

Создание оптимальных структурных алфавитов для предсказания локальной и третичной структуры белка методами машинного обучения

Мильчевский Ю.В.^{1*}, Мильчевская В.Ю.¹, Тевонян Л.Л.¹, Кретьова А.Н.¹,
Кравацкий Ю.В.^{1, 2}

¹ Институт молекулярной биологии им. В.А. Энгельгардта РАН, Москва, Россия

² Центр высокоточного редактирования и генетических технологий для биомедицины,

Институт молекулярной биологии им. В.А. Энгельгардта РАН, Москва, Россия

* milch@eimb.ru

Ключевые слова: предсказание структуры белков; локальная структура белка; структурные алфавиты (СА); выравнивание символьных последовательностей

Мотивация и цель: Предсказание структуры белка по аминокислотной последовательности является одной из центральных задач молекулярной биологии и биофизики. Несмотря на выдающиеся достижения методов машинного обучения (AlphaFold [1], ESMFold [2]), качество предсказания структуры белков, для которых данные о структуре гомологов отсутствуют в базах данных, по-прежнему часто остается низким. Более того, до сих пор нет общепризнанного представления локальной структуры [3–7]. Мы развиваем общие принципы формирования структурных алфавитов (СА) в зависимости от класса задач, для которых эти СА будут использоваться. Результатом применения этих принципов должно быть создание общего алгоритма формирования таких СА.

Методы и алгоритмы: Число структурных классов должно быть достаточным для адекватного представления хода главной цепи. Более длинные фрагменты требуют большего числа структурных классов для представления многообразия возможных конформаций белковой цепи. Ключевым параметром, определяющим выбор структурных элементов, является мера различия (метрика) между ними. В нашей работе [8] мы выбрали RMSD [9] и строго обосновали минимальное число структурных элементов заданной длины, необходимых для взаимно однозначного соответствия между координатами главной цепи и представлением локальной структуры, созданным на этих структурных элементах. В [8] приведен алгоритм, позволяющий однозначно восстановить 3D координаты по локальной структуре. Следовательно, для любой конформации фрагмента главной цепи по набору расстояний RMSD до каждого из структурных элементов можно восстановить его 3D координаты. Таким образом, мы **разложили многомерную задачу предсказания структуры белка на несколько** (по числу структурных классов) **отдельных одномерных задач**. Каждая такая задача направлена на количественную оценку взаимосвязи между последовательностью белка, статистикой его структурных конформаций и физико-химическими свойствами составляющих белок аминокислот. Для решения таких задач мы применяли пошаговый регрессионный анализ [10], который позволяет произвести отбор статистически значимых предикторов из большого первичного набора. Процедура выбора значимых предикторов позволяет проверять предположения о физико-

химических параметрах, определяющих структуру белка (физико-химические свойства аминокислот (>550) собраны в базе данных AAindex [11]). Для генерации предикторов нами предложена методика формирования и оптимизации наборов статистически значимых предикторов для последующего использования в задачах предсказания структуры и функций белка методами машинного обучения [12]. Восстановление декартовых координат по предсказанной локальной структуре, описанное в [8], на практике не работает ввиду быстрого накопления ошибок предсказания при сборке 3D структуры из отдельно предсказанных фрагментов. Для использования локальной структуры при сборке пространственной структуры всей белковой цепи мы предлагаем новый подход:

1. Предсказанная локальная структура записывается как символьная последовательность. Например, для РВ [6] получается СА из 16 символов.
2. Для обучающей выборки предсказанные символьные последовательности СА, очевидно, однозначно выравниваются с символьными последовательностями, полученными из реальных трехмерных координат без вставок и делеций. Это позволяет создать по алгоритму [13] матрицу замен (substitution matrix), которая необходима для применения программ выравнивания [14].
3. При помощи программ выравнивания проводится сканирование банка данных PDB [15] для поиска «гомологичных» в смысле СА участков. Затем из таких участков с известной трехмерной структурой создаются начальные приближения для методов молекулярной динамики и механики.

Результаты: В нашей работе [16] мы применили нейронные сети множественной регрессии (multiple regression NN) для предсказания локальной и вторичной структур белка. Наш метод превзошел все методы, описанные в литературе, для классического набора негомологичных белковых цепей CB513 [17]: для вторичной структуры DSSP [7] Q3 = 85.99 %, Q8 = 79.35 %; для 16 РВ [6] на выборке CB513 Q16 = 81.01 % [16]. Далее, для РВ с использованием полученной в [16] предсказательной модели были выполнены предсказания локальных структур обучающей выборки, после преобразования которых в СА была построена матрица замен и оптимизирована процедура выравнивания методом Smith–Waterman из пакета FASTA [14].

В целях развития нашей методики был проведен расширенный поиск оптимальных структурных алфавитов и длин фрагментов, результатом чего стал СА, состоящий из 50 структурных элементов длиной 7 аминокислотных остатков. Для этого структурного алфавита были выполнены все этапы исследования, включая обучение нейронных сетей и создание аналогичной процедуры выравнивания.

Выводы: Процедура выравнивания предсказанных символьных последовательностей СА по базе символьных последовательностей, полученных из реальных трехмерных координат из PDB банка, не зависит от наличия гомологов в общепринятом смысле этого термина, поскольку наш алгоритм предсказания не использует информацию о наличии гомологии по аминокислотной последовательности, а базируется на физико-химических свойствах аминокислот и статистике встречаемости структурных элементов. Структурные алфавиты, использующие метрику RMSD, лучше (согласно нашему предварительному анализу), чем широко известные РВ [6], описывают ход белковой цепи при длине структурных элементов от 5 до 11 остатков и позволяют

в итоге получить более адекватные начальные приближения для дальнейшего использования методов молекулярной динамики.

Финансирование: Исследование поддержано грантом Российского научного фонда РНФ (№ 24-24-00493).

Development of optimal structural alphabets for predicting protein local and tertiary structures by machine learning methods

Milchevskiy Y.V.^{1*}, Milchevskaya V.Y.¹, Tevonyan L.L.¹, Kretova A.N.¹,
Kravatsky Y.V.^{1, 2}

¹ Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia

² Center for Precision Genome Editing and Genetic Technologies for Biomedicine,
Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia

* milch@eimb.ru

Key words: protein structure prediction; local protein structure; structural alphabets (SA); sequence alignment

Motivation and Objective: Protein structure prediction from sequence stands as a central task in molecular biology and biophysics. Despite remarkable advancements in machine learning methods (such as AlphaFold [1] and ESMFold [2]), the quality of protein structure prediction remains often low for proteins lacking structural homolog data in databases. Moreover, a universally accepted representation of local structure is still absent [3–7]. Here we develop general principles for constructing such representations, further referred to as Structural Alphabets (SAs), and tailoring them to specific tasks at hand.

Methods and Algorithms: The number of structural classes should be sufficient to adequately represent any protein backbone: namely, it should be possible to reconstruct the Cartesian coordinates of the backbone from its SA representation. A key parameter determining the selection of structural elements is the measure of dissimilarity (metric) between them. In [8], we have chosen RMSD [9] and rigorously justified the minimum number of structural elements of a given length necessary for a one-to-one correspondence between the backbone coordinates and the representation of local structure created from these structural elements. Further, we presented an algorithm to reconstruct the Cartesian coordinates from the SA representation.

Thus, we decomposed a highly complex problem of protein structure prediction into several one-dimensional problems, each corresponding to a single structural class. Each one-dimensional problem is aimed at quantitatively assessing the relationship between the protein sequence, the statistics of its structural conformations (dependent on the training data), and the physicochemical properties of the constituent amino acids (independent of the training data). To address these tasks, we employed stepwise regression analysis [10], which enables the selection of statistically significant predictors from a large initial set. The procedure for selecting significant predictors allows for testing suggestions about the physicochemical parameters that determine protein structure (physicochemical properties of amino acids (>550) are assembled in the AAindex database [11]), and thus increases interpretability of the method. We have

further proposed a methodology for generating and optimizing sets of statistically significant predictors for subsequent use in protein structure and function prediction [12]. Same as the exact SA classes, the predicted structural class assignments can be used to reconstruct the Cartesian coordinates of the backbone [8]. However, it is hindered by rapidly accumulating prediction error. To mitigate this, we propose a new approach: First, we represent a protein sequence in terms of its predicted SA assignment classes for each amino acid, which gives rise to a new symbolic sequence (for instance, Protein Blocks proposed in [6] give rise to a 16-symbol alphabet, and the structural class assignment corresponds to a fragment centered at the corresponding amino acid in the protein). Second, for the proposed symbolic representations, we perform a sequence alignment against symbolic representations of a large pool of proteins with known structures (for details see [13, 14]). While the alignment algorithms use symbolic sequences, there is structural information behind each of the SA symbols. This way, we identify structural homology between the query structure and the structures in the PDB [15], and use it to construct a “first 3D approximation” that serves as a starting point for molecular dynamics algorithms.

Results: We developed a method for local protein structure prediction using SAs, and evaluated it for the secondary structure prediction, DSSP Q8 class prediction and the 16-class Protein Block class prediction tasks using the CB513 data set [6, 7, 17]. In all tasks, it outperformed the existing methods showing the following results: Q3 = 85.99 %, Q8 = 79.35 %; Protein Blocks based Q16 = 81.01 % [16]. Further, we applied the 16 class Protein Blocks alphabet to fine-tune the alignment methods: namely, we used the true and predicted SA class assignments for the training set proteins to construct a substitution matrix for the aligner.

As another branch of this research, we aimed to identify an optimal set of SA classes and an optimal length of the corresponding peptides which resulted in a SA consisting of 50 structural elements with a length of 7 amino acids. For this structural alphabet, all stages of research were performed, including training the neural networks and creating a similar alignment procedure.

Conclusion: The proposed SA-based alignment procedure, as well as the local structure prediction algorithm, does not rely on the existence of sequence homologs in terms of the amino acid sequence. Instead, it relies on structural homologs, or homologs in terms of the SA-assigned sequence, which allows for structural prediction for proteins with no existing sequence homologs among the resolved structures – a protein class poorly covered even by the best available methods. For the query protein, it would be the predicted SA classes, and for the reference structures these are the true SA class assignments. Moreover, our method for the SA class prediction does not rely on amino acid sequence homology either but is based on physicochemical properties of amino acids and the occurrence statistics of the structural elements in the PDB. Using RMSD as a measure of dissimilarity, we observed that optimal length of a basic structural fragment to best describe the trajectory of the protein backbone is between 5 and 11. It ultimately provides a more adequate initial approximations for further application of molecular dynamics methods.

Funding: This research was supported by the Russian Science Foundation grant (No. 24-24-00493).

Список литературы/References

1. Jumper J. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-589
2. Lin Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 2023;379:1123-1130
3. Han K.F. et al. Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Sci*. 1997;6:1587-1590
4. Sander O. et al. Local protein structure prediction using discriminative models. *BMC Bioinformatics*. 2006;7:14
5. Kloczkowski A. et al. Distance matrix-based approach to protein structure prediction. *J Struct Funct Genomics*. 2009;10:67-81
6. de Brevern A.G. et al. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*. 2000;41:271-287
7. Kabsch W., Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22:2577-2637
8. Milchevskaya V. et al. Structural coordinates: A novel approach to predict protein backbone conformation. *PLoS One*. 2021;16:e0239793
9. Kabsch W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr, Sect A: Found Crystallogr*. 1976;32:922-923
10. Ralston A., Wilf H.S., Enslein K. Mathematical Methods for Digital Computers. New York: Wiley, 1960
11. Kawashima S. et al. AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res*. 2008;36:D202-D205
12. Milchevskiy Y.V. et al. Method to generate complex predictive features for machine learning-based prediction of the local structure and functions of proteins. *Mol Biol*. 2023;57:136-145
13. Henikoff J.G., Henikoff S. Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci*. 1996;12:135-143
14. Pearson W.R. Blast and fasta similarity searching for multiple sequence alignment. *Methods Mol Biol*. 2014;1079:75-101
15. Berman H.M. et al. The Protein Data Bank at 40: Reflecting on the past to prepare for the future. *Structure*. 2012;20:391-396
16. Milchevskiy Y.V. et al. Effective local and secondary protein structure prediction by combining a neural network-based approach with extensive feature design and selection without reliance on evolutionary information. *Int J Mol Sci*. 2023;24(21):15656
17. Cuff J.A., Barton G.J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*. 1999;34:508-519