https://www.sworldjournal.com/index.php/swj/article/view/swj28-00-016
DOI: 10.30888/2663-5712.2024-28-00-016

UDC 004.89:577.2

ON THE DATA-SET STRUCTURE FOR TRAINING GENERATIVE NEURAL NETWORKS IN PROTEIN DESIGN

Haisha Oleksandr
c.t.s., as.prof.
ORCID: 0000-0003-3711-547X
Institut de Ciències del Mar,
Spain, Barcelona, Pg. Marítim de la Barceloneta, 37, 08003
Haisha Olena
ORCID: 0009-0000-4543-912X
Pylyp Orlyk International Classical University,
Ukraine, Mykolaiv, str. Kotelna, 2, 54003

Abstract. *The paper considers the ways of normalizing a data set that can be used to train neural networks used for designing proteins or other large organic molecules with a complex structure. The paper considers the ways of translational transformation of coordinates of atoms (or amino acid residues), as well as methods for setting universal directions of coordinate axes and transition to them by appropriate linear transformations based on the rotation matrix. The paper specifies the ways of using other information describing the protein and located in the corresponding PDB files. The advantages of using coordinates of whole amino acid residues (instead of taking into account the positions of individual atoms) are substantiated, and the methods for setting their relative coordinates are formalized. The paper can be useful in preparing data on the structure and properties of proteins for their use in the process of training a neural network using these data as input information or for comparing network's output information with them.*

Keywords: *neural network, data-set, data structure, atoms coordinates, coordinate alignment procedure.*

Introduction.

Modern computer technologies and corresponding software are actively used in chemistry to predict the properties of chemical compounds. These approaches are particularly valuable when dealing with complex organic chemical compounds, such as proteins. In other words, thanks to “in silico” calculations, it is becoming increasingly possible to design proteins with specific properties based on information from other natural or artificially designed protein molecules.

One of the most promising areas in the design of new proteins, as well as in predicting their spatial configuration and other related tasks, is the use of artificial intelligence tools. Among such tools, generative neural networks stand out, allowing the generation of new data that are structurally and otherwise similar to existing data that describe real and known objects (in our case, known proteins).

When employing neural networks, which are essentially highly advanced systems for generalizing large amounts of statistical data, a critical question arises regarding the selection of the data to be used for the training stage. Of course, the choice of specific protein samples to include in the training dataset is important. However, an even more crucial aspect is the selection of features that describe each protein; in other words, the primary task is to develop a data schema for the corresponding dataset [1].

Main text.

It is well known that one of the most common ways to represent protein structure information is through Protein Data Bank files, which have the extension PDB. These files, created based on experimental results, contain extensive information about each specific protein. In particular, they store the coordinates of the most likely position of each atom that constitutes the protein. The coordinates are represented in a traditional three-dimensional space and can be denoted by a vector of the form:

$$\vec{x} = \{x_1, x_2, x_3\}, \qquad (1)$$

where x_1 is the first coordinate of the atom (abscissa), x_2 is the ordinate of the atom, and x_3 is its applicate. All coordinates in (1) are defined relatively to an arbitrarily chosen origin of the coordinate system, meaning that the same structure of a particular protein can be described by different sets of atomic coordinates. This implies that for input data, it is more appropriate to use the relative distances between atoms or other structural elements, rather than their absolute coordinate values.

When considering many atoms (and the number of atoms in proteins typically reaches several thousands), it becomes practical to introduce indexing by the atomic number in the protein description. As a result, the coordinates of the i-th atom can be represented by a triplet:

$$\vec{x}_i = \{x_{1i}, x_{2i}, x_{3i}\}, \qquad (2)$$

where the index $i = 1..N_1$, and N_1 is the number of atoms in the group. The complete set of coordinates (2) of a group of atoms can be described by a matrix (or two-dimensional tensor) of the following form:

$$\mathbf{X} = \{\vec{x}_i \mid i = 1..N_1\} \qquad (3)$$

In some cases we can work with internal coordinates, which means that the group of atoms is formed based on some rule. Particularly it can be all the atoms in the total molecule and also any its subset, but described technique can be applied to such group of atoms in any case.

Given that the training process of a neural network considers not just one protein, but a collection of them, it is also practical to introduce indexing by sample (by PDB file). For instance, the complete set of atomic coordinates (3) for the j-th protein from the entire set under consideration can be denoted as follows:

$$\mathbf{X}_j = \{\vec{x}_{ij} \mid i = 1..N_1\}. \qquad (4)$$

Finally, if not just one PDB file is considered, but a complete dataset with M samples, then to describe all the atomic coordinate data of form (4), a three-dimensional tensor can be used of the following form:

$$\boldsymbol{\Xi} = \{\mathbf{X}_j \mid j = 1..M\}. \qquad (5)$$

Taking into account the previous remark regarding the arbitrary choice of the coordinate system origin, it is advisable to introduce adjustments to the mentioned tensor in order to input coordinates into the neural network which are relative to a unified coordinate system. Indeed, when it is necessary to input numerical coordinate values into the neural network for similar proteins, the coordinates of corresponding atoms should be close. However, even the coordinates of the same protein can differ significantly if the axes and origin of the coordinate system are defined differently. Therefore, prior to using coordinates from PDB files (regardless of whether atom-level coordinates are used, as in one method, or entire amino acid residues, or groups of amino acid residues consisting of 8-10 units, as implemented in the Combinatorial Extension method [2]), certain structural alignment procedures must be performed to align the coordinates of different proteins within a common coordinate system.

In the simplest case, the first step can be to select a universal point as the origin for all *M* samples. For example, we can choose the first atom in each PDB file as the reference point. The coordinates of this first atom can be set to {0, 0, 0} instead of the values $\vec{x}_1 = \{x_{11}, x_{21}, x_{31}\}$, which necessitates adjusting all other coordinates of (5) by the amount $\vec{x}_1$. The corresponding relative coordinates will be denoted with a prime:

$$\vec{x}_i' = \vec{x}_i - \vec{x}_1 = \{x_{1i} - x_{11}, x_{2i} - x_{21}, x_{3i} - x_{31}\},\ \mathbf{X}_j' = \{\vec{x}_{ij}' \mid i = 1..N_1\},$$
$$\boldsymbol{\Xi}' = \{\mathbf{X}_j' \mid j = 1..M\} \qquad (6)$$

In the general case, the vector to be subtracted in the given formulas does not necessarily have to correspond to the coordinates of the first atom, but can also represent any meaningful point in the protein molecule's structure (e.g., its center of mass, geometrical center of some important active site, etc.). Identifying of such a point may be a separate, complex problem; however, the procedure for using its coordinates remains the same as described in this study.

Thus, the resulting expression is no longer dependent on the arbitrary choice of the initial reference point, yet it still depends on the arbitrary selection of the coordinate axis directions. To move to a set of universal axes, it is necessary to choose directions along which the unit vectors of the orthonormal basis of the coordinate system can be most conveniently aligned. This can be achieved, for example, through the following procedure. The *Ox* axis can be oriented along the ray connecting the first atom A_1 and the second atom A_2 in the molecule, as illustrated in Fig. 1. The coordinates of the corresponding unit vector in the coordinate system used in the PDB file can then be determined using the formula of the following form:

$$\vec{i} = \frac{\overrightarrow{A_1A_2}}{\left|\overrightarrow{A_1A_2}\right|}, \qquad (7)$$

where $\overrightarrow{A_1A_2} = \{x_{12} - x_{11}, x_{22} - x_{21}, x_{32} - x_{31}\}$ is non-unit vector;

$\left|\overrightarrow{A_1A_2}\right| = \sqrt{(x_{12} - x_{11})^2 + (x_{22} - x_{21})^2 + (x_{32} - x_{31})^2}$ is the magnitude of vector $\overrightarrow{A_1A_2}$

To define the second basis vector, we may consider the plane formed by the first three atoms $A_1A_2A_3$ and choose the second basis vector so that it lies within this plane and is perpendicular to the line A_1A_2 (see Fig. 1). Using this method, to reduce the number of operations, it is advisable firstly to find the third basis vector, which would form a right-handed coordinate system with the first two. For this, the third direction vector can be taken as the result of the cross product of vectors $\overrightarrow{A_1A_2}$ (or vector (7)) and $\overrightarrow{A_1A_3}$, normalizing this result by its magnitude:

$$\vec{k} = \frac{\overrightarrow{A_1A_2} \times \overrightarrow{A_1A_3}}{\left|\overrightarrow{A_1A_2} \times \overrightarrow{A_1A_3}\right|} \qquad (8)$$

The resulting expression (8) can be used to find the second basis vector:

$$\vec{j} = \vec{k} \times \vec{i} \tag{9}$$

The last expression (9) does not require normalization since the vectors being multiplied (formulas (7) and (8)) are unit vectors.

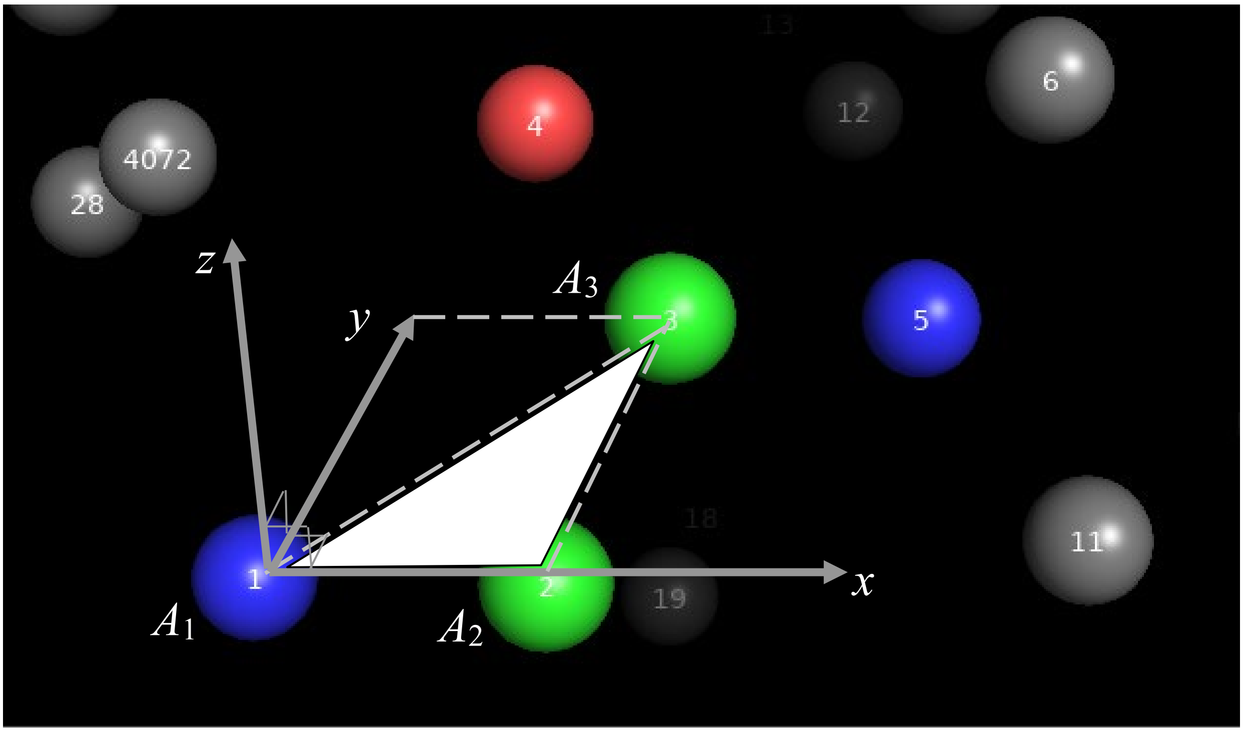

Figure 1 - Visualization of the coordinate system construction based on three specified points (demonstrated using the example of the first three atoms of the protein 7cef.pdb)

The coordinates of the three obtained basis vectors (7-9) should be combined into a rotation matrix **R**, which will be applied to all the coordinates in the given PDB file:

$$\mathbf{R} = \begin{pmatrix} i_x & i_y & i_z \\ j_x & j_y & j_z \\ k_x & k_y & k_z \end{pmatrix} \tag{10}$$

where $\vec{i} = \{i_x, i_y, i_z\}$, $\vec{j} = \{j_x, j_y, j_z\}$ and $\vec{k} = \{i_x, i_y, i_z\}$ are the coordinates of the new basis vectors in the old coordinate system, which is used in the PDB file. The coordinates $\vec{\xi} = \{\xi_1, \xi_2, \xi_3)$ of an arbitrary object from the original PDB file in the new coordinate system $\vec{\xi}' = \{\xi_1', \xi_2', \xi_3')$ will then be obtained through a linear transformation with using (10):

$$\vec{\xi}' = \mathbf{R}\vec{\xi} \tag{11}$$

It is important to clarify that the points A_1, A_2, and A_3 were selected as the first three atoms of the PDB file purely for illustrative purposes, and in general, the described procedure can be applied to move to any arbitrary orientation of the axes. In this case, point A_1 represents the new origin, point A_2 should lie on the new *Ox* axis, and point A_3 is any point in the *xOy* plane. By specifying three such points based on certain considerations (in particular, to match the structure of protein molecules or their individual parts), the coordinates of all objects from the protein file can be obtained following the specified procedure. This final expression (11) will no longer depend on the arbitrary choice of the origin and the direction of the coordinate axes and can be used as input data for a neural network in training.

Regarding the coordinates of individual atoms, if it is necessary to account for their positions with maximum precision, *B*-factors from the PDB file for all the atoms under consideration can also be introduced. These B-factors characterize the amplitude of thermal motion for each atom. From these values, a matrix of the following form can be constructed:

$$\mathbf{B} = \{b_{ji} \mid i = 1..N_1, j = 1..M\}, \tag{12}$$

where b_{ji} is an element of the matrix located in the *j*-th row at the *i*-th position, corresponding to the *B*-factor of the *i*-th atom belonging to the *j*-th protein.

This matrix **B** (12) can also be provided as input to a neural network that learns the rules and structural features of proteins.

Furthermore, despite the discussion above regarding the consideration of each atom's coordinates, it should be noted that such highly detailed information about the protein structure, as **Ξ**' (which accounts for the position of each individual atom), is, in most cases, excessive. This is especially true in the case of constructing a protein with a common functional group or active site, such as a catalytic triad. For example, in the well-studied and extensively documented mechanism of the catalytic triad [3], the key role is not played by individual atoms but rather by entire amino acid residues. In many hydrolases, the most well-known (or classic) catalytic triad consists of "Serine-Histidine-Aspartic Acid" (as, for example, in TfCut2 protein, studied by the authors in [4]). These three amino acid residues exhibit varying degrees of flexibility within the protein, particularly during interaction with the substrate. For instance, the serine residue has a flexible side chain capable of rotation around σ-bonds. In aspartic acid, the carboxyl group restricts rotation, imparting a certain rigidity to the side chain. The imidazole ring of histidine adds rigidity to its side chain. Overall, it is evident that it makes sense to consider the relative positioning and orientation of the amino acid residues within the active site. Extrapolating this approach to all amino acid residues, it is more practical to consider not their individual atoms, but certain centers, such as centers of mass or geometric centers. This approach is also justified by the fact that the average number of atoms in a standard amino acid is around 20, which reduces the volume of information input to the neural network by the same factor and also decreases the number of necessary mathematical operations for processing this information. At the same time, obtaining additional insights by considering the positions of individual atoms rather than entire amino acid residues appears unlikely. Thus, we proceed to the consideration of working with the centers of individual residues.

The position of the center of mass of the *k*-th amino acid residue can be calculated using the formula:

$$\vec{r}_k = \frac{\sum\limits_{i:\, A_i \in R_k} m_i \vec{x}_i}{\sum\limits_{i:\, A_i \in R_k} m_i} \tag{13}$$

where A_i is the i-th atom in the protein sequence listed in the PDB file;

R_k is the k-th amino acid residue for which the center of mass is being calculated, where $k = 1..N_2$, and N_2 is the number of amino acid residues forming the protein;

m_i is the mass of the i-th atom.

In certain cases, it may be more appropriate to consider the geometric center instead of the center of mass, which can be defined as the center of mass assuming that all atoms have equal mass. In this case, the geometric center will be determined by the following formula, which we get from (13) by reducing it by mass:

$$\vec{r}_k = \frac{\sum_{i:\, A_i \in R_k} x_i}{\sum_{i:\, A_i \in R_k} 1} \tag{14}$$

In addition to the positions of the centers of amino acid residues, their spatial orientation may also play an important role, which is particularly relevant for rigid static atomic structures that are not subject to rotations (and to a lesser extent for flexible ones that contain σ-bonds). As is well known, the spatial orientation of a rigid three-dimensional structure can be described by three Euler angles (α, β, γ). When considering pairs of amino acid residues, the relative angles, which show the mutual orientation of two rigid bodies in space, are of interest and are equal to the differences in the Euler angles calculated for each of the pair of amino acid residues.

This information related to the spatial geometry of complex organic molecules, such as proteins, can be fed into a neural network for training or used in the loss function, comparing it with the network's outputs.

Conclusion.

Thus, the study proposes approaches to formalizing information related to the spatial structure and properties of complex organic molecules. In particular, methods for transforming data from PDB files describing such complex structures as proteins have been considered.

References:

1. Hou, Q., Waury, K., Gogishvili, D., Feenstra, K.A. Ten quick tips for sequence-based prediction of protein properties using machine learning // PLoS Computational Biology. — 2022. — T. 18, №12. — e1010669. DOI: https://doi.org/10.1371/journal.pcbi.1010669.

2. Shindyalov, I.N., Bourne, P.E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path // Protein Engineering, Design and Selection. — 1998. — T. 11, №9. — C. 739–747. DOI: https://doi.org/10.1093/protein/11.9.739.

3. Blay, V., Pei, D. Serine proteases: how did chemists tease out their catalytic mechanism? // ChemTexts. — 2019. — T. 5. — C. 19. DOI: https://doi.org/10.1007/s40828-019-0093-4.

4. Falkenstein, P., Zhao, Z., Di Pede-Mattatelli, A., Künze, G., Sommer, M., Sonnendecker, C., Zimmermann, W., Colizzi, F., Matysik, J., Song, C. // ACS Catalysis. — 2023. — T. 13, №10. — C. 6919-6933. DOI: https://doi.org/10.1021/acscatal.3c00259.

Анотація. *У роботі розглядаються шляхи нормалізації набору даних, який може бути використаний для навчання нейронних мереж, що застосовуються для дизайну білків або інших великих органічних молекул, що володіють складною структурою. Розглянуто шляхи трансляційного перетворення координат атомів (або амінокислотних залишків), а також способи завдання універсальних напрямків осей координат та переходу до них шляхом відповідних лінійних перетворень на основі матриці повороту. Вказано шляхи використання інших відомостей, що описують білок, та розміщені у відповідних PDB-файлах. Обґрунтовано переваги використання координат цілих амінокислотних залишків (замість урахування положень окремих атомів) та формалізовано способи завдання їх відносних координат. Робота може бути корисною при підготовці даних про структуру та властивості білків для їх використання в процесі навчання нейронної мережі, що використовує ці дані як вхідну інформацію або для порівняння з ними її вихідної інформації.*

Ключові слова: *нейронна мережа, дата-сет, структура даних, координати атомів, вирівнювання координат.*

Article sent: 05.11.2024