# NOVEL COMPUTATIONAL TOOLS AND DATABASES

**DOI: https://doi.org/10.18454/jbg.2019.3.12.1**

**Rudnev V.R.*[1], Tikhonov D.A.[2], Kulikova L.I.[3], Gubin M.Yu.[4], Efimov A.V.[5]**

[1, 2] Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Pushchino, Russia
[2, 3] Institute of Mathematical Problems of Biology RAS - the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Russia
[4] National Research Tomsk Polytechnic University, Tomsk, Russia
[5] Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia

* Correspodning author (volodyarv[at]mail.ru)

Received: 12.11.2019; Accepted: 25.11.2019; Published: 29.11.2019

# DATABASE OF TWO-HELICAL MOTIFS OF PROTEIN MOLECULES AND COMPUTER SERVICES FOR THEIR ANALYSIS

Application notes

**Abstract**

This paper is devoted to the development of structural motifs database of protein molecules consisting of two elements of a secondary structure that have unique spatial stacking of a polypeptide chain. The motives investigated are pairs of any type of helix, connected by a different non-zero length and different conformation of the connection. For each structure, geometric parameters are calculated. The database interface allows you to perform searches, sorting, filtering on all parameters, which makes it possible to obtain samples of structures with given geometric characteristics. Also, the system interface allows for statistical analysis and the construction of histograms for the distribution of various characteristics of structures in the sample. Thus, it is possible to investigate the correlation of the geometry of the spatial structures of the investigated motifs with the amino acid sequences. The system allows uploading of the received samples to the file system in CSV format. The upload format and set of uploaded data is configured in the interface. Downloading can contain both all the geometric characteristics of the structures of interest, and the corresponding amino acid sequences. The database interface also allows you to view 3D models of structural motifs.

**Keywords:** protein structural motifs, database, helix pairs, inter-helix distances, helix axis angles, area of helix projection intersection.

**Руднев В.Р.*[1], Тихонов Д.А.[2], Куликова Л.И.[3], Губин М.Ю.[4], Ефимов А.В.[5]**

[1, 2] Институт теоретической и экспериментальной биофизики РАН, Пущино, Россия
[2, 3] Институт математических проблем биологии РАН – филиал Института прикладной математики им. М.В. Келдыша РАН, Пущино, Россия
[4] Национальный исследовательский Томский политехнический университет, Томск, Россия
[5] Институт белка РАН, Пущино, Россия

* Корреспондирующий автора (volodyarv[at]mail.ru)

Получена: 12.11.2019; Доработана: 25.11.2019; Опубликована: 29.11.2019

# БАЗА ДАННЫХ ДВУХСПИРАЛЬНЫХ МОТИВОВ БЕЛКОВЫХ МОЛЕКУЛ И ВЫЧИСЛИТЕЛЬНЫЕ СЕРВИСЫ ДЛЯ ИХ АНАЛИЗА

Техническая спецификация

**Аннотация**

Данная работа посвящена созданию базы данных структурных мотивов белковых молекул, состоящих из двух элементов вторичной структуры, имеющих уникальные укладки полипептидной цепи в пространстве. Исследуемые мотивы представляют собой пары любого типа спиралей, соединёнными между собой различной ненулевой длины и различной конформации перетяжки. Для каждой структуры рассчитаны геометрические параметры. Интерфейс базы данных позволяет выполнять операции поиска, сортировки, фильтрации по всем параметрам, что дает возможность получать выборки структур с заданными геометрическими характеристиками. Также интерфейс системы позволяет проводить статистический анализ и строить гистограммы распределения различных характеристик структур в выборке. Таким образом, имеется возможность исследовать корреляцию геометрии пространственных структур исследуемых мотивов с аминокислотной последовательностью. Система позволяет выгружать полученные выборки в файловую систему в формате CSV. Формат выгрузки и набор выгружаемых данных настраивается в интерфейсе.

Выгрузка может содержать как все геометрические характеристики интересующих структур, так и соответствующие аминокислотные последовательности. Интерфейс базы данных позволяет также просматривать 3D модели интересующих структурных мотивов.

**Ключевые слова:** структурные мотивы белковых молекул, база данных, спиральные пары, межспиральные расстояния, углы между осями спиралей, площадь пересечения проекций спиралей.

## 1. Introduction

The importance of creating a database of structural motifs that have unique spatial styling of the polypeptide chain and their further study follows from the interest of researchers in these structures [1], [2], [3], [4], [5], [6], [7]. Structural motifs formed by two α-helices located in the polypeptide chain one after another and interconnected by constrictions are described in [2], [3]. They are compact spatial structures. It is also known from the literature that the densest packing of two α-helices is achieved with antiparallel, perpendicular and so-called beveled orientation between the helices. Examples of such packages are super-secondary structures: α-α-corners, α-α-hairpins, L-shaped and V-shaped structures [3]. Thus, the creation of a database of all structural motifs of protein molecules registered in the bank of PDB protein structures [8], which allows a comprehensive analysis of structures, is an extremely important and urgent task [9], [10], [11], [12], [13]. This work is aimed at the creation of a database of helix pairs formed by two helices of any type located in a polypeptide chain one after another and interconnected by constrictions of different lengths having different conformations. In published works [14], [15], [16], the approach we developed for the selection of helical pairs in the structures of protein molecules presented in PDB was described. In these studies, structures were studied in the formation of which two helices of any type participate: α-helices (type H helix), 310-helices (G-helix) and π-helices (I-helix). A point model of a helix pair was proposed, which describes structures with four points in space. Indeed, if we approximate both helices by cylinders onto which the helices formed by a thread passing through Cα atoms are wound, then the beginnings and ends of the axes of the cylinders will give us those four points that completely describe this super-secondary structure.

Important characteristics of helix pairs are inter-helix distances (interplanar distance, minimum distance and distance that describes the relative location of the helices in the helix pair), angles between the axes of the helices and the intersection area of the projections of the helices.

Other main characteristics of the studied structures and their descriptions are presented below:

• PDB code - a unique code of a protein molecule in Protein Data Bank in which the structure is detected;

• beginning of structure - the number of amino acid residue corresponding to the beginning of the first helix in the protein molecule;

• beginning of the connection- the number of amino acid residue corresponding to the beginning of the banner;

• end of structure - the number of amino acid residue corresponding to the end of the structure;

• end of structure — number of amino acid residue corresponding to the end of the second helix;

• angle between the axes of the helices is the interhelix angle φ;

• distance between planes - interplanar distance d;

• projection intersection area — the area of the intersection of the projections of the cylinders;

• secondary structure of DSSP - the result of processing the structure with the DSSP program [17];

• Helix A primary structure — amino acid sequence corresponding to the first helix of the structure;

• the primary structure of the constriction is the amino acid sequence corresponding to the constriction;

• helix B primary structure — amino acid sequence corresponding to the second helix;

• connection length - the amount of amino acid residues in the hauling;

• helix A structure length - the number of amino acid residues in the first helix;

• helix B structure length - the number of amino acid residues in the second helix.

## 2. Methods

### 2.1. Server side

The server part of the portal is written in the Python programming language. This language was chosen for several reasons:

1. Python is one of the most commonly used languages for developing applications related to Data science, and has a wide range of tools for processing data;

2. Python has developed tools for working with biological data, for example BioPython;

3. the existence of ready-made modules for importing Matlab files into Python.

The simplicity and convenience of developing Internet applications in Python, especially in conjunction with existing frameworks, also played an important role. This project used the Django framework as the basis for the site.

An ORM system (Object-relational mapping, converting relational DBMS data into objects of a programming language) can significantly simplify and speed up the development of interaction with the database, as well as abstract from direct work with the database as such, which makes it possible, if necessary, to replace it with minimal effort DB.

To store data at this stage, relational DBMS (Database Management System) MySQL is used. The choice of the DBMS was determined, first of all, by the ease of installation and integration with the application. In the future, in order to support broader and more flexible capabilities, it is planned to switch to non-relational (relational DBMSs work with data in the form of tables and relationships between them; non-relational DBMS include any other DBMS) ElasticSearch DBMS, which has the following capabilities and advantages:

- high performance when working with large volumes of data;

- convenient and functional interface for integration with software;
- the ability to store and index arbitrary data structures, including nested ones;
- wide possibilities for data search.

### 2.2. Client side

The client part is divided into three parts: tabular data view interface; interface for viewing distributions; interface visualization of molecules and functional elements.

The tabular data viewer interface is written in JavaScript + HTML + CSS to adhere to the basic standards applicable to technologies for developing sites on the Internet and there is no need for complex client-side logic that could require the use of CSS preprocessors, visual HTML editors or JavaScript frameworks.

The distribution histogram viewing interface was created using the Google Charts library, which provides the necessary functionality with ease of integration and a fairly high speed. The visualization interface uses the JSmol display system, which allows viewing the visualization of molecules and fragments of molecules in a web browser. This solution is non-alternative for such functionality. In this case, the displayed fragments of molecules are placed in the file system as files in the PDB format. JSmol also uses its own server-side script in php, which in the future is planned to be replaced with its own backend, which allows you to use data from the database, and not directly as files.

## 3. Results

The work presents a database of structural motifs of protein molecules, consisting of two elements of the secondary structure, having unique stackings of the polypeptide chain in space. The studied motifs are pairs of any type of helices, interconnected by constrictions of different nonzero lengths and various conformations. For each structure, geometric parameters are calculated. The database of double-helix motifs of protein molecules is available on the Internet URL:http://protdb.org. The developed database allows you to build samples of structural motives according to geometrical parameters of interest, and also provides the opportunity to study the relationship of the spatial structure geometry with the amino acid sequence using the developed tools.

The database interface allows you to search, sort, filter by all parameters, which makes it possible to obtain samples of structures with specified geometric characteristics. Also, the system interface allows for statistical analysis and histograms of the distribution of various characteristics of structures in the sample. The system allows you to upload the resulting samples to a file system in CSV format. Unloading may contain both all geometric characteristics of the structures of interest, and the corresponding amino acid sequences. The database interface also allows you to view 3D models of structural motives. The implemented basic functionality for viewing structural elements with a limited set of analytical tools implies further expansion of capabilities.

It is also planned to implement the import functionality of structures created using the Matlab application.

## Conflict of Interest

None declared.

## Конфликт интересов

Не указан.

## Funding

The study was made with the support from the RFBR (project № 18-07-01031).

## Финансирование

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-07-01031.

## References

1. Edwards M.S. Structural and sequence patterns in the loops of βαβ units. Protein Engineering / Edwards M.S., Sternberg J.E., Thornton J.M. – 1987. –V. 1. – P. 173–181.

2. Efimov A.V. Standard structures in proteins / Efimov A.V. // Prog. Biophys. Molec. Biol. –1993. –V. 60. – P. 201–239.

3. Efimov A.V. Super-secondary structures and modeling of protein folds / Efimov A.V. // In: Methods in Molecular Biology. Ed. Kister A.E. Clifton: Humana Press, –2013. – V. 932. – P. 177–189.

4. Lin S.I. A study of 4-helix bundles–investigating protein folding via similar architectural motifs in protein cores and in subunit interfaces / Lin S.I., Tsai J., Nussinov R. //J. Mol.Biol. –1995. –V. 248. –P. 151–161.

5. Chothia C. Structure of proteins: packing of α-helices and pleated sheets / Chothia C., Levitt M., Richardson D. // Proc. Natl. Acad. Sci. –1977. –V. 74. – P. 4130–4134.

6. Trovato A. A new perspective on analysis of helix-helix packing preferences in globular proteins / Trovato A., Seno F. // Proteins: structure, function, bioinformatics. –2004. –V. 55.– P. 1014–1022.

7. Walther D. Principles of helix-helix packing in proteins: the helical lattice superposition model / Walther D., Eisenhaber F., Argos P. // Molecular Biology. –1996. –V. 255. –P. 536–553.

8. Berman H.M. The Protein Data Bank / Berman H.M., Westbrook J., Feng Z. and others // Nucleic Acids Research. – 2000. – V. 28. – P. 235–242.

9. Adamian L.I. Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins / Adamian L.I., Liang J. // J. Mol.Biol. –2001. –V. 311. – P. 891–907.

10. Calhoun J.R. Computational design and characterization of a monomeric helical dinuclear metalloprotein / Calhoun J.R., Kono H., Lahr S. and others // Journal of Molecular Biology. –2003. –V. 334. –No. 5. –P. 1101–1115.

11. Calhoun J.R. Artificial diiron proteins: From structure to function / Calhoun J.R., Nastri F., Maglio O. and others // Peptide Science. –2005. –V. 80. –No. 2–3. –P. 264–278.

12. Chino M. Artificial diiron enzymes with a de novo designed four-helix bundle structure / Chino M., Maglio O., Nastri F. and others // European Journal of Inorganic Chemistry. –2015. –P. 3371–3390.

13. Chino M. Designing Covalently Linked Heterodimeric Four-Helix Bundles / Chino M., Leone L., Maglio O. and others // Methods in enzymology. –2016. –2016(21). –V. 580. –P. 471–499.

14. Tikhonov D.A. Statistical Analysis of the Internal Distances of Helical Pairs in Protein Molecules / Tikhonov D.A., Kulikova L.I., Efimov A.V. // J. Mathematical Biology and Bioinformatics. –2019. –V. 14(S). – P. t18–t36.

15. Tikhonov D.A. The Study of Interhelical Angles in the Structural Motifs Formed By Two Helices / Tikhonov D.A., Kulikova L.I., Efimov A.V. // J. Mathematical Biology and Bioinformatics. –2019. –V. 14(S). –P. t1–t17.

16. Tikhonov D.A. Analysis Of Torsion Angles Between Helical Axes in Pairs of Helices in Protein Molecules / Tikhonov D.A., Kulikova L.I., Efimov A.V. // J. Mathematical Biology and Bioinformatics. –2018. –V. 13(S). –P. t17–t28.

17. Kabsch W. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features / Kabsch W., Sander C. // Biopolymers. –1983. –V. 22. –№ 12. –P. 2577–2637.