МОДЕЛИРОВАНИЕ БЕЛКОВЫХ КОМЛЕКСОВ С УЧЕТОМ НЕСТРУКТУРНОЙ ИНФОРМАЦИИ

А. Ю. Хадарович[1], И. В. Анищенко[2], П. Кундротас[2], И. А. Ваксер[2,3], А. В. Тузиков[1]

[1]*Объединенный институт проблем информатики НАН Беларуси*
Минск, Беларусь
[2]*Центр вычислительной биологии и* [3]*Отделение молекулярных биологических наук, Университет Канзаса*
Лоренс, США
e-mail: ahadarovich@gmail.com

Докинг, основанный на шаблонах, предсказывает структуры белок-белковых комплексов, используя известные структуры в качестве шаблонов для моделирования. Подходящие шаблоны часто выбираются по схожести последовательностей и/или структурной схожести между целевыми белками и шаблоном. Эффективность докинга, основанного на шаблонах, значительно уменьшается, когда шаблоны мало похожи на целевые белки. Поэтому предлагается функциональная мера схожести, которая дополняет структурную меру. Комбинированная функция оценки позволяет улучшить отсеивание неподходящих шаблонов, увеличивая надежность докинга.

Ключевые слова: моделирование белковых комплексов; белок-белковые взаимодействия.

MODELING OF PROTEIN COMPLEXES TAKING INTO ACCOUNT STRUCTURAL INFORMATION

A. Y. Hadarovich[1], I. V. Anishchenko[2], P. J. Kundrotas[2], I. A. Vakser[2,3], A. V. Tuzikov[1]

[1]*United Institute of Informatics Problems, National Academy of Sciences*
Minsk, Belarus
[2]*Center for Computational Biology and* [3]*Department of Molecular Biosciences*
The University of Kansas,
Lawrence, USA

Template-based docking generates structures of protein-protein complexes using known structures as the templates for modeling. Suitable templates are often detected by sequence and/or structure similarity between the target and the template. The performance of the template-based docking significantly decreases when the templates are only moderately similar to the target. Thus a functional similarity score was developed, which is complementary to the structural score. A combined scoring function allows to improve discrimination of wrong templates, enhancing the reliability of the docking.

Keywords: modeling of protein complexes; protein-protein interactions.

Белковые взаимодействия определяют большинство процессов в клетке. Знание пространственной организации белковых молекул является ключом не только к пониманию их функций и механизма работы, но и основой для разработки эффективных и безопасных лекарственных средств. В настоящее время существует большой разрыв между количеством последовательностей белков, которые можно получать быстро и относительно недорого методами секвенирования, и экспериментально определенными пространственными структурами белков, поскольку определять структуру в прямом эксперименте не всегда возможно или целесообразно – из-за сложности, дороговизны и ограниченности возможностей экспериментальных методик. Поэтому для решения данной задачи применяется докинг – метод моделирования, который позволяет предсказать наиболее выгодную для образования устойчивого комплекса ориентацию и положение одной молекулы по отношению к другой.

Алгоритмы предсказания структуры белок-белковых взаимодействий могут использовать шаблоны, т. е. белковые комплексы, для которых уже известна структура. Этот подход называется докингом, основанном на шаблонах. Данный метод моделирования осуществляет поиск комплексов по базе данных белок-белковых комплексов, пространственная структура которых определена ранее с помощью экспериментальных методов. Чтобы найти шаблон, наиболее похожий на исследуемые белки, нужно ввести некоторую оценочную функцию, которая для заданного шаблона и целевых белков задает некоторое значение, характеризующее их схожесть.

Одна из мер, которая позволяет найти структурную схожесть между белками и используется в шаблонном докинге, – TM-score (Template Modeling Score). Она вычисляется при помощи алгоритма TM-Align, в ходе которого производится наложение шаблонного белка на целевой белок (таргет) [1, 2]. Эксперименты показали, что данная величина не зависит от размеров белков и хорошо отражает схожесть белковых пар. Значение TM-score для двух случайных белков равно приблизительно 0,17. Однако существует «серая зона» (когда TM-score принимает значения на отрезке [0,4; 0,8]), в которую попадают как хорошие модели, структурно подобные исследуемому белку, так и плохие [3]. Введение меры, основанной на функциональных описаниях белков, как дополнительной к TM-score, может позволить исключить «серую зону», а значит, и неоднозначность при ранжировании и отборе шаблонов. Для этой цели был рассмотрен иерархический словарь биологических терминов Gene Ontology, которые используются для описания свойств объекта.

GO-термины (Gene Ontology terms) были введены как унификация того естественного языка, которым описываются различные соединения. Они фиксируют три направления описания (так называемые онтологии): *молекулярную функцию*; *биологический процесс*, в котором белок принимает участие; а также *клеточную компоненту*, в состав которой входит данный белок [4]. На основе данных описаний была разработана функциональная мера схожести белок-белковых комплексов (GO-score), а также комбинированная мера, включающая структурную составляющую (TM-score) и три функциональные составляющие (по одной для каждого из направлений Gene On-Ontology)(1).

$$F(TM, GO) = TM \times \sqrt{0{,}43 \cdot GO_{MF} + 0{,}21 \cdot GO_{BP} + 0{,}36 \cdot GO_{CC}}. \qquad (1)$$

TM обозначает значение меры TM-score, GO_{MF}, GO_{BP} и GO_{CC} – значения оценок для молекулярной функции, биологического процесса и клеточной компоненты соответственно.

В качестве тестового набора для рассматриваемой меры было взято множество, состоящее из 587 целевых белков (таргетов) и библиотеки шаблонов, содержащей 4950 экспериментально определенных белок-белковых комплексов [5]. Для оценки качества рассматриваемых мер был проведен докинг. При этом отсекались пары «таргет-шаблон» со значением TM-score менее 0,4, так как такие пары не рассматривались как схожие. Также отсеивались пары комплексов, для которых было невозможно вычислить GO-score хотя бы по одному направлению онтологии в силу недостаточной аннотированности белков. Результаты показали, что использование разработанной комбинированной меры целесообразно в докинге, основанном на шаблонах, так как данная мера позволяет улучшить отбор моделей для построения комплекса и их ранжирование [6, 7].

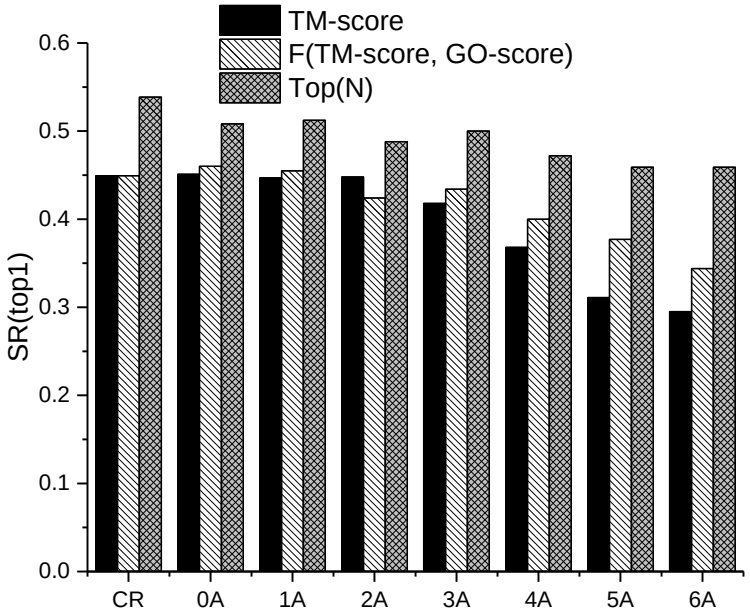
Проведенное исследование [8] говорит о том, что белок-белковых комплексов, хранящихся в PDB (Protein Data Bank), достаточно для того, чтобы найти подходящий шаблон для докинга для почти всех известных белковых взаимодействий, при условии, что компоненты сами имеют известную структуру или она может быть построена на основе гомологии. Однако экспериментально определенных структур индивидуальных белков, которые входят в состав комплекса, пока недостаточно. Поэтому возникает вопрос о том, можно ли использовать модели белков в шаблонном моделировании комплексов. Основная проблема состоит в том, что если данные модели сильно искажены, значительно отличаются от структур белков, полученных экспериментально, то сложно выбрать подходящий шаблонный белок-белковый комплекс, поскольку их отбор осуществляется с помощью структурной меры (например, TM-score). Использование комбинированной меры, использующей кроме структурной информации функциональную, которая не зависит от искажений модели, может облегчить решение данной задачи.

Для проверки данной гипотезы был выбран тестовый набор, состоящий из 165 комплексов. Для белков, входящих в эти комплексы, использовались модели с различной степенью искажения структуры (от 1 до 6 А включительно). Также в набор были включены начальные структуры комплексов (для них искажение равно 0 А, что означает «без искажений») [9].

Чтобы определить, насколько эффективно данная мера позволяет найти шаблон для создания хорошей модели в шаблонном докинге, используется величина Success rate. Она означает долю целевых белковых комплексов (таргетов), для которых в Top(n) найдена хотя бы одна хорошая модель с помощью рассматриваемой меры. Здесь Top(n) – первые n шаблонов в списке, отсортированном в порядке убывания значений рассматриваемой меры. Как видно на рисунке, значение Success rate для комбинированной меры, как правило, выше, чем для структурной.

Кроме того, эффективность заданной меры возрастает по сравнению со структурной мерой по мере увеличения степени искажения моделей. Можно сделать вывод о том, что использование разработанной комбинированной меры схожести белок-белковых комплексов, основанной на функциональной и структурной информации, позволяет улучшить отбор подходящих шаблонных комплексов в шаблонном докинге в случае, когда вместо экспериментально определенных структур используются модели белков.

Таким образом, применение разработанной меры позволяет улучшить качество отбора структур белок-белковых комплексов, как в случае использования экспериментально определенных структур, так и в случае моделей белков, что говорит о широком диапазоне применения данной меры.



Success rate для Top(1) для TM-score и комбинированной *F* меры для моделей с различной степенью искажения

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Zhang Y., Skolnick J. Scoring Function for Automated Assessment of Protein Structure Template Quality // Proteins. 2004. Vol. 57. № 4. P. 702–710.
2. Zhang Y., Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score // Nucleic Acids Research. 2005. Vol. 33. № 7. P. 2302–2309.
3. Negroni J., Mosca R., Aloy P. Assessing the Applicability of Template-Based Protein Docking in the Twilight Zone // Structure. 2014. Vol. 22. № 9. P. 1356–1362.
4. The Gene Ontology Consortium. Gene Ontology Consortium: going forward / The Gene Ontology Consortium // Nucleic Acids Research. 2015. Vol. 43. P. D1049–D1056.
5. Anishchenko I. V., Kundrotas P. J., Tuzikov A. V., Vakser I. A. Structural templates for comparative protein docking // Proteins. 2015. Vol. 83. № 9. P. 1563–1570.
6. A functional ontology-based score for template-based protein docking / A. Y. Hadarovich [et al.] // Abstracts of the 12th International Symposium Bioinformatics Research and Applications ISBRA 2016. Minsk, 2016.
7. Quantitative comparison of functional properties in protein-protein complexes / A. Y. Hadarovich [et al.] // Proceedings of the Moscow Conference on Computational Molecular Biology (MCCMB'15). Moscow, 2015.
8. Kundrotas P. J., Zhu Z., Janin J., Vakser I. A. Templates are available to model nearly all complexes of structurally characterized proteins // Proc Natl Acad Sci USA. 2012. Vol. 109. № 24. P. 9438–9441.
9. Anishchenko I. V., Kundrotas P. J., Tuzikov A. V., Vakser I. A. Protein models: The Grand Challenge of protein docking // Proteins. 2014. Vol. 82. № 2. P. 278–287.