ISSN 1561-8323 (Print)
ISSN 2524-2431 (Online)

УДК 004.021
https://doi.org/10.29235/1561-8323-2020-64-2-150-158

Поступило в редакцию 06.11.2019
Received 06.11.2019

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

[1]Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Республика Беларусь
[2]Университет Вашингтона, Сиэтл, Соединенные Штаты Америки
[3]Университет Канзаса, Лоренс, Соединенные Штаты Америки

АЛГОРИТМ ПРЕДСКАЗАНИЯ СТРУКТУР БЕЛКОВЫХ КОМПЛЕКСОВ НА ОСНОВЕ ГЕННОЙ ОНТОЛОГИИ

Аннотация. Предлагается алгоритм сравнения белок-белковых комплексов на основе их функциональных свойств в терминах генной онтологии. Мера функциональной схожести комплексов интегрируется со структурной мерой для нахождения шаблона для моделирования белковых комплексов. Приводятся результаты моделирования белковых комплексов с помощью предложенного алгоритма.

Ключевые слова: белок-белковые взаимодействия, моделирование белковых комплексов, термины генной онтологии

Для цитирования: Алгоритм предсказания структур белковых комплексов на основе генной онтологии / А. Ю. Хадарович [и др.] // Докл. Нац. акад. наук Беларуси. – 2020. – Т. 64, № 2. – С. 150–158. https://doi.org/10.29235/1561-8323-2020-64-2-150-158

Anna Yu. Hadarovich[1], Ivan V. Anishchenko[2], Petras Kundrotas[3], Ilya Vakser[3], Corresponding Member Alexander V. Tuzikov[1]

[1]United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Republic of Belarus
[2]University of Washington, Seattle, United States of America
[3]University of Kansas, Lawrence, United States of America

STRUCTURE PREDICTION ALGORITHM FOR PROTEIN COMPLEXES BASED ON GENE ONTOLOGY

Abstract. We propose an algorithm for comparing protein-protein complexes based on their functional properties in terms of Gene Ontology. The proposed measure of a functional similarity between complexes is combined with a structural measure to find templates for the template-based docking of protein complexes. We present the results on the modeling of protein complexes based on this algorithm.

Keywords: protein-protein interactions, modeling of protein complexes, gene ontology terms

For citation: Hadarovich A. Yu., Anishchenko I. V., Kundrotas P., Vakser I., Tuzikov A. V. Structure prediction algorithm for protein complexes based on gene ontology. *Doklady Natsional'noi akademii nauk Belarusi = Doklady of the National Academy of Sciences of Belarus*, 2020, vol. 64, no. 2, pp. 150–158 (in Russian). https://doi.org/10.29235/1561-8323-2020-64-2-150-158

Введение. Белковые взаимодействия определяют большинство процессов в клетке. При всем разнообразии работа белков всегда базируется на их высоко специфическом взаимодействии, для которого необходима определенная пространственная структура. Поэтому биологическая функция белков тесно связана с существованием их в виде трехмерных структур. Даже небольшие изменения этих структур часто ведут к утере или резкому изменению активности белков.

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 то же время определять структуру белков в прямом эксперименте не всегда возможно или целесообразно – из-за сложности, дороговизны или ограниченности возможностей экспериментальных методик. Иногда удается преодолеть эти сложности, подойдя к проблеме с другой стороны: структуру белков можно предсказать, используя теоретические подходы, основанные на физических или эмпирических приближениях.

Алгоритмы предсказания структуры белок-белковых взаимодействий могут использовать шаблоны (белковые комплексы, для которых структура уже известна). При таком подходе по базе данных осуществляется поиск белок-белковых комплексов, пространственная структура которых определена ранее экспериментально или с помощью методов моделирования. Чтобы найти шаблон, наиболее похожий на исследуемые белки, нужно ввести некоторую оценочную функцию, которая для заданного шаблона и целевых белков характеризует их схожесть.

Целью исследования является разработка алгоритма нахождения функциональной схожести между белками и белок-белковыми комплексами на основе их биологических описаний из словаря терминов генной онтологии (Gene Ontology). Применение разработанной меры схожести в комбинации с мерой схожести по структуре белков предназначено для поиска подходящих шаблонов при моделировании структур белок-белковых комплексов, основанном на шаблонах.

Материалы и методы исследования. *Термины генной онтологии для описания свойств белков*. Go-термины (Gene Ontology terms) – иерархический словарь терминов, которые используются для описания свойств генов и кодируемых ими белков. 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. База данных, в которой хранятся эти описания, называется GeneOntology [1]. Информация о go-терминах белка содержится также в других базах данных. Для некоторых белков описание в терминах генной онтологии может отсутствовать.

Каждый go-термин имеет свой уникальный идентификатор (который имеет вид GO:*xxxxxxx*, где *x* – цифра), название, кратко характеризующее описываемое им свойство, и подробное описание, а также занимает определенное место в иерархической структуре. Каждое направление (функция, процесс и компонента) задается ациклическим орграфом. Каждому термину поставлена в соответствие вершина графа, отождествляемая с самим термином. Термины могут быть связаны отношением включения, могут также иметь и другие типы связей, но они не рассматриваются в данной работе, так как эти связи считаются менее значимыми. Отношение включения определяет дугу в графе от вершины, задаваемой термином-потомком, в вершину, задаваемую родительским термином. Это отношение имеет следующий смысл: свойство, описываемое термином-потомком, является подтипом свойства, описываемого родительским термином. Отношение включения является транзитивным. В отличие от других иерархических словарей go-термин может иметь несколько родительских вершин, что обеспечивает большую гибкость, но усложняет систему в целом.

Гибкость и полнота словаря go-терминов позволяют использовать его для решения различных биологических задач, включающих работу с белками, таких как прогнозирование и проверка белок-белковых взаимодействий и исследование экспрессии генов. Данная онтология зарекомендовала себя как надежный и эффективный источник описания биологической информации.

Термин t_c является потомком t_p, если в графе существует путь (t_c, ..., t_p), а t_p называется предком t_c. Обозначим через ANC(t) и DES(t) множества всех предков и потомков термина t, причем термин рассматривается как предок и потомок самого себя, т. е. $t \in \mathrm{ANC}(t)$ и $t \in \mathrm{DES}(t)$.

В работе рассматривается алгоритм, основанный на «информационной значимости» терминов [2]:

$$IC(t) = -\log(P(t)) = -\log\left(\frac{N_{\mathrm{DES}}(t)}{N_{\mathrm{all}}}\right),$$

где $N_{\mathrm{DES}}(t)$ – количество элементов в DES(t) в используемом источнике данных (в данной работе рассматриваются вхождения терминов в базу данных белков UniprotKB); N_{all} – количество вхождений всех терминов в этот источник данных.

Можно заметить, что чем дальше данный термин расположен в графе от термина, не имеющего исходящих дуг (такую вершину будем называть корневой), т. е. чем длиннее путь, соединяющий рассматриваемый термин и корневой, тем большим будет значение информационной значимости для него. Для корневого термина данная величина будет равна 0, если для него все

термины в словаре являются потомками. Смысл этой величины следующий: чем больше ее значение для термина, тем он более информативен.

За схожесть между двумя терминами t_1 и t_2 можно принимать следующую величину [3]:

$$S_{\text{Resnik}}(t_i, t_j) = IC(t_i, t_j) = \max\{IC(t) \mid t \in \{\text{ANC}(t_i) \cap \text{ANC}(t_j)\}\}. \quad (1)$$

Серьезным недостатком формулы (1) является то, что значение схожести, вычисленное этим способом, не зависит от уровня местоположения терминов в графе. Общий термин, расположенный близко к корню графа, может показывать одно и то же значение схожести для всех потомков другого общего термина, что может привести к необъективным результатам.

Для того чтобы избежать этого недостатка, была предложена мера схожести, учитывающая величины информационной значимости для рассматриваемых терминов [4],

$$S_{\text{Shlicker}}(t_i, t_j) = \frac{2IC(t_i, t_j)(1 - P(t))}{IC(t_i) + IC(t_j)},$$

где t – общий предок для t_1 и t_2, имеющий максимальную информационную значимость.

Алгоритм нахождения функциональной схожести между белок-белковыми комплексами. Для нахождения схожести между белками p_k и p_l используется метод, предложенный в [5]. Обозначим через $T(p)$ множество терминов, которыми описывается белок p, а через $N_{T(p_k)}$ – количество терминов, описывающих белок p_k. Вычисление производится по формулам

$$S(t_i, T(p_l)) = \max\{S(t_i, t_j) \mid t_j \in T(p_l)\},$$

$$S(p_k, T(p_l) = \frac{1}{N_{T(p_k)}} \sum_{i=1}^{N_{T(p_k)}} S(t_i, T(p_l)),$$

$$S(p_k, p_l) = \frac{S(p_k, T(p_l)) + S(p_l, T(p_k))}{2}.$$

Схожесть между двумя белок-белковыми комплексами c_m и c_n, состоящими из белков p_{m1}, p_{m2} и p_{n1}, p_{n2} соответственно, вычисляется двумя способами в зависимости от способа наложения комплексов друг на друга:

$$S_{11}(c_m, c_n) = \frac{S(p_{m1}, p_{n1}) + S(p_{m2}, p_{n2})}{2},$$

$$S_{12}(c_m, c_n) = \frac{S(p_{m1}, p_{n2}) + S(p_{m2}, p_{n1})}{2}.$$

Рассматриваются два способа наложения комплексов вида «цель–шаблон». В первом случае первый белок первого комплекса p_{m1} накладывается на первый белок второго комплекса p_{n1}, а второй белок первого комплекса p_{m2} накладывается на второй белок второго комплекса p_{n2} (будем обозначать такое наложение как S_{11}, или прямое наложение). Во втором случае первый белок первого комплекса накладывается на второй белок второго комплекса, а второй – на первый (будем обозначать S_{12}, или перекрестное наложение).

Функциональная схожесть, вычисляемая по формулам, приведенным выше, будет обозначаться как GO-score или GO_X, где X задает молекулярную функцию *MF* или биологический процесс *BP* или клеточную компоненту *CC*.

Алгоритм нахождения структурной схожести между белок-белковыми комплексами. Одной из оценочных функций, позволяющих найти подходящий шаблон, на основе которого будет строиться модель, является TM-score (Template Modeling Score) [6], вычисляемая следующим образом:

$$\text{TM-score}(p_m, p_n) = \max\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)}\right],$$

где p_m и p_n – исследуемый и шаблонный белок соответственно; L_N – длина исследуемой белковой структуры; L_T – количество выровненных по структуре шаблона остатков; d_i – расстояние между i-й парой выровненных остатков и d_0 – параметр нормализации, максимум вычисляется по множеству шаблонов. TM-score принимает значение на отрезке от 0 до 1 и вычисляется при помощи алгоритма TM-align, который осуществляет наложение шаблонного белка на целевой белок (таргет) [7]. Наложенный таким образом шаблон будем называть моделью белка. Модель комплекса будут образовывать белки, составляющие шаблон, наложенные на белки целевого комплекса. Известно из литературы, что оценка TM-score хорошо отражает схожесть пар «целевой белок–модель». Значение TM-score для двух случайных белков близко к 0,17 по данным экспериментальных исследований. Однако существует так называемая 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 [8]. В сообщении показано, что можно совместно использовать структурную и функциональную информацию о белках для поиска подходящего шаблона с учетом «серой зоны» для оценки TM-score.

В данном исследовании TM-score применялся для систематического поиска шаблонов в наборе из 4950 комплексов [9], состоящих их двух белков. Для этого использовался протокол моделирования, основанного на шаблонах [10]. Для каждого комплекса целевые белки структурно накладывались на белки шаблона с помощью программы TM-align, и качество наложения оценивалось с помощью оценки TM-score. Только модели, сгенерированные на основе наложений с TM-score большим 0,4 для обоих белков, учитывались для дальнейших вычислений. Модели ранжировались по минимальному из двух значений TM-score:

$$TM = \min\{\text{TM-score}(S_{11}), \text{TM-score}(S_{12})\}, \quad (2)$$

где S_{11} и S_{12} – выравнивание p_{m1} на p_{n1} и p_{m2} на p_{n2} или p_{m1} на p_{n2} и p_{m2} на p_{n1} соответственно.

Моделирование белковых комплексов с учетом функциональной и структурной информации. Использование функциональной и структурной схожести в белок-белковом моделировании. Для оценки качества рассматриваемых мер схожести белков был проведен докинг – метод молекулярного моделирования, который позволяет предсказать наиболее выгодную для образования устойчивого комплекса ориентацию и конформацию одной молекулы в центре связывания другой. Данные о положении и конформации партнеров используются для предсказания силы взаимодействия посредством оценочных функций. Для каждой пары комплексов «цель–шаблон» при каждом из двух возможных наложений (прямом и перекрестном) было вычислено значение TM-score, три значения GO-score для соответствующих направлений онтологии (молекулярной функции, клеточной компоненты и биологического процесса), среднеквадратическое отклонение LRMSD (данную меру рассматриваем в качестве эталонной). Отсеивались пары комплексов, для которых отсутствовало значение функциональной схожести (GO-score) хотя бы по одному направлению онтологии в силу недостаточной информации о белках.

Рассматривалась задача классификации. Модель для целевого белок-белкового комплекса, построенная на основе рассматриваемого шаблона, считалась корректной, если значение LRMSD для нее и целевого белкового комплекса не превышало 10 Å. Требовалось распознать корректные модели при помощи вычисленных для целевых комплексов и соответствующих им моделей величин TM-score и GO-score. Предполагалось, что для целевых комплексов и корректных для них моделей будут наблюдаться высокие значения данных мер. В качестве искомой меры использовались следующие комбинации структурной и функциональных мер, обозначаемых как LIN и PC соответственно:

$$f_{\text{lin}}(\text{TM}, \text{GO}) = t\text{TM} + a\text{GO}_{MF} + b\text{GO}_{BP} + c\text{GO}_{CC}, \quad (3)$$

$$f_{pc}(\text{TM}, \text{GO}) = TM(a\text{GO}_{MF} + b\text{GO}_{BP} + c\text{GO}_{CC}). \quad (4)$$

Вторая комбинация (*pc*) рассматривалась для того, чтобы увеличить разрыв между моделями, которые проявляют наибольшую схожесть относительно геометрических и функциональных свойств с целевыми белками и остальными комплексами для лучшего детектирования подходящих шаблонов. Для нахождения весовых коэффициентов использовался алгоритм дифференциальной эволюции [13]. В качестве целевой функции, которая максимизировалась, использовалось значение площади под кривой точность–полнота (Precision–Recall). Точность в данном случае определяется как количество корректных моделей, определенных таковыми рассматриваемой мерой, деленное на общее количество моделей, определенных как корректные данной мерой. Полнота определяется как количество корректных моделей, определенных таковыми рассматриваемой мерой, деленное на фактическое количество корректных моделей. Анализ этих данных показал, что использование трех направлений онтологии по отдельности является менее эффективным для выбора наиболее подходящих шаблонов по сравнению с TM-score. Найденные коэффициенты показаны в таблице.

Таблица коэффициентов функций (3) и (4)

The table of coefficients of functions (3) and (4)

Функция Function	*a*	*b*	*c*	*t*
LIN	0,24	0,18	0,08	0,5
PC	0,47	0,34	0,19	–

Формирование тестового множества. Для получения необходимых описаний для белок-белковых комплексов использовались базы данных GeneOntology [1] и PDB [11]. Эффективность оценочных функций (3) и (4) определялась по предсказаниям структурных выравниваний, полученных для трех наборов белков из базы данных DOCKGROUND [9; 10], с использованием библиотеки шаблонов из полных структур комплексов, образованных парами белков [12; 14].

Первый набор состоял из белок-белковых структур в связанном состоянии. Это означает, что структуры компонент (белков в составе комплекса) были получены экспериментально в составе комплекса. Они отличаются от экспериментально полученных структур белков в несвязанном состоянии (не в составе комплекса) – второго набора из DOCKGROUND 4 [10]. Это объясняется тем, что белки, вступая во взаимодействие, могут изменять свою пространственную структуру в процессе формирования белок-белкового комплекса. Третий набор был образован структурами белок-белковых комплексов, полученных экспериментально с помощью рентгеновской кристаллографии, и шестью моделями с предварительно определенным среднеквадратичным отклонением альфа-атомов углерода (среднеквадратичное отклонение от реальной структуры (1, 2, …, 6Å) для каждой отдельной структуры белка в 165 белково-белковых комплексах. Необходимость третьего набора заключалась в следующем: для решения задачи докинга, т. е. предсказания структуры белкового комплекса, на вход алгоритма требуется подать структуры составляющих комплекс белков. В лучшем случае эти структуры уже были получены экспериментально (например, с помощью методов рентгеновской кристаллографии или ядерного магнитного резонанса). Однако часто экспериментально определенные структуры белков отсутствуют, и необходимо получить модели этих структур. Модели структур белков, полученные при помощи алгоритмов сворачивания (которые предсказывают структуру белка по последовательности), являются менее точными, чем экспериментальные. Для того чтобы исследовать сценарий моделирования белок-белковых комплексов, приближенный к реальной ситуации, когда используются смоделированные структуры белков и подаются на вход алгоритма предсказания структур белок-белковых комплексов, был введен третий набор данных. Это было сделано для

того, чтобы оценить степень влияния искаженных входных данных на конечный результат моделирования.

Результаты и их обсуждение. График, представленный на рис. 1, показывает эффективность оценочных функций в контексте кривых точность–полнота. В терминах площади под данными кривыми функции LIN и PC показывают почти идентичную эффективность, обе значительно превосходят TM-score, которая обозначена на рисунке как TM. Тем не менее, значительная разница в площади под кривыми между TM-score и комбинированными функциями в наборе структур в связанном состоянии обеспечивается в основном целевыми белками с простыми структурными мотивами. Функция PC была выбрана для идентификации более надежных шаблонов из-за более высоких значений точности. Ранее было показано, что комбинированная оценочная функция значительно увеличила долю корректных моделей среди предсказанных при шаблонном моделировании белок-белковых комплексов [15]. При отборе шаблонов в протоколе моделирования важным этапом является их ранжирование. Полученные результаты показали, что использование комбинированной меры сходства, объединяющей структурную составляющую (TM-score) и функциональную (GO-score), целесообразно при сравнительном моделировании, так как данная мера позволяет улучшить ранжирование моделей для построения белкового комплекса (рис. 2), причем количество корректных моделей, для которых ранг был улучшен с помощью

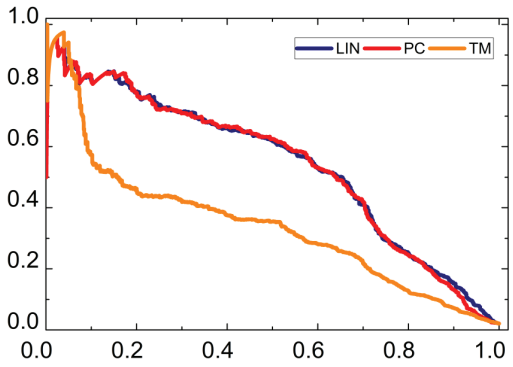

Рис. 1. Сравнение характеристик кривых точность–полнота для различных оценочных функций. Кривые точность–полнота показаны для результатов моделирования с применением функций TM, PC и LIN для структур в связанном состоянии. Функции TM, PC и LIN вычислены с помощью формул (2)–(4) соответственно с коэффициентами из таблицы

Fig. 1. Comparison of precision-recall curves for the different scoring functions. Precision-recall curves are shown for simulation results using the TM, PC, and LIN functions for structures in a bound state. The functions TM, PC and LIN are calculated using formulas (2)–(4), respectively, with the coefficients from Table

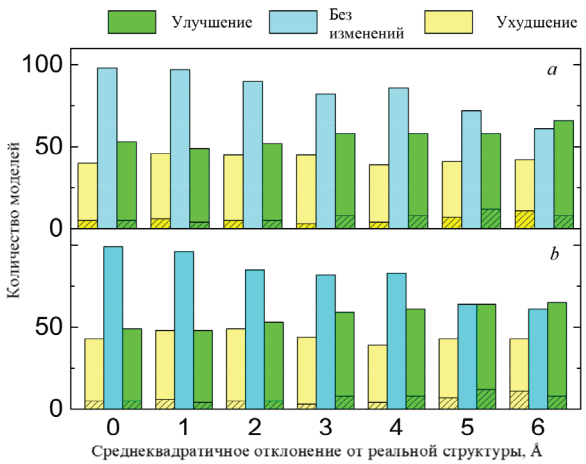

Рис. 2. Сравнение ранжирования моделей белок-белковых комплексов, оцениваемых только по TM и функциям LIN (*a*) и PC (*b*) в наборе данных смоделированных структур. Данные приведены для всего набора данных отдельно для структур, определенных экспериментально (0 Å), и различных уровней искажения моделируемых структур (от 1 до 6 Å). Заштрихованные участки столбцов соответствуют моделям, для которых изменение ранга превышает 10 позиций. Функции TM, LIN и PC рассчитывались по формулам (2)–(4) соответственно. Коэффициенты для функций LIN и PC были взяты из таблицы

Fig. 2. Comparison of the ranking of models of protein-protein complexes, evaluated only on the TM and the LIN (*a*) and PC (*b*) functions in the dataset of modeled structures. The data are presented for the entire dataset separately for structures determined experimentally (0 Å) and different levels of distortion of the modeled structures (from 1 to 6 Å). The stroked sections of the columns correspond to models for which the rank change exceeds 10 positions. The functions TM, LIN, and PC were calculated using formulas (2)–(4), respectively. The coefficients for the LIN and PC functions were taken from Table

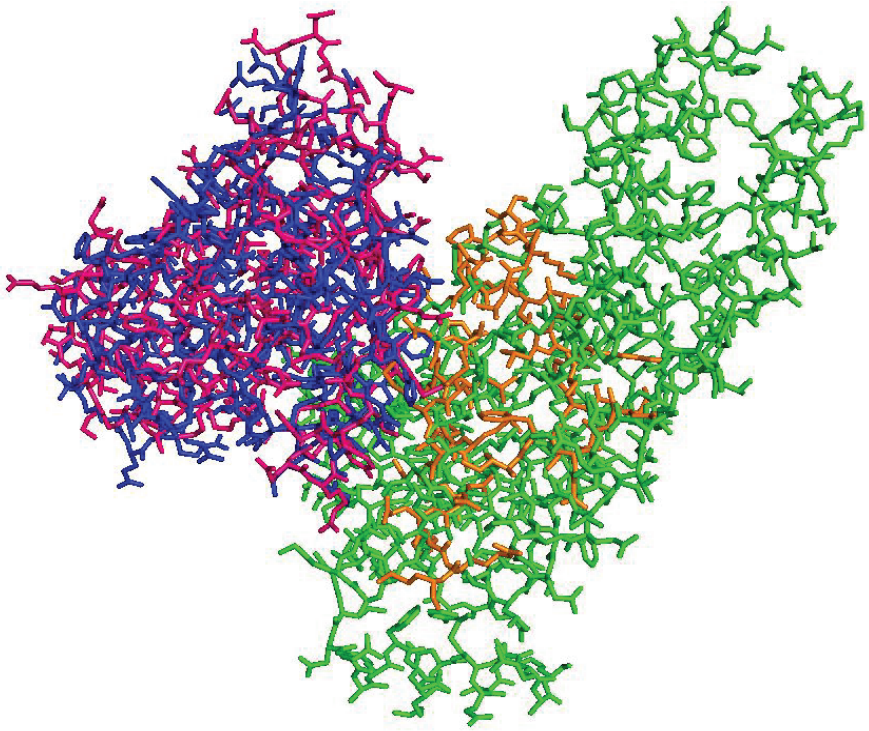

Рис. 3. Пример пары целевых белков, для которой комбинированная функция привела к значительному улучшению ранжирования хороших моделей. Выровненные структуры белков и шаблона показаны синим и оранжевым цветом для целевого комплекса 1j2j (соответственно цепочки A и B) и зеленым и фиолетовым цветом для шаблонного комплекса 2efd (цепочки A и B соответственно)

Fig. 3. The example of a pair of target proteins, for which the combined function has led to a significant improvement in the ranking of good models. The aligned structures of the proteins and the template are shown in blue and orange for the target complex 1j2j (chains A and B, respectively) and green and purple for the template complex 2efd (chains A and B, respectively).

комбинированных функций оценки, также становится большим при увеличении уровня искажения модели. Это увеличение является монотонным с увеличением уровня искажения модели. Основной вклад в этот эффект вносят шаблоны со слабым и умеренным структурным сходством с целевыми белками, т. е. для входных данных, которые представляют собой модели с низким значением структурной схожести с реальной структурой, комбинированная мера позволяет выбрать подходящие шаблоны за счет функциональной оценки.

Примером улучшения ранжирования, благодаря мере, основанной на терминах генной онтологии, может служить модель с низким LRMSD для целевого белок-белкового комплекса 1j2j, имеющего две цепи A и B в записи PDB, построенная на основе шаблонного комплекса между белками, обозначенными как цепи A и B в записи PDB 2efd. Согласно структурной мере TM-score модели комплекса был присвоен низкий ранг (рис. 3). Тем не менее, целевые белки и шаблон имеют довольно высокое сходство во всех трех онтологических доменах, что позволяет комбинированной функции присвоить данной модели высокий ранг.

Заключение. Использование терминов генной онтологии в сравнительном моделировании белок-белковых комплексов увеличивает вероятность выбора корректной модели из числа предсказаний в реальном сценарии, когда структура реального комплекса неизвестна. Комбинированная оценочная функция разделила корректные и некорректные модели значительно лучше, чем оценка, основанная исключительно на структурном выравнивании. В то же время проведенное исследование свидетельствует о том, что функциональные свойства белок-белковых комплексов не могут быть использованы в качестве независимого критерия для выбора шаблонов, так как при этом возможно исключение из рассмотрения «корректных» моделей с низким значением функциональной меры. Комбинированная мера имеет преимущество по сравнению с использованием только структурной меры, даже в случае, когда свойства белков представлены

описанием только молекулярной функции и не включают описания направлений биологического процесса и/или клеточной компоненты. Это существенно для практического применения, так как многие белки не имеют аннотации во всех трех областях одновременно. Предложенный подход к распознаванию корректных предсказаний сравнительного моделирования будет становиться все более эффективным по мере того, как все больше белков будут получать высококачественную аннотацию генной онтологии, описывающую их специфические особенности. Также разработанный алгоритм нахождения подходящего шаблона (структуры белкового комплекса) на основе структурной и функциональной информации позволяет эффективно ранжировать модели. Применение данного алгоритма особенно актуально, когда в качестве входных данных используются модели белковых структур, которые менее точны по сравнению с экспериментально определенными. Так как данные модели имеют такое же функциональное описание в терминах онтологии, как и структуры белков, функция оценки, основанная на терминах генной онтологии, при выборе структурных шаблонов присваивает им высокий ранг в случае, когда их свойства похожи на свойства целевых белков, на основе которых моделируется белок-белковый комплекс. Данный алгоритм может быть успешно встроен в существующие алгоритмы определения структур белковых комплексов, основанных на методологии шаблонного моделирования, в качестве одного из этапов.

Список использованных источников

1. Gene Ontology Consortium: going forward / The Gene Ontology Consortium // Nucleic Acids Research. – 2015. – Vol. 43. – P. D1049–D1056. https://doi.org/10.1093/nar/gku1179

2. Metrics for GO based protein semantic similarity: a systematic evaluation / C. Pesquita [et al.] // BMC Bioinformatics. – 2008. – Vol. 9. – P. S4. https://doi.org/10.1186/1471-2105-9-s5-s4

3. Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy / P. Resnik // Proceedings of the 14th International Joint Conference on Artificial Intelligence. – 1995. – Vol. 1. – P. 448–453.

4. A new measure for functional similarity of gene products based on Gene Ontology / A. Schlicker [et al.] // BMC Bioinformatics. – 2006. – Vol. 7, N 1. – Art. 302. https://doi.org/10.1186/1471-2105-7-302

5. Couto, F. M. Measuring semantic similarity between Gene Ontology terms / F. M. Couto, M. J. Silva, P. M. Coutinho // Data & Knowledge Engineering. – 2007. – Vol. 61, N 1. – P. 137–152. https://doi.org/10.1016/j.datak.2006.05.003

6. Zhang, Y. Scoring Function for Automated Assessment of Protein Structure Template Quality / Y. Zhang, J. Skolnick // PROTEINS: Structure, Function, and Bioinformatics. – 2004. – Vol. 57, N 4. – P. 702–710. https://doi.org/10.1002/prot.20264

7. Zhang, Y. TM-align: a protein structure alignment algorithm based on the TM-score / Y. Zhang, J. Skolnick // Nucleic Acids Research. – 2005. – Vol. 33, N 7. – P. 2302–2309.

8. Negroni, J. Assessing the Applicability of Template-Based Protein Docking in the Twilight Zone / J. Negroni, R. Mosca, P. Aloy // Structure. – 2014. – Vol. 22, N 9. – P. 1356–1362. https://doi.org/10.1016/j.str.2014.07.009

9. DOCKGROUND resource for studying protein-protein interfaces / D. Douguet [et al.] // Bioinformatics. – 2006. – Vol. 22, N 21. – P. 2612–2618. https://doi.org/10.1093/bioinformatics/btl447

10. DOCKGROUND: A comprehensive data resource for modeling of protein complexes / P. J. Kundrotas [et al.] // Protein Sci. – 2018. – Vol. 27, N 1. – P. 172–181. https://doi.org/10.1002/pro.3295

11. The Protein Data Bank / H. M. Berman [et al.] // Nucleic Acids Research. – 2000. – Vol. 28, N 1. – P. 235–242. https://doi.org/10.1093/nar/28.1.235

12. Structural templates for comparative protein docking / I. Anishchenko [et al.] // Proteins. – 2014. – Vol. 83, N 9. – P. 1563–1570. https://doi.org/10.1002/prot.24736

13. Das, S. Particle Swarm Optimization and Differential Evolution Algorithms: Technical Analysis, Applications and Hybridization Perspectives / S. Das, A. Abraham, A. Konar // Advances of Computational Intelligence in Industrial Systems. – 2008. – Vol. 116. – P. 1–38. https://doi.org/10.1007/978-3-540-78297-1_1

14. Sinha, R. Docking by structural similarity at protein-protein interfaces / R. Sinha, P. J. Kundrotas, I. A. Vakser // Proteins. – 2010. – Vol. 78, N 15. – P. 3235–3241. https://doi.org/10.1002/prot.22812

15. Gene ontology improves template selection in comparative protein docking / A. Hadarovich [et al.] // Proteins. – 2019. – Vol. 87, N 3. – P. 245–253. https://doi.org/10.1002/prot.25645

References

1. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 2015, vol. 43, pp. D1049–D1056. https://doi.org/10.1093/nar/gku1179

2. Pesquita C., Faria D., Bastos H., Ferreira A. E. N., Falcão A. O., Couto F. M. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 2008, vol. 9, pp. S4. https://doi.org/10.1186/1471-2105-9-s5-s4

3. Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, vol. 1, pp. 448–453.

4. Schlicker A., Domingues F. S., Rahnenführer J., Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 2006, vol. 7, no. 1, art. 302. https://doi.org/10.1186/1471-2105-7-302

5. Couto F. M., Silva M. J., Coutinho P. M. Measuring semantic similarity between Gene Ontology terms. *Data & Knowledge Engineering*, 2007, vol. 61, no. 1, pp. 137–152. https://doi.org/10.1016/j.datak.2006.05.003

6. Zhang Y., Skolnick J. Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins: Structure, Function, and Bioinformatics*, 2004, vol. 57, no. 4, pp. 702–710. https://doi.org/10.1002/prot.20264

7. Zhang Y., Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 2005, vol. 33, no. 7, pp. 2302–2309. https://doi.org/10.1093/nar/gki524

8. Negroni J., Mosca R., Aloy P. Assessing the Applicability of Template-Based Protein Docking in the Twilight Zone. *Structure*, 2014, vol. 22, no. 9, pp. 1356–1362. https://doi.org/10.1016/j.str.2014.07.009

9. Douguet D., Chen H. C., Tovchigrechko A., Vakser I. A. DOCKGROUND resource for studying protein-protein interfaces. *Bioinformatics*, 2006, vol. 22, no. 21, pp. 2612–2618. https://doi.org/10.1093/bioinformatics/btl447

10. Kundrotas P. J., Anishchenko I., Dauzhenka T., Kotthoff I., Mnevets D., Copeland M. M., Vakser I. A. DOCKGROUND: A comprehensive data resource for modeling of protein complexes. *Protein Science*, 2018, vol. 27, no. 1, pp. 172–181. https://doi.org/10.1002/pro.3295

11. Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E. The Protein Data Bank. *Nucleic Acids Research*, 2000, vol. 28, no. 1, pp. 235–242. https://doi.org/10.1093/nar/28.1.235

12. Anishchenko I., Kundrotas P. J., Tuzikov A. V., Vakser I. A. Structural templates for comparative protein docking. *Proteins: Structure, Function, and Bioinformatics*, 2014, vol. 83, no. 9, pp. 1563–1570. https://doi.org/10.1002/prot.24736

13. Das S., Abraham A., Konar A. Particle Swarm Optimization and Differential Evolution Algorithms: Technical Analysis, Applications and Hybridization Perspectives. *Advances of Computational Intelligence in Industrial Systems*, 2008, vol. 116, pp. 1–38. https://doi.org/10.1007/978-3-540-78297-1_1

14. Sinha R., Kundrotas P. J., Vakser I. A. Docking by structural similarity at protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 2010, vol. 78, no. 15, pp. 3235–3241. https://doi.org/10.1002/prot.22812

15. Hadarovich A., Anishchenko I., Kundrotas P. J., Tuzikov A. V., Vakser I. A. Gene ontology improves template selection in comparative protein docking. *Proteins: Structure, Function, Bioinformatics*, 2019, vol. 87, no. 3, pp. 245–253. https://doi.org/10.1002/prot.25645

Информация об авторах

Хадарович Анна Юрьевна – науч. сотрудник. Объединенный институт проблем информатики НАН Беларуси (ул. Сурганова, 6, 220012, Минск, Республика Беларусь). E-mail: ahadarovich@gmail.com.

Анищенко Иван Владимирович – канд. техн. наук, науч. сотрудник. Университет Вашингтона (Сиэтл, США). E-mail: aivan@uw.edu.

Кундротас Петрас – профессор. Университет Канзаса (66045, 2030 Бекер Драйв, Канзас, США). E-mail: pkundro@ku.edu.

Ваксер Илья – профессор, директор. Центр вычислительной биологии. Университет Канзаса (66045, 2030 Бекер Драйв, Канзас, США). E-mail: vakser@ku.edu.

Тузиков Александр Васильевич – член-корреспондент, д-р физ.-мат. наук, профессор, генеральный директор. Объединенный институт проблем информатики НАН Беларуси (ул. Сурганова, 6, 220012, Минск, Республика Беларусь). E-mail: tuzikov@newman.bas-net.by.

Information about the authors

Hadarovich Anna Yu. – Researcher. United Institute of Informatics Problems of the National Academy of Sciences of Belarus (6, Surganov Str., 220012, Minsk, Republic of Belarus). E-mail: ahadarovich@gmail.com.

Anishchanka Ivan V. – Ph. D. (Engineering), Researcher. University of Washington (Seattle, USA). E-mail: aivan@uw.edu.

Kundrotas Petras – Assistant research professor. University of Kansas (66045, 2030 Becker Drive, Lawrence, Kansas, USA). E-mail: pkundro@ku.edu.

Vakser Ilya – Professor, Director. University of Kansas Center for Computational Biology (66045, 2030 Becker Drive, Lawrence, Kansas, USA). E-mail: vakser@ku.edu.

Tuzikov Alexander V. – Corresponding Member, D. Sc. (Physics and Mathematics), Professor, General director. United Institute of Informatics Problems of the National Academy of Sciences of Belarus (6, Surganov Str., 220012, Minsk, Republic of Belarus). E-mail: tuzikov@newman.bas-net.by.