. Among currently available biophysical tools, NMR spectroscopy is able to provide unique access to atomic-level dynamic behavior of protein molecules in solution at physiological conditions (such as temperature and pH, etc.), as a result, this article focuses on protein structures experimentally determined by NMR spectroscopy.

Two Geometrical Systems for protein 3D Structure Description and Representation

In PDB-format files, Cartesian coordinate system (xyz coordinates) is the default system to define atomic positions in protein structures.

In 2011 and 2015, it was proposed for the first time that protein 3D structures be represented in spherical coordinates (r, φ, θ), with an aim to express all protein 3D structures deposited in the PDB in spherical coordinates [2]. Indeed, this 2011-2015 approach is a global spherical coordinate system one, where the protein geometric centroid is taken as the unique original point for all atoms in a protein molecule, resulting in two applications, i.e., the separation of the protein outer layer from its inner core, and the identification of protrusions and invaginations on the protein surface [2].

Here, this article proposes an alternative local spherical coordinate system approach (refereed to below as the 2018 ap-

proach) to establish a two-parameter description of protein main chain 3D structure. In protein, each amino acid residue's main chain is constituted by NN (amide nitrogen), CA and the carboxyl carbon (with a double-bonded oxygen), i.e., N-CA-CO. To simplify the representation of protein 3D structure, hiding the amino acid residue side chains leaves only the backbone, i.e., three main chain atoms and another three backbone atoms, Hα, HN (amide hydrogen) and O=C (the carboxyl oxygen with a double-bonded carbon). Thus, protein main chain geometry can be abstracted as a linear sequence, -[Ni-CAi-COi-Ni+1-CAi+1-COi+1]-, where $0 < i < n$, n represents the number of amino acid residues in a protein. In this linear sequence, with an arbitrary position of the starting atom j specified in a coordinate system, the next j + 1 atom's position can be determined with three spherical coordinate system parameters, r (the bond length), θ (the polar angle) and φ (the azimuthal angle), which constitutes a local vector in a three-dimensional space from atom j as the beginning (i.e., the original) point to atom j+1 as the ending point, where $0 < j < m$, m represents the number of atoms in protein main chain. Thus, one distinction between this 2018 approach (a local spherical coordinate system approach) and the 2011-2015 approach (a global spherical coordinate system approach) is that no atomic bonding information is included in the 2011-2015 approach, while every bit of the atomic bonding information is included in the latter 2018 approach towards the description and representation of protein main chain 3D structure.

Materials and Methods

As of March 12, 2018, the PDB database contains 10656 entries for protein structures determined by NMR spectroscopy. All 10656 entries of the NMR structures were downloaded from the PDB website, with a total size of PDB files at approximately 22.2 GB. During the data processing, a set of python scripts were employed (please contact the corresponding author for the python scripts). First, a python script was employed to take the 10656 NMR PDB file with multiple models (i.e., NMR ensemble) and splits each model into individual PDB files, which yielded 191490 new PDB files. Next, another python script was employed to extract the Cartesian coordinates, which are subsequently processed by another python script to convert Cartesian coordinates to spherical coordinates. Finally, three further python scripts were used in a set of statistical analysis and plotting processes and data fitting for the 191490 PDB files.

A Local Spherical Coordinate System-Directed Modeling of Protein Main Chain Geometry

As discussed above, protein main chain geometry can be abstracted as a linear atom sequence, -[Ni-CAi-COi-Ni+1-CAi+1-COi+1]-, where $0 < i < n$, n represents the number of amino acid residues in a protein. Thus, there briefly are three types of directly bonded atom pairs in protein main chains,

- The atom pair of Ni and CAi
- The atom pair of CAi and COi
- The atom pair of COi and Ni+1

As a result of the three types of directly bonded atom pairs and the three spherical parameters, there are twelve variables in total, as listed below in (Table 1).

Fig.	Subfig.	Atom Pair	Variable (X-axis)	Frequency distribution modeling
1	1	N_i,CA_i	r = 1.46 ± 0.06 (Å)	$x = 1.46$
1	2	CA_i,CO_i	r = 1.53 ± 0.06 (Å)	$x = 1.53$
1	3	CO_i,N_{i+1}	r =1.35 ± 0.62 (Å)	$x = 1.35$
1	4	All pairs combined	r = 1.45 ± 0.37 (Å)	$x = 1.45$
2	1	N_i,CA_i	ϑ = 1.56 ± 0.69 [Radian]	$y = -0.0007 + 0.0210x - 0.0071x^2$
2	2	CA_i,CO_i	ϑ = 1.56 ± 0.69 [Radian]	$y = -0.0007 + 0.0209x - 0.0070x^2$
2	3	CO_i,N_{i+1}	ϑ = 1.57 ± 0.68 [Radian]	$y = -0.0008 + 0.0211x - 0.0069x^2$
2	4	All pairs combined	ϑ = 1.56 ± 0.69 [Radian]	$y = -0.0008 + 0.0210x - 0.0070x^2$
3	1	N_i,CA_i	φ = 0.01 ± 1.81 [Radian]	$y = 0.01$
3	2	CA_i,CO_i	φ = 0.01 ± 1.81 [Radian]	$y = 0.01$
3	3	CO_i,N_{i+1}	φ = 0 ± 1.82 [Radian]	$y = 0.01$
3	4	All pairs combined	φ = 0 ± 1.81 [Radian]	$y = 0.01$

Table 1: A list of figures and subfigures that illustrate the statistics of the three spherical coordinate system parameters. For the 191490 PDB files, this table summarizes the results of the statistical analysis and modeling of protein main chain geometry as experimentally determined by NMR spectroscopy, including the average values and the standard deviations of the twelve variables, and twelve equations (representing the frequency distributions) from the data fitting processes for the twelve variables.

As shown by the four frequency distribution plots in (Figure 1), the atomic bond lengths (as experimentally determined by NMR spectroscopy) of the three types of atomic pairs appear rather stable, all with sharp peaks centering at the average r values listed in Table 1. In light of this experimentally observed atomic bond length stability, a dimensionality shifts from three to two arises from this 2018 approach, which effectively allows a two-parameter (only φ and θ, instead of the xyz coordinates) geometric description of the three-dimensional structure of protein main chains. Intrinsically a simpler approach, this 2018 one causes a dimensionality reduction, which can be of potential significance for a wide range of protein structure-centered research fields.

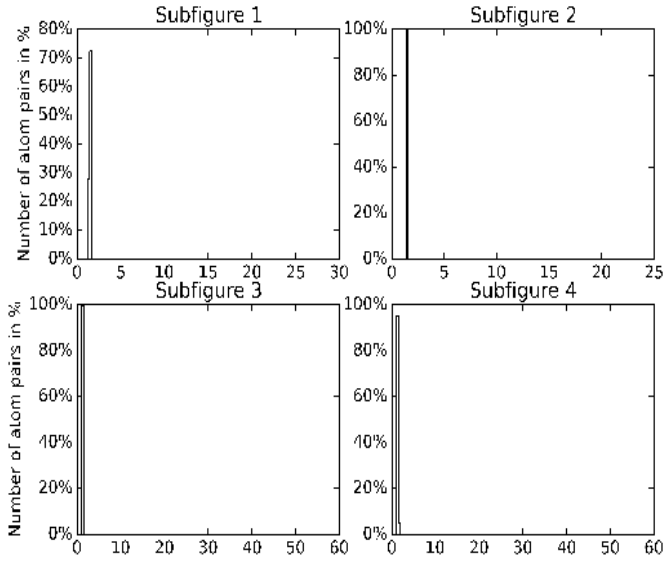

Figure 1: Frequency distribution of r for the protein main chain structures. In this histogram, the number of bins is 100. In this figure, the parameter along with its unit for x-axis is defined in Table 1.

With the results of statistical analysis in place, I present below the result from a set of statistical modeling of the protein main chain geometry as determined by (mainly solution-state) NMR spectroscopy. In Table 1, a set of equations are listed to model the frequency distributions of θ and φ. From a close visual inspection of (Figure 2) and a quantitative analysis, the frequency distribution of θ exhibits a largely symmetric parabolic pattern, with its value ranging from 0 to π, the distribution frequency of θ reaches its peak when θ ≈0.5 π, i.e., the local vector starting from the original atom j to the ending atom j+1 is perpendicular to the xy plane, either parallel (upwards) or anti-parallel (downwards) to the z axis.

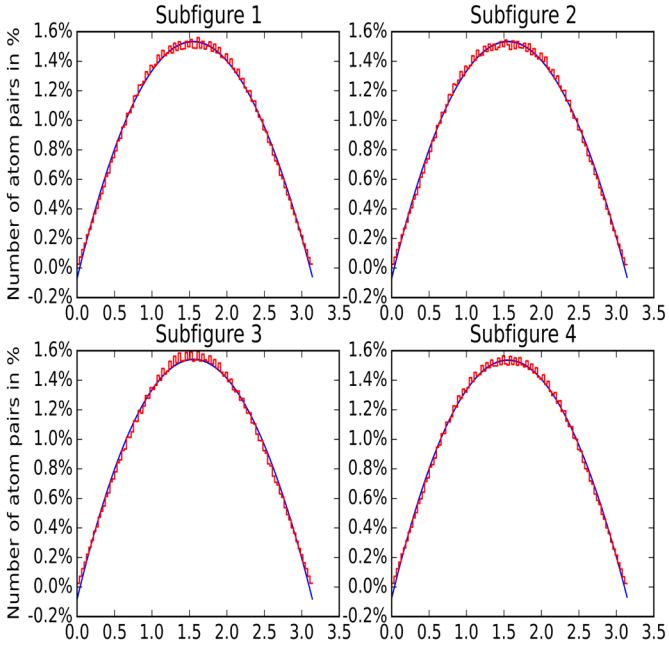

Figure 2: Statistical models (blue) and frequency distributions (red) of θ. In this histogram, the number of bins is 100. In this figure, the parameter along with its unit for x-axis is defined in Table 1.

From a visual inspection of Figure 3, the frequency distribution of φ appears largely random, with its value ranging from -π to π, i.e., for the local vector which starts from the original atom j to the ending atom j+1, its projection in the xy plane is largely random.

After the data fitting process as shown in (Figures 2 and 3), a set of Chi-square tests (the goodness-of-fit test) were conducted to examine the fitness between the observed distribution and the expected distribution. In the Chi-square tests, the p-values were found to be 1.0 for all mathematical models listed in Table 1, i.e., the observed distributions of the two spherical parameters fit to the expected distributions as described using the mathematical equations in Table 1, and statistically the fitness is totally acceptable for both θ and φ.

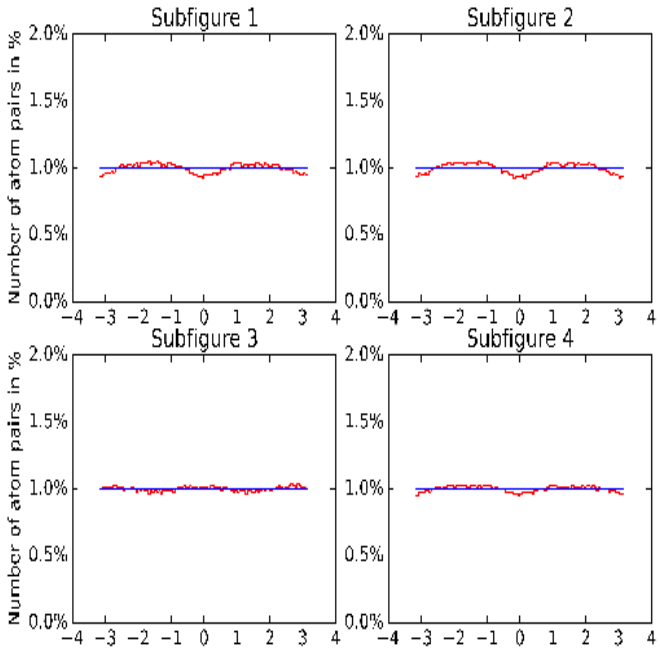

Figure 3: Statistical model (blue) and frequency distribution (red) of φ. In this histogram, the number of bins is 100. In this figure, the parameter along with its unit for x-axis is defined in Table 1.

Conclusion

To the default Cartesian coordinate system of the PDB database and the 2011-2015 global approach, this article proposes an alternative local spherical coordinate system approach to describe protein main chain geometry with only two parameters (θ and φ), resulting in a dimensionality shift from three to two and consequently a viable and simpler approach for protein main chain structure description and representation. Intrinsically a simpler approach with this dimensionality reduction, this 2018 one can find its potential application in increasing the efficiency of protein structure-centered research fields, such as protein structure alignment [3], comparison [4,5] and molecular dynamics simulations.

Discussion

In light of the basic chemistry dictations that the distances between directly bonded atoms should be relatively constant throughout the protein structure, it is conceivable that this 2018 approach be applied to describe and map protein side chain 3D structure, too, provided that the atomic bonding pattern is clearly and sequentially predefined for the side chain3D structure of each and every amino acid residue, as is done here with the linear atom sequence mentioned above for protein main chain 3D structure description and representation.

In the four histograms in Figure 1, however, there are a range of outlying r values, according to the computational analysis of the protein main chain atomic bond length distributions. For example, in PDB file 1bhb.pdb, between Val34 and Pro37, there is a gap of two amino acid residues, whose atomic coordinates information is missing. As a result, the r value for the atom pair CO34-N37 was calculated to be 24.876 Å. As another example, in PDB file 2rop.pdb, between Leu90 and Thr120, there is a gap of 29 amino acid residues, whose atomic coordinates information is missing, too. As a result, the r value for the atom pair CO90-N120 was calculated to be 50.102 Å. Both examples arise from the gaps in the experimentally determined protein structures, calling for continued comprehensive (instead of fragmented [6] or with gaps) experimental structure determination for the proteios building block of life.

Conflict of Interest and Ethical Statement

Regarding the publication of this article, the author declares that there is no conflict of interests, no ethical approval is required.

References

1. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nature Structural & Molecular Biology 10: 980.
2. Reyes VM (2015) Implementation of the Spherical Coordinate Representation of Protein 3D Structures and its Applications Using FORTRAN 77/90 Language. Quantitative Biology.
3. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 33: 2302-2309.
4. Betancourt MR, Skolnick J (2001) Universal similarity measure for comparing protein structures. Biopolymers59: 305-309.
5. Ortiz AR, Strauss CEM, Olmea O (2009) MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. Protein Science 11: 2606-2621.
6. Li W (2017) How do SMA-linked mutations of *SMN1* lead to structural/functional deficiency of the SMA protein? PLOS ONE 12: e0178519.