ФУНКЦИОНАЛЬНАЯ АННОТАЦИЯ АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ НА ОСНОВЕ ЛОКАЛЬНОГО СХОДСТВА

К.Е. Александров, Б.Н. Соболев, Д.А. Филимонов, В.В. Поройков

ГУ НИИ биомедицинской химии им. В.Н. Ореховича РАМН, Москва, Россия,
e-mail: dzimmu@yandex.ru

Разработан новый метод распознавания функциональных классов белков на основе оригинального способа описания аминокислотной последовательности. Каждая последовательность из обучающей выборки сравнивается с аннотируемой последовательностью, на основе чего вычисляются оценки локального сходства для каждой аминокислотной позиции. Эти оценки используются в качестве входных данных для оригинального классификатора. Метод тестировался на 56 классах белков «Золотого стандарта» («Gold Standard» (Brown *et al*., 2006)), сериновых протеазах и других классах белков. Разработанная нами программа показала высокую точность предсказания – для большинства классов 100 %-я точность. При отнесении белков к классам Международной классификации ферментов (EC) наша программа превосходит по точности программу SVMProt (на основе метода опорных векторов) и сопоставима с программами HMMer (на основе скрытых Марковских моделей) и PROF_PAT (на основе паттернов множественных мотивов). Описанный метод предполагается использовать как для предсказания функциональных классов белков, так и для поиска сайтов функциональной специфичности в аминокислотных последовательностях.

Ключевые слова: функциональная аннотация, аминокислотная последовательность, точность предсказания.

Введение

Функциональная аннотация вновь секвенированных генов представляет собой одну из наиболее важных задач в биоинформатике. Лишь малая часть белков, кодируемых известными нуклеотидными последовательностями, охарактеризована экспериментально. Это обусловливает необходимость развития методов компьютерной аннотации аминокислотных последовательностей. Действительно, в базе данных UniProt содержится более 4 млн аминокислотных последовательностей; база данных Gene Ontology содержит около 160 000 функциональных аннотаций, из которых экспериментально охарактеризовано менее 10 000. Первоначально компьютерная аннотация аминокислотных последовательностей производилась на основе гомологии с экспериментально охарактеризованными белками (Devos, Valencia, 2000, 2001). Этот подход обеспечивает особенно точное предсказание при использовании филогенетических методов, которые зачастую требуют ручной корректировки. Весьма полезны методы, представляющие наборы выровненных последовательностей (белковые семейства) в обобщенном виде с помощью регулярных выражений, позиционных частотных матриц (профилей) и скрытых Марковских моделей. Этот подход позволяет одновременно решать задачи классификации и функционального картирования. Методы этой группы используются в популярных информационных ресурсах: PROSITE (Hofmann *et al*., 1999), BLOCKS (Henikoff S., Henikoff J., 1991; Henikoff *et al*., 1999), PRINTS (Attwood *et al*., 1999), PROF_PAT (Bachinsky *et al*., 1997), PFAM (Finn *et al*., 2008). В связи с необходимостью классификации непрерывно растущего числа последовательностей в последнее время все более популярными становятся автоматизированные методы, основанные на обучении

с использованием выборки экспериментально аннотированных белков без процедуры выравнивания. В этой группе методов используются такие подходы, как наивный Байесовский классификатор, искусственные нейронные сети, метод *k*-ближайших соседей, дерево решений и метод опорных векторов (Han *et al*., 2006). Методы машинного обучения показывают высокую точность распознавания функциональных классов – для ряда классов более 95 %. Однако многие функциональные группы не могут быть предсказаны с необходимой точностью (Cai *et al*., 2003). Поэтому задача предсказания функции белка далека от окончательного решения.

В данной работе мы предлагаем новый метод машинного обучения PAAS (Проекции аминокислотных последовательностей, Projections of Amino Acid Sequences), который основан на новом способе представления аминокислотных последовательностей и оригинальном алгоритме классификации.

Представление аминокислотных последовательностей

Мы предлагаем описывать аннотируемую последовательность A как совокупность оценок локального сходства данной последовательности с последовательностью B, взятой из обучающей выборки. Эта выборка содержит аминокислотные последовательности белков с установленными функциональными характеристиками. Первичные оценки локального сходства рассчитываются с помощью серии сдвигов последовательности A относительно последовательности B (рис. 1). Каждый фрагмент последовательности A (длина которого определяется параметром «frame») сравнивается с каждым совмещенным фрагментом последовательности B.

Первичная оценка R_k для позиции k последовательности A равна мере сходства между фрагментом последовательности A и наиболее сходным с ним фрагментом последовательности B (в простейшем случае – числу совпадающих остатков) согласно следующему выражению:

$$R_k = \max_j (I_{k+F-1,j} - I_{kj}),$$
$$I_{kj} = \sum_{i=1}^{k} s(a_i, b_{i+j}), \quad M_l \le j \le M_r, \tag{1}$$

где R_k – первичная оценка сходства последовательности A с последовательностью B в позиции k, F – значение параметра «frame», $s(a_i, b_{i+j})$ – мера сходства остатков a_i и b_{i+j} из последовательностей A и B, j – величина сдвига. Величина сдвига j ограничена справа и слева величинами M_l и M_r, которые определяются параметром *maxBand*. Если последовательности равны по длине, то $M_l = -maxBand / 2$, а $M_r = maxBand / 2$.

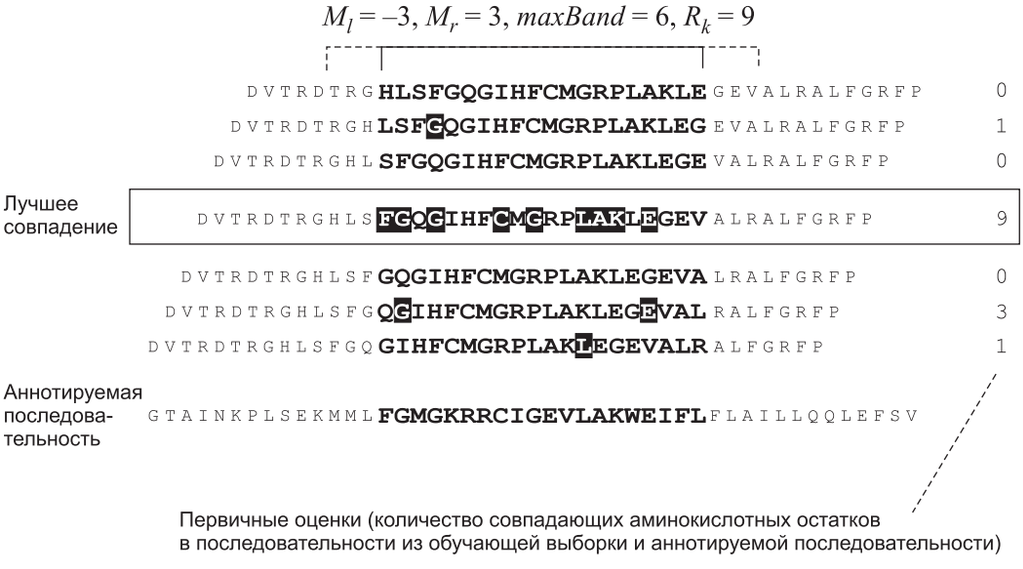

Рис. 1. Расчет первичных оценок локального сходства.

Сглаженная оценка локального сходства S_k для позиции k определяется как максимальная из всех оценок R_k, рассчитанных для всех участков последовательности A, которые включали данную позицию k.

Описанная процедура поиска сходных участков напоминает построение матрицы сходства (McLachlan, 1971). Как известно, парное выравнивание может рассматриваться как отбор таких диагональных фрагментов, которые обеспечивают наилучшую оценку выравнивания. При этом отобранные фрагменты сосредоточены в достаточно узкой полосе. Наш подход позволяет учитывать такие фрагменты, которые могут соответствовать функционально важным участкам, но игнорируются при построении оптимального выравнивания. Выявление отдельных гомологичных фрагментов на матрице сходства, не обязательно совпадающих с общим выравниванием, применялось и ранее (Туманян, Поройков, 1984). Новизна нашего подхода состоит в том, что рассчитанные для всех позиций аннотируемой последовательности оценки локального сходства S_k применяются как входные данные для программы-классификатора.

Оценки локального сходства могут использоваться и для функционального картирования аминокислотных последовательностей. На рис. 2 показано распределение усредненных оценок локального сходства аннотируемой последовательности и последовательностей трех классов. Локальные максимумы на соответствующей кривой позволяют отличить класс, к которому действительно принадлежит аннотируемый белок, от других классов.

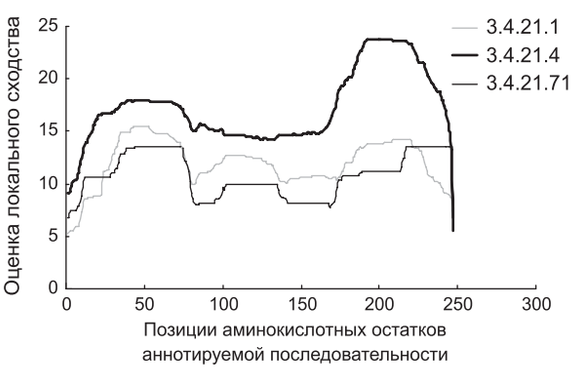

Рис. 2. Распределение усредненных (по классам ЕС) оценок локального сходства.

Утолщенная линия соответствует классу, к которому в действительности относится аннотируемая последовательность.

В данной работе мы использовали простейшую меру сходства аминокислотных остатков: 1 для идентичных остатков и 0 для различающихся (было показано, что использование матрицы замен не приводит к повышению точности предсказания).

Алгоритм классификации

Используемый алгоритм классификации основан на наивном Байесовском классификаторе (Alexandrov *et al*., 2008). Принадлежность аннотируемой последовательности к классу A оценивается с помощью специальной статистики (B-статистики), рассчитанной по следующим формулам:

$$t_0 = \frac{\sum_{k=1}^{N}\left[W_k(A) - W_k(\neg A)\right]}{\sum_{k=1}^{N}\left[W_k(A) + W_k(\neg A)\right]}, \quad (2)$$

$$t_i = \frac{\sum_{k=1}^{N} S_{ik}\left[W_k(A) - W_k(\neg A)\right]}{\sum_{k=1}^{N} S_{ik}\left[W_k(A) + W_k(\neg A)\right]}, \quad (3)$$

$$t = Sin\left[\frac{1}{n}\sum_{i=1}^{n} ArcSin(t_i)\right], \quad (4)$$

$$B = \frac{t - t_0}{1 - tt_0}, \quad (5)$$

где N – количество последовательностей в обучающей выборке, $W_k(A)$ и $W_k(\neg A)$ – веса последовательности k в классе A и в его дополнении $\neg A$, S_{ik} – оценка локального сходства аннотируемой последовательности в позиции i с последовательностью k из обучающей выборки, n – количество аминокислотных остатков в аннотируемой последовательности.

Оценка точности предсказания

Для оценки точности предсказания мы использовали скользящий контроль с исключением по одному. На каждом этапе этой процедуры из обучающей выборки мы удаляли одну последовательность и использовали ее в качестве

аннотируемой. Для каждого класса A на основе полученных значений B-статистики рассчитывался независимый критерий точности прогноза (Independent Accuracy of Prediction, IAP):

$$IAP = \frac{\sum_{i,j} \theta(B_{i \in A} - B_{j \in \neg A})}{N_A \cdot N_{\neg A}}, \qquad (6)$$

где B_i – оценка принадлежности последовательности i классу A, если i действительно принадлежит классу A; B_j – оценка принадлежности последовательности j классу A, если j на самом деле принадлежит его дополнению $\neg A$; $\theta(x) = 1$, если $x > 0$, $\theta(x) = 1/2$, если $x = 0$, $\theta(x) = 0$, если $x < 0$; N_A – количество последовательностей в классе A, $N_{\neg A}$ – количество последовательностей в дополнении A. При 100 %-й точности предсказания значение IAP = 1.

Проверка точности метода

Мы провели проверку точности метода с использованием процедуры скользящего контроля с исключением по одному на двух выборках. Одна из них представляла сериновые протеазы – 566 последовательностей, разделенных на классы согласно EC. Другая выборка содержала 817 последовательностей из так называемого «Золотого стандарта» (Brown *et al*., 2006), представляющих 56 семейств, различающихся по функциональным характеристикам и объединенных в 5 надсемейств. На рис. 3 показаны значения IAP, усредненные по всем группам выборки сериновых протеаз, при разных значениях frame и maxBand.

Результаты проверки точности метода на выборке «Золотого стандарта» схожи с таковыми для сериновых протеаз; 45 из 56 семейств «Золотого стандарта» предсказывались со 100 %-й точностью (рис. 4). Однако наблюдаются различия для семейств и надсемейств. В случае с надсемействами достижение высоких значений IAP происходит при значении frame = 50, в то время как для семейств точность предсказания достигает максимума при меньшем значении этого параметра (30) (рис. 5). По-видимому, это отражает большую дивергенцию после-

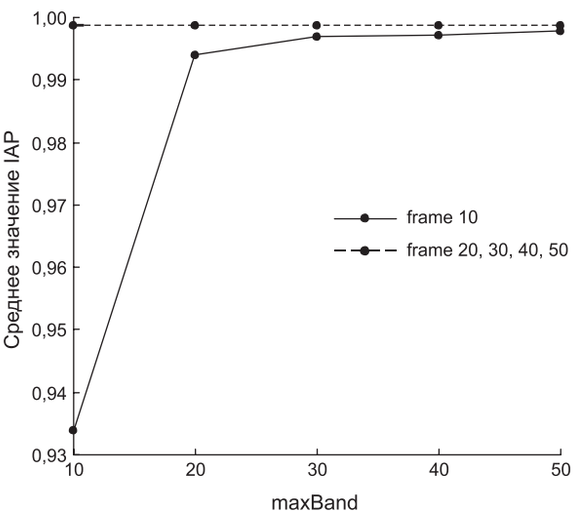

Рис. 4. Проверка точности метода на выборке семейств «Золотого стандарта».

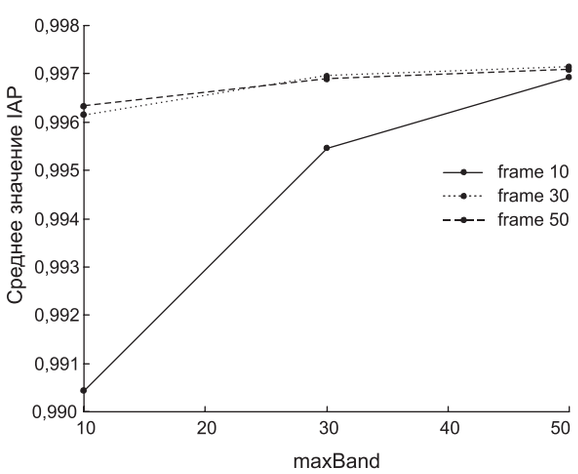

Рис. 3. Проверка точности метода на выборке сериновых протеаз.

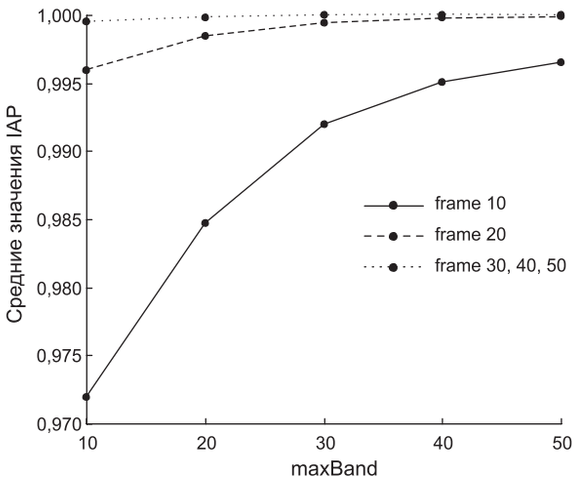

Рис. 5. Проверка точности метода на выборке надсемейств «Золотого стандарта».

довательностей, входящих в надсемейства по сравнению с семействами.

Сравнение с другими методами

Для сравнения точности предсказания функциональных классов с помощью нашего метода мы выбрали программы, основанные на известных методах: HMMer, SVMProt и PROF_PAT.

Программа SVMProt (Cai *et al*., 2003) реализует один из методов машинного обучения (метод опорных векторов) и так же, как PAAS, работает с обучающей выборкой из невыровненных последовательностей. Сравнение предложенного метода с методом опорных векторов проводилось на основе данных о чувствительности и специфичности предсказания для 46 различных классов белков, приведенных авторами программы SVMProt. На основании этих данных мы рассчитали значения IAP.

$$IAP = \frac{\frac{TP}{TP + FN} + \frac{TN}{FP + TN}}{2}, \qquad (7)$$

где TP – число истинно положительных, TN – число истинно отрицательных, FP – число ложноположительных и FN – число ложноотрицательных результатов.

Для расчета значений IAP при предсказании тех же 46 классов нашим методом мы применили процедуру скользящего контроля с исключением по одному. Расчеты производились при значении frame = 50 и различных значениях maxBand. Источником последовательностей для обучающей выборки была база данных Brenda (http://www.brenda.unikoeln.de).

Средняя точность предсказания с помощью нашего метода сравнима с точностью предсказания программы SVMProt и даже превышает ее при значении maxBand = 5000. При этом значении наш метод предсказывает 32 класса с более высоким IAP и 14 классов с более низким IAP, чем SVMProt. Стоит отметить, что с помощью PAAS 4 класса белков были предсказаны со 100 %-й точностью. Однако есть классы, для которых точность предсказания с помощью PAAS относительно низка. Например для класса EC 4.4 величина IAP составляет 0,5971. В данном случае низкая точность прогноза объясняется тем, что данный класс обучающей выборки состоит из 7 белков, относящихся к 5 разным белковым семействам (согласно классификации PFAM), так что 4 белка из 7 относятся к 4 различным классам. По-видимому, высокая точность прогноза наблюдается для тех функциональных классов, которые включают одну или несколько хорошо представленных групп гомологичных последовательностей.

Данные сравнения нашего метода с SVMProt приведены в табл. 1.

Программа HMMer (http://hmmer.janelia.org/) использует скрытые Марковские модели. Для оценки этого метода из «Золотого стандарта» были отобраны 5 надсемейств и 8 семейств, обладающих номерами EC. Представительность отобранных классов составляла от 7 до 215 последовательностей. Используя программу ClustalW со стандартными настройками, мы получили выравнивания для каждого класса; «ручная» коррекция выравниваний не проводилась. На основе полученных выравниваний были построены скрытые Марковские модели для соответствующих классов, а затем проводилось сопоставление всех отобранных последовательностей «Золотого стандарта» с построенными моделями. На основе выходных данных программы HMMer для каждого класса можно было определить пороговое значение: все последовательности, действительно относящиеся к данному классу, получали оценки выше порога, а все неотносящееся – ниже порога. Можно сказать, что HMMer продемонстрировал 100 %-ую точность распознавания отобранных классов. В табл. 2 приведены оценки точности предсказания (на основе рассчитанных значений IAP) для тех же классов, полученные с помощью нашего метода (frame = 50 и maxBand = 50) при проведении скользящего контроля с исключением по одному.

Таким образом, программная реализация предложенного метода уступает программе HMMer только в 1 из 13 классов.

Программа PROF_PAT (http://wwwmgs.bionet.nsc.ru/mgs/programs/prof_pat/) использует паттерны белковых семейств, создаваемые в автоматическом режиме на основе выровненных последовательностей из баз данных SwissProt и TrEMBL. База данных PROF_PAT содержит паттерны более чем 13 000 белковых семейств. Для ориентировочного сопоставления точности данного метода с точностью PAAS мы исполь-

Таблица 1

Сравнение PAAS и SVMProt*

Функциональный класс	IAP для SVMProt	IAP для PAAS	ΔIAP
EC 1.1	0,9164	0,9746	**0,0582**
EC 1.2	0,9195	0,9719	**0,0524**
EC 1.3	0,8397	0,9273	**0,0876**
EC 1.4	0,881	0,9542	**0,0732**
EC 1.5	0,7469	0,8766	**0,1297**
EC 1.6	0,9469	0,7756	–0,1713
EC 1.8	0,8246	0,8755	**0,0509**
EC 1.9	0,976	0,7673	–0,2087
EC 1.10	0,8438	0,6175	–0,2263
EC 1.11	0,9161	0,9942	**0,0781**
EC 1.13	0,9087	0,8195	–0,0892
EC 1.14	0,9141	0,9682	**0,0541**
EC 1.15	0,9426	0,8597	–0,0829
EC 1.17	0,8956	0,5836	–0,312
EC 2.1	0,8581	0,8209	–0,0372
EC 2.2	0,9179	0,8652	–0,0527
EC 2.3	0,8947	0,942	**0,0473**
EC 2.4	0,8722	0,9399	**0,0677**
EC 2.5	0,8746	0,9081	**0,0335**
EC 2.6	0,8885	0,9913	**0,1028**
EC 2.7	0,8259	0,95	**0,1241**
EC 2.8	0,7869	0,984	**0,1971**
EC 3.1	0,5747	0,9505	**0,3758**
EC 3.2	0,897	0,9675	**0,0705**
EC 3.3	0,9375	0,8125	–0,125
EC 3.4	0,8717	0,9623	**0,0906**
EC 3.5	0,8355	0,8921	**0,0566**
EC 3.6	0,9268	0,9623	**0,0355**
EC 4.1	0,8959	0,8463	–0,0496
EC 4.2	0,8632	0,9129	**0,0497**
EC 4.3	0,9196	0,7773	–0,1423
EC 4.4	0,7495	0,5971	–0,1524
EC 4.6	0,7604	1	**0,2396**
EC 5.1	0,8362	0,802	–0,0342
EC 5.2	0,8255	0,969	**0,1435**
EC 5.3	0,9293	0,7567	–0,1726
EC 5.4	0,8537	0,9479	**0,0942**
EC 6.1	0,9406	0,9836	**0,043**
EC 6.2	0,8717	0,9985	**0,1268**
EC 6.3	0,8961	0,9124	**0,0163**
EC 6.4	0,9389	1	**0,0611**
EC 6.5	0,8446	1	**0,1554**

Окончание таблицы 1

Функциональный класс	IAP для SVMProt	IAP для PAAS	ΔIAP
Рецепторы, сопряженные с G-белком	0,9616	0,9977	**0,0361**
Ядерные рецепторы	0,9352	1	**0,0648**
Тирозинкиназные рецепторы	0,8562	0,955	**0,0988**
Управляемые электрохимическим потенциалом транспортеры (симпортеры, унипортеры, антипортеры)	0,923	0,9872	**0,0642**

*IAP для SVMProt – значения IAP, рассчитанные по данным, приведенным в статье Cai *et al*. (2003); IAP для PAAS – значения IAP для результатов предсказания, полученных с помощью метода PAAS при значениях frame = 50 и maxBand = 5000; ΔIAP – разность между значениями IAP для SVMProt и PAAS (в тех случаях, когда точность PAAS выше точности SVMProt, т. е. ΔIAP > 0, значения выделены).

Таблица 2

Сравнение точности предсказания PAAS и HMMer

Функциональный класс	PAAS	HMMer
Амидогидролазы (sf)	++	++
Кротоназы (sf)	++	++
Енолазы (sf)	++	++
VOC (sf)	++	++
Галоацидные дегалогеназы (sf)	+	++
Гистоновая ацетилтрансфераза(f)	++	++
Енолаза (f)	++	++
АМФ-деаминаза (f)	++	++
D–гидантоиназа (f)	++	++
Дигидрооротаза 2 (f)	++	++
Гуаниновая деаминаза (f)	++	++
АТФаза р-типа (f)	++	++
Уреаза (f)	++	++

«++» – точность предсказания 100 %; «+» – точность предсказания > 98 %; «f» – семейства, «sf» – надсемейства.

зовали часть последовательностей «Золотого стандарта», для которых известен номер ЕС. Это условие было введено для облегчения сопоставления результатов предсказания с помощью программы PROF_PAT. Были установлены следующие значения параметров поиска – матрица аминокислотных замен Blosum62 и 100 %-й уровень сходства (similarity level). Часть семейств «Золотого стандарта» не имеет номеров ЕС, а часть результатов предсказания с помощью программы PROF_PAT представлены в виде названий функциональных классов, а не в виде номеров ЕС. С помощью программы PROF_PAT для 89,9 % последовательностей «Золотого стандарта» были предсказаны номера ЕС, которые действительно относятся к соответствующим белкам. Таким образом, эффективность программы PAAS сопоставима с эффективностью метода PROF_PAT.

Выводы

Предложенный подход показал высокую точность предсказания при различных уровнях функциональной классификации белков, сопоставимую с точностью трех широко используемых методов функциональной аннотации. Метод обладает рядом преимуществ: а) формализация данных о локальном сходстве последовательностей в сочетании с использованием оригинального классификатора обеспечивает высокую точность предсказания; б) предоставляется возможность как классифицировать новые аминокислотные последовательности, так и строить их функциональные карты; в) предложенный алгоритм может быть реализован в виде автоматизированного инструмента функциональной аннотации белков; г) программная реализация метода требует небольших вычислительных ресурсов, что позволяет проводить массовую аннотацию последовательностей.

Литература

Туманян В.Г., Поройков В.В. Установление оптимального соответствия между аминокислотными (нуклеотидными) последовательностями // Биофизика. 1984. Т. 24. № 6. С. 917–920.

Alexandrov K., Sobolev B., Filimonov D., Poroikov V. Recognition of protein function using the local similarity // J. Bioinf. Comp. Biol. 2008. V. 6. № 4. P. 709–725.

Attwood T.K., Flower D.R., Lewis A.P. *et al*. PRINTS prepares for the new millennium // Nucl. Acids Res. 1999. V. 27. P. 220–225.

Bachinsky A.G., Yarigin A.A., Guseva E.H. *et al*. A bank of protein family patterns for rapid identification of possible functions of amino acid sequences // Comput. Appl. Biosci. 1997. V. 13. № 2. P. 115–122.

Brown S.D., Gerlt J.A., Seffernick J.L., Babbitt P.C. A gold standard set of mechanistically diverse enzyme superfamilies // Genome Biol. 2006. V. 7. № 1. R8.

Cai C.Z., Han L.Y., Ji Z.L. *et al*. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence // Nucl. Acids Res. 2003. V. 31. P. 3692–3697.

Devos D., Valencia A. Intrinsic errors in genome annotation // Trends Genet. 2001. V. 17. № 8. P. 429–431.

Devos D., Valencia A. Practical limits of function prediction // Proteins. 2000. V. 41. P. 98–107.

Finn R.D., Tate J., Mistry J. *et al*. The Pfam protein families database // Nucl. Acids Res. 2008. V. 36 (Database issue):D281-8. Epub 2007. Nov 26.

Han L., Cui J., Lin H. *et al*. Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity // Proteomics. 2006. V. 6. № 14. P. 4023–4037.

Henikoff S., Henikoff J.G. Automated assembly of protein blocks for database searching // Nucl. Acids Res. 1991. V. 19. P. 6565–6572.

Henikoff J.G., Henikoff S., Pietrovski S. New features of the Blocks Database servers // Nucl. Acids Res. 1999. V. 27. P. 226–228.

Hoffmann K., Bucher P., Falquet L., Bairoch A. The PROSITE database, in status in 1999 // Nucl. Acids Res. 1999. V. 27. P. 215–219.

McLachlan A.D. Test for comparing related amino acid sequences: Cytochrome C and cytochrome C551 // J. Mol. Biol. 1971. V. 61. P. 409–424.

FUNCTIONAL ANNOTATION OF THE AMINO ACID SEQUENCES USING LOCAL SIMILARITY

K. Alexandrov, B. Sobolev, D. Filimonov, V. Poroikov

Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow, Russia,
e-mail: dzimmu@yandex.ru

Summary

We have developed a new method, which enables to recognize the protein functional classes based on the original description of an amino acid sequence. Each sequence of the training set is compared with the annotated sequence and local similarity scores for all amino acid positions are calculated. These scores are used as input data for the original classifier. The method was tested on 56 classes of proteins included into the Gold Standard (Brown *et al*., 2006), serine proteases and other protein classes. Protein classes mentioned above were predicted with high accuracy; most of them were predicted with 100 % accuracy. Our program predicted Enzyme classification classes with the accuracy superior to SVMProt program (based on the support vector machine) and comparable with HMMer (based on hidden Markov models) and PROF_PAT (based on the multiple motif patterns). We suppose that the suggested method can be used in the prediction of the functional classes of proteins and in the revealing of functional specificity sites in the amino acid sequences.