Half-a-century burial of ρ, θ and ϕ in PDB

Wei Li *[1]

[1]*Institute of Special Environmental Medicine, Nantong University, No. 9, Seyuan Road, Nantong City, 226000, Jiangsu Province, P. R. China*

February 9, 2021

Abstract

Since the launch of Protein Data Bank (PDB) in 1971, Cartesian coordinate system (CCS) has been the default approach to specify atomic positions in biomolecular experimental structures with X, Y and Z. In 2020, a local spherical coordinate system (LSCS) approach was proposed as an alternative to CCS, i.e., ρ, θ and ϕ. Recently, the continued application of deep learning technology in protein structure prediction (PSP) saw a leap forward in the accuracy of PSP, as evidenced by AlphaFold of Google's DeepMind. However, there still is room for the improvement of the performances of PSP algorithms to date. Given that geometrically, CCS and LSCS are like the two sides of a coin, this short article puts forward a hypothesis that the time is now ripe to end the half-a-century burial of ρ, θ and ϕ in PDB, and use them as LSCS features for the design of novel PSP algorithms in future.

Keywords: Protein structure prediction; Deep learning; AlphaFold; Cartesian coordinate system; Spherical coordinate system

*Corresponding author

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2021 doi:10.20944/preprints202103.0590.v1

Atomic position specification: from Cartesian to spherical coordinate systems

To date, it has been half a century since the launch of Protein Data Bank (PDB) in 1971 [1, 2]. Meanwhile, it has been half a century, too, since Cartesian coordinate system (CCS) has been the default approach to specify atomic positions in PDB-format coordinate files. Geometrically, CCS and SCS are like the two sides of a coin, i.e., two interconvertible geometric coordinate systems for atomic position specification. In 2020, a local spherical coordinate system (LSCS) approach was proposed [3] as an alternative to the default CCS approach and a global SCS (GSCS) approach proposed a decade ago [4]. In comparison, the 2020 LSCS approach requires only two geometric parameters (θ and ϕ), instead of three parameters (X, Y and Z) as required by the default CCS approach, because it takes as ρ the equilibrium atomic bond length, which is defined as the inter-nuclear distance at which the system energy minimum occurs [5, 6].

As discussed in [3], the time is now ripe to flip the coin over, and take a look at the other side of it, i.e., the LSCS approach, which possesses an intrinsically lower degree of descriptional complexity than those of the default half-a-century old CCS approach and the 2011 GSCS approach [4]. Additionally, the LSCS approach constitutes a potentially useful alternative for protein structure description and feature extraction for protein structure prediction (PSP), which has been a grand challenge in computational biophysics for decades, owing to its intrinsic scientific interest and also to its potential applications in structural biology-related fields [7].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2021 doi:10.20944/preprints202103.0590.v1

Application of big scientific data and AI in PSP

With the advent of AlphaFold by Google's DeepMind [8], 2020 saw a watershed moment in the improvement of PSP accuracy [9, 10, 11, 12]. In fact, given the continued growth in the number of available protein sequences and experimentally determined structures, big scientific data and artificial intelligence (AI, such as deep learning) are to play a role in the design of increasingly robust, accurate and fast PSP algorithms [11]. Additionally, it is possible, too, that the coarse structures (with $\sim$ 4 Å resolution) of most proteins that consist of a single folded domain will become available in the next decade from computational structural predictions [13]. Such broad availability of structural information might transform the life sciences, just as sequence information did in the preceding decades [13], leading to a golden age of structural biology, where quantitative, mechanistic and biophysical basis for life sciences is broadly and firmly grounded in experimental structural observations.

In the application of big scientific data and AI in PSP, nonetheless, challenge remain still. Take AlphaFold for example. Algorithmically, AlphaFold predicts the probabilities of residues being separated by different distances. Because probabilities and energies are interconvertible, AlphaFold predicts an energy landscape, making it possible to use a simple procedure to find the most favorable conformation, rather than the complex search algorithms employed by other methods [9]. Nonetheless, AlphaFold is not accurate enough yet for some applications, such as working out the catalytic mechanisms of enzymes or how drugs bind to proteins, which both typically require 2-3 Å resolution. Furthermore, although AlphaFolds search procedure is much simpler than most modern methods, it can still take tens to hundreds of hours for one PSP. In applications such as protein design, which require a series of protein sequences to be modeled structurally, the lack of speed clearly is an impediment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 March 2021 doi:10.20944/preprints202103.0590.v1

ρ, θ and ϕ: missing pieces in the PSP puzzle

Mathematically and biophysically,

1. the distances between points determine their relative locations, be the coordinate system Cartesian or spherical.

2. in CCS, predictions of distances (calculated from X, Y and Z) can predict protein structure and energy landscape of the molecular system [9].

3. in LSCS [3], predictions of distances and angles (calculated from ρ, θ and ϕ) can predict protein structure and energy landscape of the molecular system.

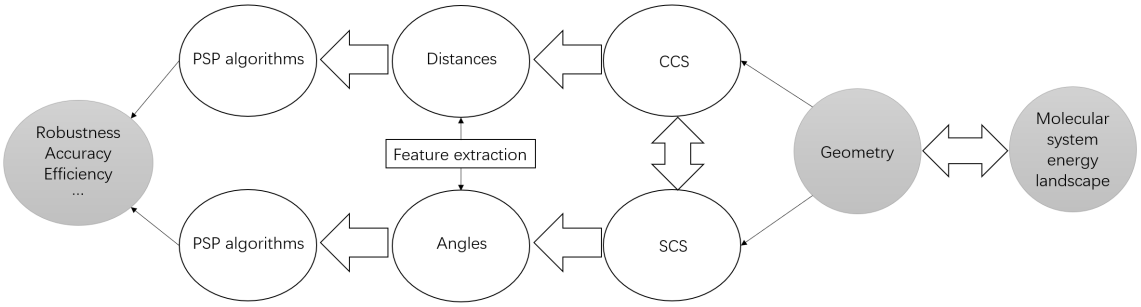

Figure 1: ρ, θ and ϕ are the missing pieces in the PSP puzzle. This flowchart highlights: 1), geometry and energy landscape are interconvertible; 2), geometrically, CCS and SCS are like the two sides of a coin; 3), distances and angles are interconvertible both in CCS and LSCS; 4), both distance and angle could be used as LSCS features in the design of novel PSP algorithms to further improve their performances [9].

In brief, this short article argues that ρ, θ and ϕ have been the missing pieces in the PSP puzzle for half a century since the launch of PDB [2], as illustrated in the minimalist flowchart (Figure 1) of the two geometric approaches (the default CCS and the LSCS [3]) towards feature extraction in the design of PSP algorithms.

 doi:10.20944/preprints202103.0590.v1

A hypothesis

The time is now ripe for the structural biology community to end the half-a-century burial of ρ, θ and ϕ in PDB, such that the trapped values of the three LSCS features could be released towards the improvement of the performance (Figure 1) of PSP algorithms [9, 10, 14] in future.

Acknowledgements

The author thanks the whole structural biology community, who have contributed pricelessly to the continued development of Protein Data Bank [2].

Conflict of Interest

None.

References

[1] Crystallography: Protein Data Bank. Nature New Biology. 1971;233(42):223–223.

[2] Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nature Structural & Molecular Biology. 2003;10(12):980–980.

[3] Li W. A Local Spherical Coordinate System Approach to Protein 3D Structure Description. 2020;.

[4] Reyes VM. Representation of protein 3D structures in spherical (ρ, ϕ, θ)

coordinates and two of its potential applications. Interdisciplinary Sciences: Computational Life Sciences. 2011;3(3).

[5] Batsanov SS. Calculation of van der Waals radii of atoms from bond distances. Journal of Molecular Structure: THEOCHEM. 1999;468(1-2):151–159.

[6] Caine BA, Bronzato M, Popelier PLA. Experiment stands corrected: accurate prediction of the aqueous pKa values of sulfonamide drugs using equilibrium bond lengths. Chemical Science. 2019;10(25):6368–6381.

[7] Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology. 2019;20(11):681–697.

[8] AlQuraishi M. AlphaFold at CASP13. Bioinformatics. 2019;35(22):4862–4865.

[9] AlQuraishi M. A watershed moment for protein structure prediction. Nature. 2020;577(7792):627–628.

[10] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577(7792):706–710.

[11] Hey T, Butler K, Jackson S, Thiyagalingam J. Machine learning and big scientific data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2020;378(2166):20190054.

[12] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). Proteins: Structure, Function, and Bioinformatics. 2019;87(12):1141–1148.

 doi:10.20944/preprints202103.0590.v1

[13] Heo L, Feig M. Modeling of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Proteins by Machine Learning and Physics-Based Refinement. 2020;.

[14] Torrisi M, Pollastri G, Le Q. Deep learning methods in protein structure prediction. Computational and Structural Biotechnology Journal. 2020;18:1301–1310.