

Прогноз свойств биологических последовательностей на основе их структурных формул – новый подход к представлению данных в биоинформатике

Лагунин А.А.^{1, 2*}, Смирнов А.С.¹, Задорожный А.Д.¹, Лебедев Н.В.¹,
Захаров О.С.¹, Полусмак И.В.¹, Журавлева С.И.¹, Степанян А.А.¹, Алимова А.Р.³,
Рудик А.В.², Филимонов Д.А.²

¹ Российский национальный исследовательский медицинский университет им. Н.И. Пирогова,
Москва, Россия

² НИИ биомедицинской химии им. В.Н. Ореховича, Москва, Россия

³ Московский государственный университет им. М.В. Ломоносова, Москва, Россия

* alexey.lagunin@ibmc.msk.ru

Ключевые слова: биологические последовательности; структурные формулы; прогноз; связь структура-свойство

Мотивация и цель: Анализ структурных и функциональных свойств молекул белков, РНК и ДНК является одной из важнейших задач в молекулярной биологии. Несмотря на активное исследование в этой области, огромное количество белков, РНК и участков геномов разных организмов нуждаются в оценках этих свойств для понимания их роли в различных биологических процессах. Экспериментальные методы ограничены в своих возможностях из-за сложности, стоимости и продолжительности таких исследований. Начиная с середины XX в. для анализа структурных и функциональных свойств биологических макромолекул используют аминокислотные (а.к.) и нуклеотидные (н.т.) последовательности, описанные в виде однобуквенного кода. И это является основой большинства современных методов биоинформатики, с помощью которых можно эффективно оценивать свойства биологических макромолекул. Несмотря на значительные успехи в создании таких методов, само представление а.к. и н.т. последовательностей в виде однобуквенного кода является ограничением для дальнейшего совершенствования биоинформатических методов, так как не отражает особенностей структуры этих последовательностей на уровне связей между атомами. С другой стороны, развитие экспериментальных методов определения 3D структур биологических макромолекул позволяет описывать их атомы в трехмерном пространстве. Такое описание позволяет оценивать их характеристики и свойства намного качественней, но требует применения значительных вычислительных ресурсов, а также ограничено количеством известных трехмерных структур.

Методы и алгоритмы: Наши исследования предлагают альтернативу представления а.к. и н.т. последовательностей белков в виде структурных формул их фрагментов, что позволяет использовать разработанные ранее в хемоинформатике методы выявления связей «структура-свойство». В качестве описания структурных формул фрагментов макромолекул используются подструктурные дескрипторы многоуровневых атомных окрестностей MNA (Multilevel Neighborhoods of Atoms) высокого уровня, реализованные в специально

разработанной для работы с такими данными программе MultiPASS, использующей байесовский алгоритм для выявления связи «структура-свойство».

Результаты: Эффективность такого подхода продемонстрирована нами в методах по прогнозу: патогенных аминокислотных замен (<https://www.way2drug.com/SAV-Pred/>) [1, 2], аминокислотных замен, связанных с лекарственной резистентностью опухолей [3], специфичности Т-клеточных рецепторов к эпитопам (<https://www.way2drug.com/TCR-Pred/>) [4] и посттрансляционных модификаций [5]. В докладе будет также продемонстрировано использование нашего подхода для предсказания вторичных структур и функциональных мотивов белков, а также мишеней микроРНК и сайтов связывания транскрипционных факторов человека.

Выводы: Предложенный подход представляет новое направление биоинформатики в исследовании свойств биологических макромолекул на основе структурных формул их фрагментов.

Финансирование: Работа выполнена в рамках Программы фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021–2030 гг.) (№ 122030100170-5).

Prediction of properties of biological sequences based on their structural formulas – a new approach to data representation in bioinformatics

Lagunin A.A.^{1,2*}, Smirnov A.S.¹, Zadorozhny A.D.¹, Lebedev N.V.¹, Zakharov O.S.¹, Polusmak I.V.¹, Zhuravleva S.I.¹, Stepanyan A.A.¹, Alimova A.R.³, Rudik A.V.², Filimonov D.A.²

¹ Pirogov Russian National Research Medical University (RNRMU), Moscow, Russia

² Institute of Biomedical Chemistry (IBMC), Moscow, Russia

³ Lomonosov Moscow State University, Moscow, Russia

* alexey.lagunin@ibmc.msk.ru

Key words: biological sequences; structural formulas; prediction; structure-property relationships

Motivation and Aim: Analysis of the structural and functional properties of protein, RNA and DNA molecules is one of the most important tasks in molecular biology. Despite active research in this area, a huge number of proteins, RNAs and regions of the genomes of different organisms require assessments of these properties to understand their role in various biological processes. Experimental methods are limited in their capabilities due to the complexity, cost, and duration of such studies. Since the mid-20th century, amino acid (aa) and nucleotide (nt) sequences, described as a single-letter code, have been used to analyze the structural and functional properties of biological macromolecules. This is the basis of most modern bioinformatics methods, which can be used to effectively evaluate the properties of biological macromolecules. Despite significant advances in the creation of such methods, the very idea of the representation of aa and nt sequences in the form of a single-letter code is a limitation for further improvement of bioinformatics methods, since it does not reflect the structural features of these sequences at the level of connections between atoms. On the other hand, the development of experimental methods for determining the 3D structures of biological macromolecules makes it possible to describe their atoms in three-dimensional space. Such a description

makes it possible to evaluate their characteristics and properties much better, but requires the use of significant computing resources, and is also limited by the number of known three-dimensional structures.

Methods and Algorithms: Our research offers an alternative to representing aa and nt sequences of proteins in the form of structural formulas of their fragments, which makes it possible to use methods previously developed in chemoinformatics for identifying structure-property relationships. To describe the structural formulas of fragments of macromolecules, high-level substructural descriptors of multilevel atomic neighborhoods MNA (Multilevel Neighborhoods of Atoms) are used, implemented in specially developed for working with such data MultiPASS program, which uses the Bayesian algorithm to identify the “structure-property” relationships.

Results: We have demonstrated the effectiveness of this approach in methods for predicting: pathogenic amino acid substitutions (<https://www.way2drug.com/SAV-Pred/>) [1, 2], amino acid substitutions associated with drug resistance of tumors [3], specificity of T-cell receptors for epitopes (<https://www.way2drug.com/TCR-Pred/>) [4] and post-translational modifications [5]. The talk will also demonstrate the use of our approach to predict secondary structures and functional motifs of proteins, as well as microRNA targets and binding sites for human transcription factors.

Conclusion: The proposed approach represents a new direction in bioinformatics in the study of the properties of biological macromolecules based on the structural formulas of their fragments.

Funding: This work has been supported within the framework of the Program for Basic Research in the Russian Federation for a long-term period (2021–2030) (No. 122030100170-5).

Список литературы/References

1. Zadorozhny A., Smirnov A., Filimonov D., Lagunin A. Prediction of pathogenic single amino acid substitutions using molecular fragment descriptors. *Bioinformatics*. 2023;39(8):btad484. doi 10.1093/bioinformatics/btad484
2. Zadorozhny A.D., Rudik A.V., Filimonov D.A., Lagunin A.A. SAV-Pred: a freely available web application for the prediction of pathogenic amino acid substitutions for monogenic hereditary diseases studied in newborn screening. *Int J Mol Sci*. 2023;24:2463. doi 10.3390/ijms24032463
3. Zhuravleva S.I., Zadorozhny A.D., Shilov B.V., Lagunin A.A. Prediction of amino acid substitutions in abll protein leading to tumor drug resistance based on “Structure-Property” relationship classification models. *Life*. 2023;13:1807. doi 10.3390/life13091807
4. Smirnov A.S., Rudik A.V., Filimonov D.A., Lagunin A.A. TCR-Pred: A new web-application for prediction of epitope and MHC specificity for CDR3 TCR sequences using molecular fragment descriptors. *Immunology*. 2023;169(4):447-453. doi 10.1111/imm.13641
5. Karasev D.A., Savosina P.I., Sobolev B.N., Filimonov D.A., Lagunin A.A. Application of molecular descriptors for recognition of phosphorylation sites in amino acid sequences. *Biomed Khim*. 2017;63(5):423-427. doi 10.18097/PBMC20176305423