RMSD CALCULATIONS AND COMPUTER MODELLING OF PROTEIN STRUCTURES

Fatima Sapundzhi, Valentin Slavov

South-West University "Neofit Rilski", 66 Ivan Michailov str. 2700, Blagoevgrad, Bulgaria,
E-mail: sapundzhi@swu.bg

Received 11 January 2019
Accepted 30 July 2019

ABSTRACT

One of the important topics in structural bioinformatics refers to the analysis of protein sequences and their biological functions, as well as the assessment of protein structural similarities. These investigations play a critical role in the drug design, the homology modelling and the protein structure prediction.

Therefore, it is important to evaluate the structures similarity and to identify the similar predictions. The degree of similarity of two protein 3D structures is usually measured by the root-mean-square distance (RMSD) between the equivalent atom pairs. The objective of this research is to present a simple procedure to calculate the RMSD between pairs of 3D structures and to align the structures in order to find the minimal value of RMSD. A web service for calculating the RMSD in Perl programming language is developed. The tool can be used in the field of bioinformatics research and computer modelling of protein structures.

Keywords: computer modelling, root-mean-square distance, RMSD, protein sequences, ligand-receptor interactions, Perl.

INTRODUCTION

Measuring the difference between 3D structures is useful in bioinformatics, where comparing different molecule conformations is needed. The researchers usually measure the difference between two conformations of a molecule by computing the Root Mean Square Deviance (RMSD). It is a measure of the similarity of the three-dimensional (3D) structures often used in bioinformatics, which computes the sum of the squared distances between the corresponding atoms [1, 2].

It is important to obtain the minimum values of RMSD in these calculations. Aiming this the conformations studied must be initially aligned, including their translation and rotation. The molecule conformations, i.e. the coordinates of its atoms are its important characteristic feature. A specific molecule may have different conformations, which in turn requires to examine the differences between them.

A widely used approach to compare the 3D structures of biomolecules is to align - translate and rotate a structure with respect to another one to minimize the value of RMSD.

There are several procedures aiming to find the optimal solution alignment. Some of them refer to the algorithms presented by Kabsh [3,4] and by McLachlan [5].

The purpose of the presented work is to create a web service and a web interface related to this surface, which calculates the RMSD and returns the result.

EXPERIMENTAL

An immensely popular measure used to express the structural similarity of 3D structures refers to the RMSD calculated between equivalent atoms. The RMSD can be defined for two structures containing identical number and types of atoms. Let the two sets atomic coordinates be V_i and W_i, ($i = 1,2, \ldots, n$). The RMSD formula is presented by Eq. 1:

$$RMSD\,(V,W) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|V_i - W_i|^2} \qquad (1)$$

where

$$|V_i - W_i|^2 = (v_{i,x} - w_{i,x})^2 + (v_{i,y} - w_{i,y})^2 + (v_{i,z} - w_{i,z})^2$$

is the square distance between each of the n pairs of

equivalent atoms in two optimal superposed structures.

The sets atomic coordinates V_i and W_i, can refer to a subset of the entire molecule, such as Cα atoms, backbone atoms or heavy atoms. In order to reflect the internal motions of the proteins, it is necessary to have a similar orientation in the space. This is most often achieved by fitting to a given model.

The superposition of the V_i and W_i sets of the identical atomic positions can be mathematically treated as a minimization of RMSD between the atoms (weighted RMSD).

A rotation matrix R minimizing RMSD is required for the V_i and W_i sets of cantered atomic positions:

$$minRMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|V_i - R.W_i|^2} \quad (3)$$

Kabsch algorithm

The Kabsh algorithm is one of the earliest algorithms applied to achieve a mathematically exact solution of minimum RMSD [2, 3]. For the given sets of V_i and W_i W_i of atomic positions, if the points are not centred, it is important to translate them, prior to the operation, in a way that their average coincides with the origin.

The rotation matrix $R_{3\times3}$ is required to rotate the points of V_i into W_i. The matrixes $P_{n\times3}$ and $Q_{n\times3}$ are obtained. They contain the coordinates of V_i and W_i as row vectors, respectively:

$$P = \begin{pmatrix} x_1 & y_1 & z_1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & z_n \end{pmatrix} \quad (3)$$

The cross-covariance matrix A is calculated in accordance with:

$$A = P^T Q \quad (4)$$

It is possible to calculate the optimal rotation matric R, which is given as:

$$R = \sqrt{(A^T A)}.A^{-1} \quad (5)$$

It is not guaranteed that the matrix A has an inverse. A singular value decomposition (SVD) of matrix A is carried out to account for all special cases:

$$A = TSU^{-1} \quad (6)$$

Then the rotation matrix R is given as:

$$R = T\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix}U^{-1} \quad (7)$$

where $d = sign(det(TU^{-1}))$.

Perl (Practical Extraction and Report Language)

Perl is a popular programming language used in bioinformatic environments [6, 7]. It is an interpreted and dynamic programming language. It is important for bioinformatics to have a scripting language to quickly develop scripts (short programs) for scanning or transforming large amounts of data. Perl is a good language for scripting, because of its compact syntax, a broad array of functions, and data orientation. This language provides powerful ways to match and manipulate strings through the use of regular expressions.

BioPerl [7] is a part of Perl modules that facilitates the development of Perl scripts for bioinformatics applications. It provides a set of sequence analysis functions accessing the biological databases for data retrieval, reading and converting the major sequence file formats, as well as providing an interface to ClustalW, BLAST, FastA, and other standard bioinformatics applications representing the protein structure (PDB) data, etc.[8].

RESULTS AND DISCUSSION

The current research presents a program that can calculate RMSD. The toll calculates RMSD from two sets of vectors V and W. The Kabsch algorithm [3, 4] is applied. It follows three steps:

- the first one translates both structures at the center of the coordinate system;
- the second one computes the cross-covariance matrix;
- the third one computes the optimal rotation matrix.

The optimal rotation matrix R is used to rotate matrix P unto matrix Q so the minimum RMSD can be calculated.

A calculator with a command interface (CLI) is realized to achieve the goal. The interpreted programming language Perl [6, 7] is chosen as the implementation language.

The script calculates RSMD in Angstroms between the two structures. The program accepts two files as arguments. They are described in a specific format of the atomic structure - .pdb or .xyz.

The RMSD measures the average distance between the atoms of two protein or ligand structures according to Eq.1. The subscripts x, y, z denote the $x - y - z$ coordinates of every atom.

Note that both files (both structures) must contain the same number of atoms in a similar order to provide the work of the script.

The server side of the project is implemented through the NodeJS-based Express Web server, while the client part is realized through the functional language Elm. A Command Interface Calculator (CLI) is used to implement the project written on Perl.

The server side of the application is a quite simple web service that accepts the files, executes the commands, and returns the response to the client with the calculated values or error (Fig. 1).

The program displays the RMSD calculated in three ways (Fig. 2).

The result of the Perl tool using the Kabsch algorithm is shown for a comparison. The known .pdb files are used for testing the Kabsch algorithm (ci2_1.pdb and ci2_2.pdb). The response that is sent to the client is JSON (JavaScript Object Notation) format [9]. It is a standard text-based format for representing structured data based on JavaScript object syntax. It is commonly used for transmitting data in web applications.

The client component is a simple application written with the help of language Elm - a purely functional language based on Haskel [10]. The architecture of an Elm project sets a structure that divides the logic of the application in three distinct parts:

A *Model* – it presents the status of the application.

Fig. 1. RMSD Calculator.

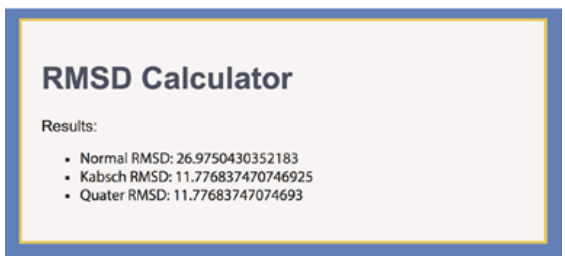

Fig. 2. Results provided by the RMSD Calculator.

In this case, the essential part of the model refers to the files as well as the server and several help variables that work with the drag & drop control on files.

- An *Update* – it shows how the status will be changed describing the logic that will update the model and the model presentation in HTML format. The update happens via the update function, which receives a message and returns a model and command. Messages can be obtained from the user interface or at execution of a command in the application.

- A *View* – it presents the status as HTML.

The results of the program implementation on the ground of a given example are shown as follows:

```
Normal RMSD: 26.9750429930231
Translated RMSD: 15.8457291113576
Rotated RMSD: 11.776837467101
```

The result obtained by the *Python* tool [11] using the Kabsch algorithm is given for a comparison as:

```
$calculate_rmsd ci2_1.pdb ci2_2.pdb
Normal RMSD: 26.9750430352183
Kabsch RMSD: 11.776837470746925
```

The result using the Perl module Bio::PDB:Structure::Atom [12] is:

RMSD-biopdb: 11.7895493912548.

This program has been used in previous studies with opioid receptors [13-18]. It has been developed on the Perl scripting language using the BioPerl bioinformatics package that facilitates the development of molecular biological software.

The developed tool is designed to be easy to imbed in other applications [19-25]. This tool will be uploaded to a server and can be freely used by researchers in the field of bioinformatics research and computer modelling of protein structures.

CONCLUSIONS

The bioinformatics considers the question of a structural comparison of two molecules whose atoms spatial positions are known. This is done by algorithms described in this investigation which center and rotate the molecules investigated providing metrics for the differences (RMSD) between them.

Acknowledgements

This paper is partially supported through Projects RPY – B4/19; RP – B7/20 by SWU "Neofit Rilski", Bulgaria, Project of the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)" and Project BNSF H27/36 by the National Science Fund at the Ministry of Education and Science, Bulgaria.

REFERENCES

1. E. Coutsias, C. Seok, K. Dill, Using quaternions to calculate RMSD, J Comput Chem. 25, 15, 2004, 1849-1857.
2. F. Armougom, S. Moretti, V. Keduas, C. Notredame, The iRMSD: a local measure of sequence alignment accuracy using structural information, Bioinformatics, 22, 14, 2006, e35-39.
3. W. Kabsch, A solution for the best rotation to relate two sets of vectors, Acta Crystallographica Section A, 32, 5, 1976, 922-923.
4. W. Kabsch, A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A, 34, 5, 1978, 827-828.
5. A. McLachlan, Gene duplications in the structural evolution of chymotrypsin Journal of Molecular Biology, 128, 1, 1979, 49-79.
6. L. Wall, T. Christiansen, J. Orwant Programming Perl, O'Reilly Media; 3rd edition, 2000.
7. J.Tisdall, MasteringPerl for Bioinformatics: Perl Programming for Bioinformatics, O'Reilly Media; 1 ed., 2003.
8. J. Daugelaite, A. O' Driscoll, R. Sleator, An overview of multiple sequence alignments and cloud computing in bioinformatics, ISRN Biomathematics, 2013, 2013, 1-14.
9. T. Marrs, JSON at Work: Practical Data Integration for the Web, O'Reilly Media, 1 edition, 2017.
10. J. Fairbank, Programming Elm: build safe sane and maintainable front end applications, Pragmatic Bookshelf, 2017.
11. P. Joshi, Artificial Intelligence with Python: A Comprehensive Guide to Building Intelligent Apps for Python Beginners and Developers, Packt, 2017.
12. S. Kearsley, On the orthogonal transformation used for structural comparisons. Acta Cryst. A45, 1989, 208-210.
13. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, Determination of the relationship between the docking studies and the biological activity of δ-selective enkephalin analogues, Journal of Computational Methods in Molecular Design, 5, 2015, 98-108.
14. T. Dzimbova, F. Sapundzhi, N. Pencheva, P. Milanov, Computer modeling of human delta opioid receptor, Int. J. Bioautomation, 17, 2013, 5-16.
15. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, Comparative evaluation of four scoring functions with three models of delta opioid receptor using molecular docking, Der Pharma Chemica, 8, 2016, 118-124.
16. I. Nedyalkov, A. Stefanov, G. Georgiev, Modelling and studying of cloud infrastructures, International Conference on High Technology for Sustainable Development, HiTech 2018 – Proceedings, 2018, 8566664.
17. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, Modeling the relationship between biological activity of delta-selective enkephalin analogues and docking results by polynomials, Bulgarian Chemical Communications, 49, 4, 2017, 768-774.
18. P. Apostolov, A. Stefanov, S. Andonova, Application of Hausdorff window for array antennas design, 27th National Conference with International Participation TELECOM, Sofia, Bulgaria, 2019, 4-7.
19. F. Sapundzhi, M. Popstoilov, C# implementation of the maximum flow problem, 27th National Conference with International Participation TELECOM, Sofia, Bulgaria, 2019, 62-65.
20. F. Sapundzhi, T. Dzimbova, Computer modelling of the CB1 receptor by molecular operating environment, Bulgarian Chemical Communications, 50, Special Issue B, 2018, 15-19.
21. F. Sapundzhi, T. Dzimbova, N. Pencheva, P. Milanov, Molecular docking experiments of cannabinoid receptor, Bulgarian Chemical Communications, 50, Special Issue B, 2018, 44-48.
22. F. Sapundzhi, M. Popstoilov, Optimization algorithms for finding the shortest paths, Bulgarian Chemical Communications, 50, Special Issue B, 2018, 115-120.
23. R. Topalska, F. Sapundzhi, Chemical structure computer modelling, Journal of Chemical Technology and Metallurgy, 55, 4, 2020, 715-719.
24. F. Sapundzhi, Computer modelling and optimization of the structure-activity relationship by using surface fitting methods, Bulgarian Chemical Communications, 51, 4, 2019, 569-579.
25. F. Sapundzhi, T. Dzimbova, A study of QSAR based on polynomial modeling in Matlab, International Journal of Online and Biomedical Engineering, 15, 15, 2019, 39-56.