# PENTAFOLD 2.0 – АЛГОРИТМ ДЛЯ ПРЕДСКАЗАНИЯ ВТОРИЧНОЙ СТРУКТУРЫ БЕЛКОВ И ПЕПТИДОВ

**[1]Побойнев В. В., [1]Хрусталёв В. В., [2]Хрусталёва Т. А.**

*[1]Белорусский государственный медицинский университет, кафедра общей химии, г. Минск*
*[2]ГНУ «Институт физиологии НАН Беларуси», лаборатория клеточных технологий, г. Минск*

***Ключевые слова:*** компьютерный алгоритм, вторичная структура, вероятностная шкала, чувствительность, специфичность.

***Резюме:*** *создан оригинальный алгоритм для предсказания вторичной структуры белков по аминокислотной последовательности, основанный на вероятностных шкалах. По сравнению с существующими методами данный оригинальный алгоритм отличается более высокой чувствительностью к выявлению бета-тяжей.*

***Abstract:*** *the original algorithm for secondary structure of proteins prediction based on probability scales using amino acid sequence as an input has been created. This original algorithm has a higher sensitivity for beta-strands than currently existing methods.*

**Актуальность.** Предсказание вторичной структуры белка является одной из наиболее важных задач, преследуемых биоинформатикой, поскольку имеет большое значение для медицины (например, в разработке лекарственных препаратов белковой природы) и для биотехнологии (например, при создании новых вакцинных пептидов). Все алгоритмы для предсказания вторичной структуры можно разделить на две большие группы: основанные на гомологии и не основанные на гомологии (вероятностные). Первые используют в своей работе сведения о белках, которые уже были изучены методами рентгеноструктурного анализа или с помощью ядерно-магнитного резонанса, из международной базы данных PDB (www.pdb.org). Работа алгоритмов, не основанных на гомологии, заключается в сопоставлении аминокислотной последовательности с вероятностными шкалами. Последний тип алгоритмов представляет гораздо больший интерес для исследователей, так как позволяет предсказать участки белка, в которых возможны структурные переходы.

Все аминокислотные остатки можно разделить на три группы: аминокислотные остатки, которые достоверно чаще образуют альфа-спирали – альфа-образователи (аланин, глутамин, глутаминовая кислота, аргинин, лейцин, метионин, лизин); аминокислотные остатки, которые чаще образуют бета-тяжи – бета-образователи (валин, изолейцин, цистеин, фенилаланин, тирозин, триптофан, треонин) и аминокислотные остатки, которые чаще находятся за пределами элементов вторичной структуры, т.е. находятся в неструктурированных участках (так называемый койл) белков и пептидов – пролин, глицин, аспарагиновая кислота, аспарагин, серин, гистидин [1]. Однако перечисленные выше предпочтения для большинства аминокислотных остатков нельзя считать достаточно сильно выраженными: в частности, вероятность того, что остаток лизина войдёт в состав альфа-спирали равна 36%, бета-тяжа – 30%, койла – 34%.

На формирование элементов вторичной структуры влияют ближние и дальние взаимодействия между аминокислотными остатками одной и той же полипептидной цепи [2]. Ближние взаимодействия можно предсказать с высокой долей вероятности, в отличие от дальних [2]. Влияние дальних взаимодействий на образование элементов вторичной структуры при работе с аминокислотной последовательностью не учитывается. Так, в большой прионном белке человека в норме третья и вторая альфа-спирали стабилизируются при помощи контактов друг с другом. При исследовании только первичной структуры прионного белка человека по методам из NPS@Consensus на месте второй альфа-спирали предсказывается бета-тяж (Рис. 1).

```
CC36xx0    CPNITIKQHTVTTTTKGENWTETDVKMRERVREQMC
DPM        cccee eeeeeeeeeccctcccchehehhhhhhhhcc
DSC        cccee eeeeeeeeeeeeeeeeeeeeeeeeeeeeec
GOR1       ettee eeeeeeeeeccccchhhhhhhhhhhhhhhh
GOR3       cccce eeeeeeeeeeeccccchhhhhhchhhhhhhe
GOR4       cccccccceeeeeeccccceeeeeeeeeeeeeeec
HNNC       cccee eeeeeeeeeccccccccchhhhhhhhhhcc
PHD        cccee eeeeceeeeccccccccceeechhhhhhhhecc
Predator   cccccccccccccccccccchhhhhhhhhhhhccc
SIMPA96    cccce eeeceeeeccccccccchhhhhhhhhhhhc
SOPM       cccee eccee eeeeeccccccchhhhhhhhhhhhhh
SOPMA      cccee eccee eeeeeccccccchhhhhhhhhhhhhh
Sec.Cons.  cccee eeeeeeeeeeeccccccchhhhhhhhhhhhhc
```

***Рис. 1*** **–** Предсказанная вторичная структура С-конца второй и N-конца третьей альфа-спиралей по данным алгоритмов из NPS@Consensus

**Цель** настоящего исследования: сравнение результатов предсказания вторичной структуры белков и пептидов с помощью оригинального алгоритма PentaFOLD 2.0 и алгоритмов из NPS@Consensus.

**Задачи** исследования включали: 1. Создать вероятностный алгоритм PentaFOLD 2.0 для предсказания вторичной структуры белков и пептидов; 2. Обработать выборку преимущественно альфа-спиральных, а также выборку преимущественно бета-структурных белков и пептидов человека с помощью разработанного алгоритма; 3. Обработать две названные выше выборки белков и пептидов человека с помощью алгоритмов из NPS@Consensus; 4. Сравнить результаты предсказаний NPS@Consensus и алгоритма PentaFOLD 2.0.

**Материалы и методы.** Материалом для создания алгоритма PentaFOLD 2.0 (http://chemres.bsmu.by) послужила выборка трёхмерных структур бактериальных белков из PDB. Максимальный процент сходства аминокислотных последовательностей этих белков друг с другом не превышал 25% по алгоритму Decrease Redundancy. Общая численность выборки составила 542 белка. На основании анализа аминокислотного состава альфа-спиралей, бета-тяжей и участков

полипептидной цепи, находящихся в неструктурированном состоянии, была получена первая вероятностная шкала, являющаяся по сути обновлённым вариантом шкалы Chou и Fasman [2]. В алгоритме PentaFOLD 2.0 вероятность включения аминокислотного остатка в тот или иной элемент вторичной структуры рассчитывается как среднее значение для пентапептида, в центре которого находится рассматриваемый остаток. Чередования гидрофобных и гидрофильных аминокислотных остатков изучали во фрагментах длиною в пять аминокислотных остатков (в пентапептидах). Вторая вероятностная шкала PentaFOLD основана на частотах использования 32 типов пентапептидов в альфа-спиралях, бета-тяжах и неструктурированных фрагментах. Результаты предсказаний вторичной структуры алгоритм выдаёт по двум паттернам – альфа-спиральному и бета-структурному.

Особенностью обновлённой версии алгоритма является наличие дополнительных дипептидных шкал, содержащих вероятности образования той или иной структуры парой аминокислотных остатков в рамках 1-2, 1-3, 1-4 и 1-5. Поскольку наиболее характерными особенностями обладает койл между двумя бета-тяжами, на первом этапе работы алгоритм сравнивает вероятность его образования с таковой для бета-структуры, расположенной между двумя бета-тяжами. Другими словами, по результатам изучения аминокислотного состава койла между двумя бета-тяжами и бета-тяжей, которые с двух сторон окружены другими бета-тяжами (вне зависимости от длины койла, отделяющего два бета-тяжа), составлена дополнительная вероятностная шкала. По этой шкале алгоритм и определяет вероятность образования койла на протяжении всей аминокислотной последовательности. Если образование неструктурированного участка для данной аминокислоты имеет высокую вероятность (более 0,6), то алгоритм предсказывает койл. Далее, для бета-паттерна алгоритм выбирает наиболее характерные (вероятность выше 0,75) альфа-спиральные пентапептиды (по данным сравнения состава альфа-спиралей, расположенных между двумя бета-тяжами, с таковым для бета-структуры между двух бета-тяжей, и состава альфа-спиралей, расположенных между альфа-спиралью и бета-тяжем в сравнении с аналогичным случаем для бета-структуры), а затем предсказывает наиболее характерные бета-тяжи (вероятность выше 0,6) по результатам сравнения состава бета-тяжей между двумя спиралями с таковым для альфа-спиралей между двумя спиралями. Для альфа-паттерна, наоборот, сначала выбираются наиболее характерные бета-структурные пентапептиды, а потом – с более низким порогом – характерные альфа-спиральные фрагменты. На заключительном этапе по альфа-паттерну как альфа-спиральные предсказываются все остатки, которые являются таковыми хотя бы по одной из двух вероятностных шкал, по бета-паттерну – наоборот.

Для проверки работы алгоритма в данной работе использовались три независимые выборки белков и пептидов человека. Первая выборка состояла из 50 белков и пептидов: альфа-спиральных, бета-структурных и смешанных. Вторая и третья «чистые» выборки состояли из 45 трёхмерных структур каждая и включали только альфа-спиральные или только бета-структурные белки и пептиды человека. Процент сходства между первичными последовательностями во всех

использованных выборках не превышал 25% (по алгоритму Decrease Redundancy), т. е. белки и пептиды в выборках были негомологичными.

Алгоритм для предсказания консенсусной вторичной структуры включает одиннадцать различных вероятностных методов, таких как DPM; DSC; GORI; GORIII; GORIV; HNN; PHD; PREDATOR; SIMPA96; SOPM; SOPMA [3]. После обработки первичной последовательности каждым из алгоритмов формируется консенсусная вторичная структура. Результаты работы алгоритмов сравнивали с описанием вторичной структуры по методу DSSP.

**Результаты и их обсуждение.**

Результаты наших in silico экспериментов показали, что по выборке, состоящей из 50 первичных структур белков и пептидов алгоритмы из NPS@Consensus лучше всего предсказывают неструктурированные участки (чувствительность – 69,31%; специфичность – 76,54%). Несколько хуже предсказанию поддаётся альфа-спираль (чувствительность – 68,02%; специфичность – 59,06%). Бета-тяжи по алгоритмам из NPS@Consensus предсказываются хуже всего: чувствительность – 51,73%; специфичность – 56,41%. Для того, чтобы сравнить алгоритмы из NPS@Consensus между собой по достоверности предсказаний вторичной структуры мы использовали понятие эффективности, которое рассчитывается как среднее арифметическое суммы чувствительности и специфичности конкретного алгоритма. В результате мы выяснили, что среди данных алгоритмов есть те, которые очень хорошо предсказывают альфа-спирали – PHD, PREDATOR, SIMPA96. Эффективность данных методов для альфа-спирали составляет 64,71%, 63,49% и 63,29% соответственно. Алгоритмы, которые очень хорошо предсказывают бета-тяжи – PREDATOR, PHD, SIMPA96. Эффективность данных методов для бета-тяжей составляет 56,98%, 54,39% и 53,61% соответственно. Алгоритмы, которые очень хорошо предсказывают неструктурированные участки белков и пептидов – PREDATOR, SIMPA96, HNN. Эффективность данных методов для койла составляет 76,88%, 73,72% и 72,03% соответственно. Далее мы рассчитали чувствительность и специфичность алгоритмов из NPS@Consensus по «чистым» альфа-спиральной и бета-структурной выборкам. Чувствительность при этом по альфа-спиральной выборке составила 77,13%, специфичность – 81,99%. Чувствительность по альфа-спиральной выборке для койла составила 62,66%, специфичность – 70,21%. Чувствительность для бета-тяжей по бета-структурной выборке равняется 51,28%, специфичность – 76,64%. Чувствительность и специфичность по данной выборке для неструктурированного состояния составила 73,43% и 74,27% соответственно.

Результаты предсказаний алгоритма PentaFOLD 2.0 по альфа-спиральной выборке следующие: по альфа-паттерну – чувствительность составила 65,64%, специфичность – 74,32%; по бета-паттерну – чувствительность составила 28,20%, специфичность – 75,20%. В сравнении с алгоритмами из NPS@Consensus чувствительность и специфичность для альфа-спирали у PentaFOLD 2.0 ниже, соответственно и по эффективности (79,56% и 69,98%) алгоритм PentaFOLD 2.0 уступает алгоритмам из NPS@Consensus. Этот результат можно объяснить тем, что методы из NPS@Consensus настроены на выявление альфа-спиралей, так как

количество остатков в спиралях для альфа-бета белков, как правило, выше, чем в бета-тяжах.

Результаты работы алгоритма PentaFOLD 2.0 для бета-тяжей по бета-структурной выборке следующие: по альфа-паттерну – чувствительность составила 27,16%, специфичность – 70,89%; по бета-паттерну – чувствительность составила 60,12%, специфичность – 63,29%. Как видно, в сравнении с результатами алгоритмов консенсусной вторичной структуры чувствительность PentaFOLD 2.0 выше (60,12% против 51,28%). Специфичность, а также эффективность (63,96% против 61,71%) при этом у PentaFOLD 2.0 ниже. Таким образом, наш алгоритм лучше предсказывает участки первичной структуры белков, где может сформироваться бета-тяж. Это особенно важно при разработке вакцин от заболеваний, в основе которых лежит переход альфа-спирали или неструктурированного участка в бета-тяж с последующим формированием бета-амилоида.

Что касается неструктурированных участков белков и пептидов, то результаты предсказаний алгоритма PentaFOLD 2.0 по альфа- и бета-паттернам отличаются незначительно, но по бета-структурной выборке данные участки белков предсказываются лучше. При обработке альфа-спиральной выборки чувствительность для койла по альфа-паттерну составила 61,33%, специфичность – 60,64%; по бета-паттерну – чувствительность составила 62,15%, специфичность – 59,89%. При обработке бета-структурной выборки чувствительность для койла по альфа-паттерну составила 66,23%, специфичность – 69,87%; по бета-паттерну – чувствительность составила 66,67%, специфичность – 70,08%. Алгоритмы из NPS@Consensus койл предсказывают лучше, чем PentaFOLD 2.0.

Полученные результаты говорят о том, что оригинальный алгоритм PentaFOLD 2.0 способствует выявлению тех участков белков, которые склонны к структурным переходам. Это, в свою очередь, способствует совершенствованию критериев отбора фрагментов белков возбудителей инфекционных болезней, на месте которых, при определённых условиях могут сформироваться бета-тяжи, для включения их в состав коротких вакцинных пептидов.

**Выводы:** 1. Алгоритмы из NPS@Consensus лучше всего предсказывают неструктурированные участки, несколько хуже – альфа-спирали, бета-тяжи предсказываются хуже всего; 2. Среди алгоритмов из NPS@Consensus можно выделить те, которые очень хорошо предсказывают альфа-спирали – PHD, PREDATOR, SIMPA96; алгоритмы, которые очень хорошо предсказывают бета-тяжи – PREDATOR, PHD, SIMPA96; алгоритмы, которые очень хорошо предсказывают неструктурированные участки в белках и пептидах – PREDATOR, SIMPA96, HNN; 3. Чувствительность алгоритма PentaFOLD 2.0 для бета-тяжей выше, чем у методов из NPS@Consensus, т.е. разработанный нами алгоритм лучше предсказывает участки первичной структуры белков и пептидов, где может сформироваться бета-тяж. 4. Как алгоритмы из NPS@Consensus, так и оригинальный алгоритм PentaFOLD 2.0 лучше предсказывают неструктурированные участки по бета-структурной выборке.

**Литература**

1. Chou, P. Y. Prediction of the secondary structure of proteins from their amino acid sequence / P. Y. Chou, G. D. Fasman // Adv. Enzymol. Relat. Areas. Mol. Biol. – 1978. – Vol. 47. – P. 45-48.

2. Хрусталёв, В. В. О вкладе специфических чередований гидрофобных и гидрофильных аминокислотных остатков в формирование элементов вторичной структуры белков / В. В. Хрусталёв, В. В. Побойнев, Т. А. Хрусталёва // Фундаментальная наука в современной медицине: материалы саттел. дистанционной научно-практич. конф. студентов и молодых учёных . – 2016. – С. 312-317.

3. Combet, C. NPS@: Network Protein Sequence Analysis / C. Combet [et al.] // TIBS. – 2000. – Vol. 25. – P.147-150.