Литература

1. *Bollerslev T.* Generalized Autoregressive Conditional Heteroscedasticity // J. of econometrics. Vol. 31. 1986. № 3. P. 307–327.
2. *Терех В. С.* О стационарности процесса APGARCH(p,q) // Теория вероятностей, случайные процессы, математическая статистика и приложения: материалы Междунар. науч. конф., Минск, 2015. С. 326–328.
3. *Strauman D.* Estimation in Conditionally Heteroscedastic Time Series Models. Berlin, 2005.
4. *Huber P.* Robust statistics. New York, 1981.

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИИ ДЛЯ ПРЕДСКАЗАНИЯ СТРУКТУРЫ БЕЛОК-БЕЛКОВЫХ КОМПЛЕКСОВ

А. Ю. Хадарович

1. ВВЕДЕНИЕ

Знание пространственной организации белковых молекул является основой для разработки эффективных и безопасных лекарственных средств. Определять структуру белков в прямом эксперименте не всегда возможно или целесообразно из-за сложности и дороговизны экспериментальных методик. Однако можно попытаться предсказать структуру белков (комплексов), используя теоретические подходы, основанные на приближениях. Алгоритмы предсказания структуры белок-белковых комплексов могут использовать шаблоны (белковые комплексы, структура которых уже известна). При таком подходе по базе данных осуществляется поиск белок-белковых комплексов, пространственная структура которых определенна ранее с помощью экспериментальных методов. Чтобы найти шаблон, наиболее похожий на исследуемые белки (таргеты), нужно ввести некоторую оценочную функцию, которая для заданного шаблона и целевого белка дает некоторое значение, характеризующее их схожесть. Одной из таких мер является TM-score (Template Modeling Score), которая основывается на информации о координатах атомов белка в пространстве [1]. Данная величина хорошо отражает схожесть пар «целевой белок-модель». Но 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. Задача состоит во введении меры, базирующейся на функциональных свойствах рассматриваемых белков, как дополнительной к TM-score, которая позволит исключить «серую зону», а, значит, и неоднозначность при ранжировании и отборе шаблонов.

GO-термины (Gene Ontology terms) – иерархический словарь терминов, который используется для описания свойств объекта. Термины фиксируют три направления описания (онтологии): *молекулярную функцию*, *биологический процесс* и *клеточную компоненту*. Каждое из трех направлений задаётся ациклическим орграфом. Каждому термину поставлена в соответствие вершина графа. Термины могут быть связны специальным отношением, тогда от вершины, задаваемой термином-потомком, идет дуга в вершину, задаваемую родительским термином. На основании go-терминов можно сравнивать описываемые ими белки и вычислять меру их сходства.

2. АЛГОРИТМ НАХОЖДЕНИЯ СХОЖЕСТИ МЕЖДУ БЕЛОК-БЕЛКОВЫМИ КОМПЛЕКСАМИ

На первом этапе требуется задать схожесть между терминами. Он состоит из следующих шагов [2].

Для каждого термина вычисляется «вероятность», задаваемая формулой (1):

$$P(t) = \frac{N_{base}(t)}{N_{all}}, \tag{1}$$

где $N_{base}(t)$ – количество упоминаний термина и всех его потомков в каком-либо источнике (рассматривалась база данных Uniprot), N_{all} – количество вхождений всех терминов. Вычисляется величина, называемая «информационной значимостью» – $IC(t)$ (2):

$$IC(t) = -log\big(P(t)\big). \tag{2}$$

Множество предков для термина t будем обозначать $Ancestors(t)$. Тогда для терминов t_1 и t_2 множество $CommonAnc(t_1, t_2)$ содержит общих предков терминов t_1 и t_2 (3):

$$CommonAnc(t_1, t_2) = Ancestors(t_1) \cap Ancestors(t_2) \tag{3}$$

За схожесть между терминами принимается величина, вычисленная по формулам (4–5):

$$Share(t_1, t_2) = Max\{IC(t) \mid t \in CommonAnc(t_1, t_2)\}, \tag{4}$$

$$Sim(t_1, t_2) = \frac{2Share(t_1, t_2)(1 - P(t))}{IC(t_1) + IC(t_2)}. \tag{5}$$

На следующем этапе определяется схожесть между белками. Белки задаются соответствующими наборами терминов, и для вычисления схожести между наборами используется best-match average метод (6–8):

$$GO - score(p_1, p_2) =$$

$$= \frac{GO-score(p_1, Terms(p_2)) + GO-score(p_2, Terms(p_1))}{2}, \quad (6)$$

$$GO - score(p_1, Terms(p_2)) =$$

$$= \frac{\sum_{i=1}^{n} GO-score(t_i, Terms(p_2) \mid t_i \in Terms(p_1))}{n}, \quad (7)$$

$$GO - score(t_1, Terms(p_2)) = Max\{Sim(t_1, t_2) | t_2 \in Terms(p_2)\}. \quad (8)$$

Для нахождения схожести между двумя комплексами рассматриваются два наложения комплексов: первое – первый белок первого комплекса («p11») накладывается на первый белок второго комплекса («p21»), второй белок первого комплекса («p21»), соответственно, накладывается на второй белок второго комплекса («p22») (будем обозначать такое наложение «1-2»); второе – первый белок первого комплекса накладывается на второй белок второго комплекса, второй – на первый (будем обозначать «2-1»). Обозначим первый комплекс, состоящий из белков «p11» и «p12», как «C_1», второй, состоящий из белков «p21» и «p22», как «C_2». Для каждого из двух наложений (1-2 и 2-1 соответственно) схожесть между комплексами вычисляется по формулам (9-10):

$$GO - score(C_1, C_2) =$$

$$= \frac{GO-score(p11, p21) + GO-score(p12, p22)}{2}, \quad (9)$$

$$GO - score(C_1, C_2) =$$

$$\frac{GO-score(p11, p22) + GO-score(p12, p21)}{2}. \quad (10)$$

Итоговая функция, включающая геометрическую составляющую (TM-score) и функциональную составляющую (GO-score) выглядит следующим образом:

$$TM + GO - Score(C, C_2) = TM - score(C_1, C_2) \times$$

$$\times \sqrt{\begin{matrix} 0.5 * GO - score_{mol}(C_1, C_2) + 0.23 * GO - score_{bio}(C_1, C_2) \\ +0.27 * GO - score_{cell}(C_1, C_2) \end{matrix}} \quad (11)$$

Для нахождения весовых коэффициентов использовался алгоритм DEPS (Differential Evolution & Particle Swarm Optimization) [3]. График

Precision-Recall (рисунок 1) иллюстрирует преимущество использования комплексной меры по сравнению с «чистым» TM-score.

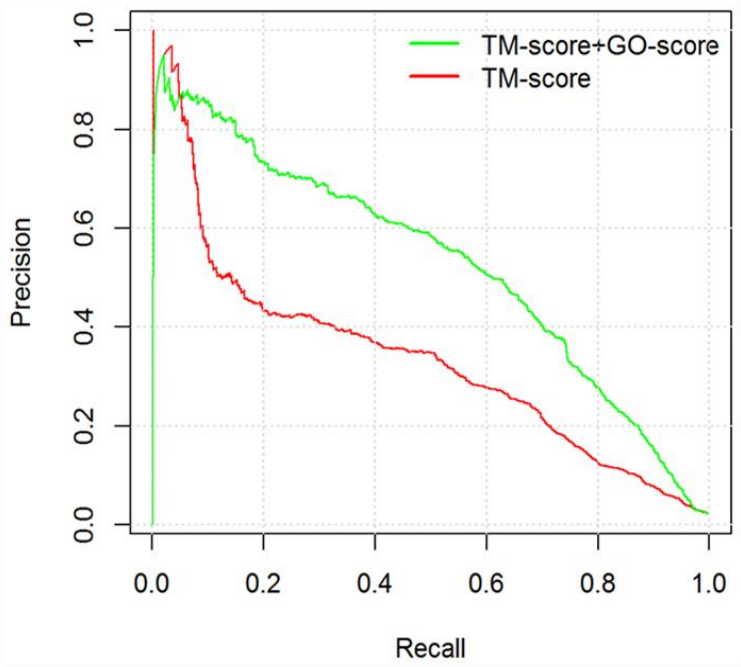

Рис. 1. Precision-Recall кривые для TM-score и итоговой меры

3. ЗАКЛЮЧЕНИЕ

На основании произведенного анализа можно сделать вывод о том, что использование меры, основанной на функциональных свойствах белков, целесообразно при отборе и ранжировании моделей белок-белковых комплексов при сравнительном моделировании, так как дает преимущество по сравнению с использованием только геометрической меры (TM-score), которое наиболее ярко выражено в «серой зоне». Ожидается, что с увеличением объема аннотируемых с помощью рассмотренного словаря GeneOntology белков эффективность разработанной меры будет возрастать.

Литература

1. *Zhang Y., Skolnick J.* Scoring function for automated assessment of protein structure template quality // PROTEINS: Structure, Function, and Bioinformatics. 2002. V. 57. P. 702–710.
2. *Schlicker A. et al.* A new measure for functional similarity of gene products based on Gene Ontology // BMC Bioinformatics. 2006. V. 7. P. 302.

3. *Das S, Abraham A., Konar A.* Particle swarm optimization and differential evolution algorithms: technical analysis, applications and hybridization perspectives // Studies in Computational Intelligence (SCI). 2008. V. 116. P. 1–38.

СОЗДАНИЕ АВТОМАТИЗИРОВАННОЙ ДАКТИЛОСКОПИЧЕСКОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

В. К. Чугунов

ВВЕДЕНИЕ

В дактилоскопии ученые-криминалисты изучают распределение папиллярных узоров на внутренней стороне кисти человека для решения двух основных прикладных задач:

- разработки регистрационных систем как с использованием одного пальца (монопальцевые), так и десятипальцевых;
- поиска преступников по следам пальцев рук, обнаруженных на месте преступления, что и стало фундаментом дактилоскопии.

С развитием технических средств первая задача, ранее решаемая с использованием бумажных картотек, была переведена на автоматический режим считывания и поиска отпечатков пальцев.

В современных устройствах технологии распознавания отпечатков пальцев находят применение уже не с целью поиска преступника, а с целью идентификации пользователя. Так, технологии дактилоскопии применяются и в биометрии. Биометрия – наука, изучающая возможности использования различных характеристик человеческого тела для идентификации человека.

В статье рассматривается способ выявления особых точек (минюций) и сопоставления отпечатков пальцев. Также рассматриваются аспекты реализации автоматизированной дактилоскопической информационной системы на основе определения ориентации изображения, его последующей обработки и выделения особых точек в соответствии с алгоритмом M3GL.

ЭТАП ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ОТПЕЧАТКА

Отпечаток пальца образует так называемые папиллярные линии на гребешковых выступах кожи, разделенных бороздками. Эти линии формируют более сложные узоры (дуги, петли или завитки), уникальные у