ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ
(Вычислительные системы)
2002 год Выпуск 171

УДК 519.95+577.2

РАСПОЗНАВАНИЕ САЙТОВ РАЗРЕЗАНИЯ В АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ[1]

Н.Г.Загоруйко[2] О.А.Кутненко, В.А.Иванисенко [3], С.В.Николаев

Введение

Автоматическое распознавание сигнальных пептидов и сайтов разрезания в белках является актуальной задачей как для распознавания их внутриклеточной локализации, так и для решения прикладных задач в медицине и биотехнологии. Существующие методы распознавания используют матрицы аминокислотных замен [1] или нейросетевые алгоритмы, работающие с двадцатибуквенным аминокислотным кодом [2].

В работе исследована возможность привлечения физико-химических характеристик аминокислот для распознавания сайтов разрезания. Для решения задачи применены методы интеллектуального анализа данных (Data Mining), используемые для автоматического обнаружения скрытых эмпирических закономерностей при распознавании образов [3]. Особенность данных методов состоит в их ориентации на решение задач, для которых

[1]Работа частично поддержана РФФИ, проекты 02-01-00082 и 01-07-90376, а также Интеграционным проектом СО РАН (2002-65).

[2]e-mail: zag@math.nsc.ru

[3]e-mail: salix@bionet.nsc.ru

использование традиционных статистических методов вызывает большие затруднения из–за их специфики. Указанная задача из области биоинформатики относится именно к этому типу задач в силу как большого объема анализируемых данных, так и отсутствия оснований для выдвижения гипотез о законах распределения. Для решения задачи разработан метод выбора информативного подмножества характеристик и построено решающее правило, основанное на использовании функции принадлежности.

Постановка задачи

Данные представлены наборами фрагментов последовательностей белков эукариот, грамм–положительных и грамм–отрицательных бактерий. Кроме того, отдельно рассматривались человеческие белки и белки Ecoli. Наборы данных подразделялись на поднаборы, содержащие сайты разрезания (образ «Сигнальный пептид»), якоря (образ «Якорь»). Сигнальные якоря по физико–химическим свойствам близки к сигнальным пептидам, но в отличии от последних не отрезаются сигнальными пептидазами [4]. В качестве негативной выборки использовались фрагменты ядерных и цитоплазматических белков, не содержащих ни сайтов, ни якорей (образ «Негатив»). Все данные были взяты с веб–сайта[4].

Фрагменты белка (протеины или аминокислотные последовательности) представлены текстами различной длины — от нескольких десятков до нескольких тысяч символов. Алфавит символов содержит 20 элементов $\{A, C, D$ и т. д.$\}$, где символ обозначает определенную аминокислоту. Каждая аминокислота описывается своим набором физико–химических свойств.

При исследовании использовались два набора характеристик аминокислот: 1) набор из 10–и свойств Kidera [5], т.е. некоррелируемых между собой линейных комбинаций из структурных и физико–химических свойств аминокислот и 2) 434 структурных и физико–химических свойства, взятые из базы данных[5].

[4] http://www.cbs.dtu.dk./services/SignalP/

[5] http://www.genome.ad.jp/dbget/aaindex.html

Задача состоит в распознавании принадлежности некоторого фрагмента аминокислотной последовательности к одному из этих трех образов и указании наиболее вероятного места локализации сайта разрезания (якоря) для фрагментов образов «Сигнальный пептид» («Якорь»).

Методы и алгоритмы

Формирование обучающей выборки. Окно анализа. На первом этапе решения задачи изучалось качество распознавания сайта разрезания в зависимости от ширины окна анализа, т.е. от количества рассматриваемых символов слева и справа от точки разрезания. В качестве признаков был использован набор свойств Kidera. Ширина окна анализа менялась от 6 до 36 символов.

Данные (фрагменты эукариотических белков) были разделены на два образа: «Сигнальный пептид» и «Негатив». Каждый образ был представлен выборкой из $N_1 = N_2 = 506$ последовательностей. При формировании обучающей выборки первого образа («Сигнальный пептид») окно шириной L символов размещалось на протеиновой цепочке так, чтобы сайт разрезания совпадал с центром окна. Каждому элементу окна анализа соответствует вектор 10 свойств Kidera, так что набору символов в окне соответствует вектор размерностью $L \times 10$. В итоге обучающая выборка для первого образа представляет собой таблицу размерностью в N_1 строк (объектов) и $L \times 10$ столбцов (признаков). Второй образ («Негатив») был представлен цепочками из L аминокислот, попавших в окно на случайно выбранном участке каждого из N_2 белков, не содержащих сайтов разрезания. В результате был сформирован массив данных из представителей двух образов в виде таблицы размером в $N = N_1 + N_2$ строк и $L \times 10$ столбцов.

Учитывая ограниченность объема данных и результаты предварительных экспериментов, была выбрана ширина окна анализа $L = 18$ символов, и для этой ширины проводились дальнейшие исследования.

Выбор решающего правила. На данном этапе все N объектов использовались для синтеза классификатора (выбора подпространства информативных признаков) и для последующей

классификации (распознавания). Распознавание проводилось последовательно для каждого из N объектов первого и второго образов. Очевидно, что данный подход дает заниженную оценку вероятности ошибки.

В качестве решающего правила использовалось правило *"k ближайших соседей"*. В соответствии с этим правилом при классификации неизвестного объекта X среди объектов первого и второго классов находим k ближайших к точке X объектов. Пусть k_1 и k_2 — соответственно число объектов из первого и второго классов среди этих k ближайших соседей. Если $k_1 \geq k_2$, то объект X принадлежит первому классу, в противном случае X принадлежит второму классу. Сравнивались решения при разных значениях параметра k — от 1 до 1001. Было обнаружено (см. табл. 1), что с ростом k меняется как общее количество ошибок (P), так и соотношение между ошибками первого рода (P_1, пропуск сайта) и ошибками второго рода (P_2, ложная тревога). При увеличении k от 1 до 101 растет общее количество ошибок и количество ложных тревог, но уменьшается число пропущенных сайтов разрезания. Решение о том, на каком значении k нужно остановиться, зависит от соотношения стоимости ошибок I и II рода.

Таблица 1

Изменение числа ошибок в зависимости от параметра k

k	P	%ошибок	P_1	P_2
1	125	12.4	31	94
5	123	12.2	27	96
7	127	12.6	28	99
21	125	12.4	27	98
41	128	12.7	22	106
101	137	13.6	18	119
201	133	13.2	19	114
301	132	13.1	22	110
401	136	13.5	26	110
601	137	13.6	35	102
1001	138	13.7	53	85

Для сокращения машинного времени и предполагая стоимость ошибок одинаковой, в данных исследованиях было решено положить $k = 1$, т.е. использовать правило ближайшего соседа.

Проверка гипотезы "нечетности". Эта гипотеза опирается на геометрическое представление о пространственной форме белка и утверждает, что аминокислоты, следующие друг за другом через один символ и имеющие одинаковую ориентацию, могут совместно участвовать в определенных биохимических процессах. Следовательно, при фиксированном положении окна и фиксированной нумерации символов в окне интересующая нас функция разрезания может по-разному проявляться на четных и нечетных подпоследовательностях аминокислот.

Полагаем, что сайт разрезания совпадает с центром окна анализа или нулевым элементом последовательности символов, попавших в окно анализа. Тогда для окна шириной $L = 18$ символы нумеровались слева направо следующим образом: $-8,-7,\ldots,0,\ldots,8,9$. Исследовалась зависимость результатов распознавания от того, какие из попавших в окно символов используются для распознавания. Рассматривались три варианта: 1) все 18 символов; 2) только 9 символов с четными номерами $(-8,-6,\ldots,6,8)$; 3) только 9 символов с нечетными номерами $(-7,-5,\ldots,7,9)$. В ходе экспериментов выяснилось, что более предпочтительно использовать только элементы окна анализа с нечетными номерами (табл. 2).

Таблица 2

Надежность распознавания в зависимости от набора символов

Набор символов	Надежность распознавания (%)
Все символы	83.8
9 четных символов	80.4
9 нечетных символов	85.4

Выбор признаков. При ширине окна $L = 18$ из 10 признаков набора Kidera методом полного перебора было выбрано 7 наиболее информативных признаков (исключены признаки с номерами 3, 5 и 10). Качество распознавания при изменении полученно-

го набора признаков снижалось. Надежность распознавания на обучающей выборке двух образов была равна 87,6%.

Далее информативный набор признаков выбирался непосредственно из всех имеющихся в распоряжении 434 физико-химических характеристик аминокислот плюс 10 свойств набора Kidera. Предварительно все характеристики нормировались по дисперсии. Для поиска наиболее информативного подмножества признаков использовался алгоритм AddDel [3], который сочетает в себе идеи методов "последовательного добавления наиболее ценных" (Addition) и "последовательного удаления наименее ценных" (Deletion) признаков. Вначале информативность всех признаков оценивается по отдельности и выбирается самый информативный признак. Затем к нему по очереди подбирается такой второй признак, сочетание с которым оказывается наиболее информативным. Процесс добавления повторяется n_Add раз. Затем каждый из отобранных признаков по очереди удаляется из этого набора и определяется такой признак, удаление которого в наименьшей степени ухудшает качество выбранной подсистемы. Таким способом за n_Del шагов ($n\_Del < n\_Add$) удаляется n_Del признаков. После этого процессы добавления и удаления признаков повторяются. В данных экспериментах использовалась тактика "два шага вперед, один шаг назад" ($n\_Add = 2, n\_Del = 1$).

Сравнительный анализ различных (по мощности и качеству распознавания на этапе обучения) наборов информативных признаков показал, что признаки, определяющие существенные различия между образами, попадают в искомый набор в первую очередь, а признаки, определяющие более тонкие различия между образами на этапе обучения, значительно увеличивают длину набора, и эти последние признаки на этапе контрольного распознавания выступают как шум, т.е. оказывают отрицательное (маскирующее) воздействие на существенные отличия. В результате численного моделирования установлено, что оптимальное (для распознавания) число информативных признаков лежит в интервале от 2 до 7. На рис. 1 показано изменение качества распознавания (при обучении) в зависимости от количества выбранных признаков.

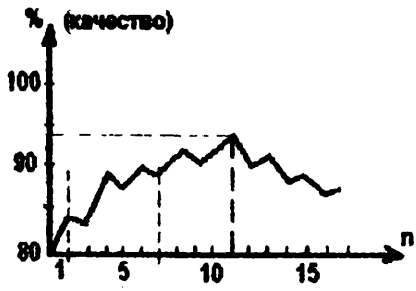

Рис. 1

Функция принадлежности. В реальных условиях интерес представляет не только решение о принадлежности контрольного объекта образу «Сигнальный пептид» или «Негатив», но и оценка вероятности наличия сайта разрезания в каждом окне наблюдения. Такую оценку можно получить с помощью *функции принадлежности* [3]. Для каждого контрольного объекта X найдем расстояния r_1 и r_2 до двух ближайших соседей — по одному из каждого образа. *Функция принадлежности* к некоторому образу задается в виде $f = 1 - \frac{2r_1}{r_1 + r_2}$, значение функции f меняется в пределах от -1 до $+1$. Если $f \geq 0$, то объект X принадлежит первому образу, в противном случае — второму образу.

Коллективные решения. При проведении экспериментов было обнаружено, что результаты распознавания зависят от того, как сформирована обучающая выборка образа «Негатив» (из-за ограниченности данных для образа «Сигнальный пептид» обучающая выборка этого образа была одной и той же). Было решено исследовать устойчивость получаемых результатов по отношению к этому фактору. Было сформировано 7 обучающих выборок, состоящих из одного и того же набора элементов первого образа и семи разных наборов элементов второго образа. И для каждой из семи выборок решалась задача поиска наиболее информативного набора признаков.

Распознавание контрольного объекта X проводилось по этим 7 наборам признаков параллельно. Решение в пользу первого

или второго образа принималось по большинству из 7 голосов. Решение о принадлежности контрольного объекта принималось также и в зависимости от общего порогового значения функции принадлежности $F = \frac{\sum_1^I f_i}{I}$, где f_i — функция принадлежности для i-го набора информативных признаков, $i = 1, \ldots, I$; I — количество наборов. Эксперименты показали, что при ответе на вопрос о принадлежности объекта одному из двух образов, данные методы принятия решений дают близкие результаты. Поэтому, учитывая бóльшую информативность решения по значению функции принадлежности, в дальнейшем решение о принадлежности контрольного объекта принималось данным образом.

Распознавание контрольной последовательности. Для контрольного распознавания предъявлялись данные, не участвовавшие в обучении. Окно анализа двигалось вдоль аминокислотной последовательности слева направо со сдвигом в один символ. И для каждого окна принималось решение о том, есть здесь сайт разрезания или нет. На рис. 2, 3 приведены примеры обработки отдельных фрагментов белков: на рис. 2 представлен график функции принадлежности для фрагмента, содержащего сайт разрезания в позиции 0, на рис. 3 — то же для фрагмента, не содержащего сайта разрезания.

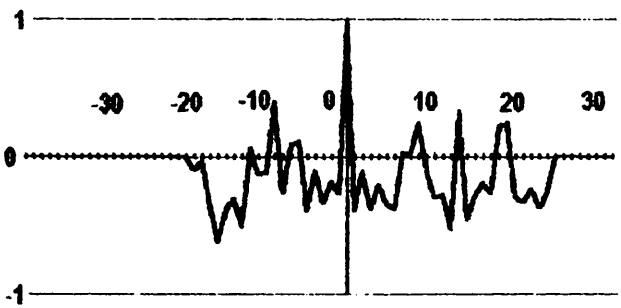

Рис. 2.

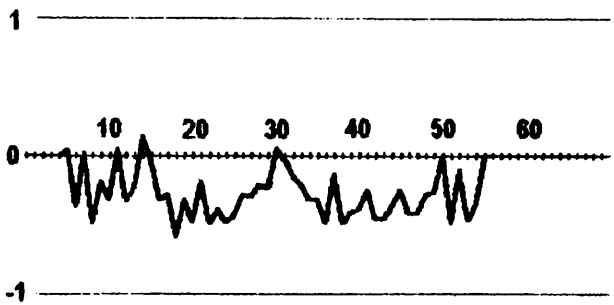

Рис. 3.

Использование дополнительной информации. При распознавании помимо физико-химических свойств аминокислот использовалась также информация о частоте встречаемости аминокислот в позициях -3, -1 окна анализа [2].

Пусть в позициях −3, −1 окна анализа (контрольного объекта) записана некоторая пара аминокислот. Обозначим через α_i частоту встречаемости данной пары аминокислот в позициях −3, −1 обучающих реализаций i-го образа; через $g = k(\alpha_i - \alpha_j)$ — "вклад" информации о частоте встречаемости данной пары аминокислот в принятие решения о принадлежности контрольного объекта образу i или j. Если $g > 1$, то положим $g = 1$ (если $g < -1$, то положим $g = -1$). Определим функцию принадлежности F^* следующим образом: $F^* = \dfrac{F+g}{2}$. Весовой коэффициент $k = 0.5$ получен в результате численного моделирования. Отметим, что из-за ограниченности объема данных предложенное правило требует уточнения.

Распознавание трех образов. Разработанная техника была применена для решения задачи распознавания трех образов: «Сигнальный пептид», «Якорь», «Негатив» (фрагменты цитоплазматических и ядерных белков). Сигнальный пептид распознавался по наличию сайта разрезания. Сигнальный

якорь — по С–терминальной границе мембранного участка белка. Для распознавания этих трех образов использовался алгоритм попарного сравнения. Обозначим указанные выше образы через I_j, $j = 1,2,3$. Для каждой пары образов (I_i, I_j), $i \neq j$, $i = 1,2,3$, $j = 1,2,3$, было сформировано свое "компетентное" признаковое пространство, в котором эти два образа максимально отличаются друг от друга. Сравнивая распознаваемую реализацию X с эталонами образов I_i и I_j (в данном случае в качестве эталонов используются все обучающие реализации сравниваемых образов), мы определяем, на какой из этих двух образов более похожа реализация X в подпространстве, оптимальном именно для этого случая. Будем говорить, что образ I_i является "победителем" при сравнении образов I_i и I_j, если функция принадлежности для данной пары $f_{ij} \geq 0$, в противном случае "победителем" является образ I_j. Если в двух парных сравнениях (I_i, I_j) и (I_i, I_k), $i \neq j \neq k$, $i,j,k \in \{1,2,3\}$, "победителем" оказывается один и тот же образ I_i, то "победитель" среди проигравших образов I_j и I_k не может победить образ I_i, таким образом, сравнение между проигравшими образами не проводится. В случае если при парных сравнениях получены три "победителя" — I_1, I_2, I_3, что не исключено, так как расстояния и соответственно функции принадлежности, вычисляются в разных признаковых пространствах, то решение о принадлежности контрольного объекта X к одному из трех образов принимается, исходя из максимального абсолютного значения функции принадлежности $f_{ij}, i \neq j$, $i,j \in \{1,2,3\}$.

Результаты экспериментов и их обсуждение

Был проведен сравнительный анализ результатов при использовании окон разной ширины и конфигурации. Оказалось, что наилучшие результаты получаются при окне, в котором используются четыре нечетных символа — по два с каждой стороны от центра окна. Использование при распознавании только нечетных позиций окна анализа согласуется с известным правилом "–1, –3" [6].

На рис. 4–5 приведены графики функции принадлежности, усредненной по 252 контрольным белкам, содержащим сайт разрезания. Фрагменты белков позиционированы по точке локализации сайта (вертикальная линия). На рис. 6–7 — то же для 252 случайно выбранных фрагментов контрольных белков без сайтов разрезания. Приведенные на данных рисунках результаты получены при ширине окна анализа $L = 18$ (рис. 4,6) и $L = 8$ (рис. 5,7).

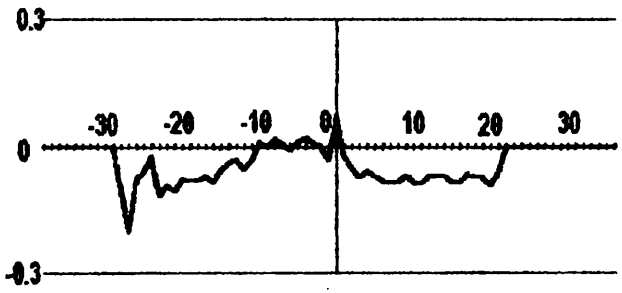

Рис. 4

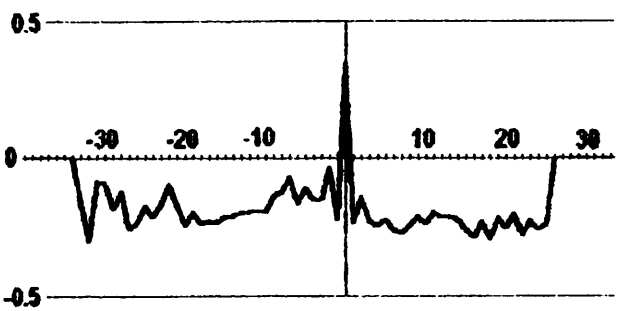

Рис. 5

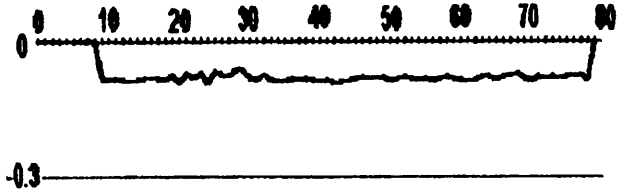

Рис. 6

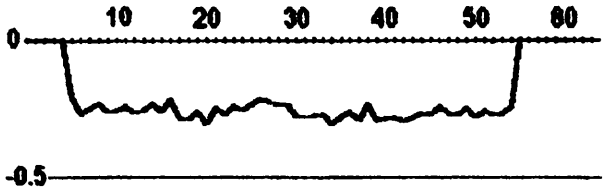

Рис. 7

Из рисунков видно, что сайт разрезания в среднем хорошо обнаруживается и локализуется, при этом результаты распознавания при ширине окна анализа $L = 8$ существенно лучше, чем при ширине окна $L = 18$.

Приведенные ниже (см. приложения 1 и 2) результаты экспериментов получены при ширине окна $L = 8$. В приложении 1 приведены результаты распознавания двух образов из имеющихся трех: «Сигнальный пептид», «Якорь», «Негатив», для разных белков. В приложении 2 — результаты распознавания трех указанных выше образов для белков эукариот и человека.

Проведенные эксперименты показали, что учет дополнительной информации о частоте встречаемости аминокислот в позициях -3, -1 окна анализа повышает качество распознавания. Полученные результаты по распознаванию трех образов демонстрируют удовлетворительную дискриминирующую способность между сигнальными пептидами и сигнальными якорями. Тестирование экспериментальных данных большого объема показало, что сайты разрезания правильно обнаруживаются и локализуются как минимум в 85% случаев.

Заключение

В работе продемонстрировано применение методов интеллектуального анализа данных и обнаружения скрытых эмпирических закономерностей для решения задачи биоинформатики — распознавания сайтов разрезания сигнальных пептидов. Полученные результаты показывают, что использование описанной методики позволяет получить высокую надежность обнаружения и локализации сайтов разрезания.

Литература

1. von HEIJNE G. A new method for predicting signal sequence cleavage sites //Nucleic Acids Res. — 1986. — Vol. 14. — P. 4683–4690.

2. NIELSEN H., ENGELBRECHT J., BRUNAK S., VON HEIJNE G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites //Protein Engineering. — 1997. — Vol. 10, № 1. P. 1–6.

3. ЗАГОРУЙКО Н.Г. Прикладные методы анализа данных и знаний. — Новосибирск: Изд. ИМ. 1999. — 270 с.

4. SAKAGUCHI M., TOMIYOSHI R., KUROIWA T., MIHARA K., OMURA T. Functions of Signal and Signal-Anchor Sequences are Determined by the Balance Between the Hydrophobic Segment and the N-Terminal Charge// PNAS. — 1992. — Vol. 89. — P. 16-19.

5. KIDERA A., KONISHI Y., OKA M., OOI T., SCHERAGA H.A. Statistical analysis of the physical properties of the 20 naturally occuring amino-acids //J. Prot. Chem. — 1985. — Vol. 4. — P. 23–55.

6. von HEIJNE G. Life and death of a signal peptide// Nature. — 1998. — Vol. 396. — P. 112–113.

Поступила в редакцию
12 сентября 2002

ПРИЛОЖЕНИЕ 1

На рис. 8–9 представлены результаты экспериментов по распознаванию сигнальных пептидов в контрольных фрагментах белков Ecoli, как содержащих сайты разрезания (52 фрагмента), так и без сайтов разрезания (66 фрагментов). Было выделено 6389 объектов ширины $L = 8$: 52 объекта образа «Сигнальный пептид» и 6337 объектов образа «Негатив». На рис. 8 приведены графики ошибок I–го и II–го рода в зависимости от порогового значения функции принадлежности F: с учетом только физико–химических свойств аминокислот (а), и с дополнительным использованием информации о частоте встречаемости аминокислот (б). На рис. 9 показано изменение ошибки II–го рода в зависимости от ошибки I–го рода: кривая 1 — учитывались только физико-химические свойства аминокислот, кривая 2 — дополнительно использовалась информация о частоте встречаемости аминокислот.

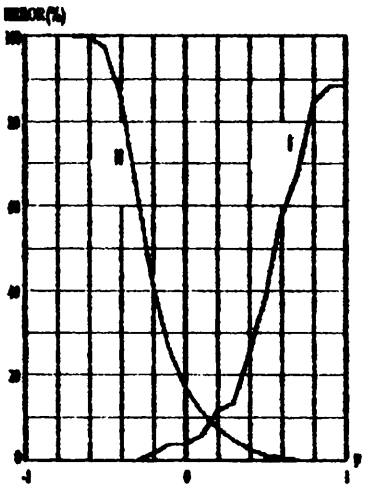

Рис. 8,а

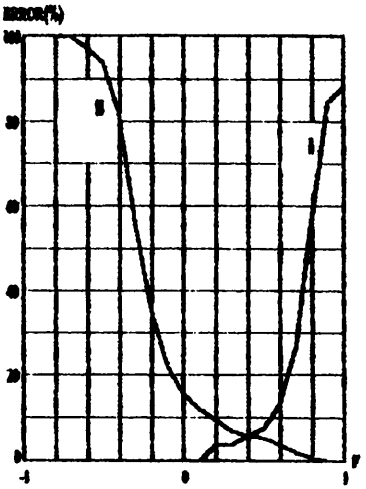

Рис. 8,б

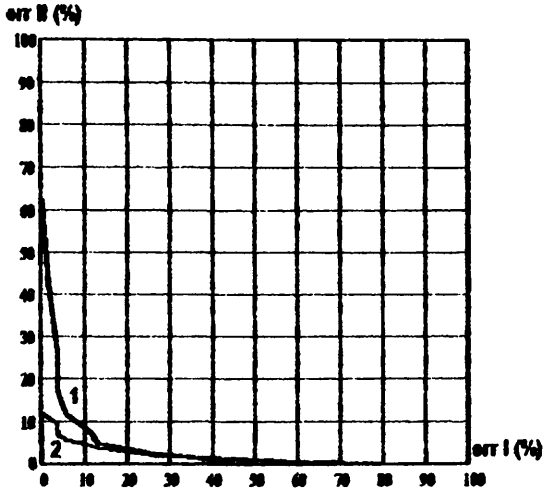

Рис. 9

На рис. 10–11 представлены результаты экспериментов по распознаванию сигнальных пептидов в контрольных фрагментах белков грамм-положительных бактерий, как содержащих сайты разрезания (108 фрагментов), так и без сайтов разрезания (32 фрагмента). Было выделено 7576 объектов ширины $L = 8$: 108 объектов образа «Сигнальный пептид» и 7468 объектов образа «Негатив». На рис. 10 приведены графики ошибок I–го и II–го рода в зависимости от порогового значения функции принадлежности F: с учетом только физико–химических свойств аминокислот (а), и с дополнительным использованием информации о частоте встречаемости аминокислот (б). На рис. 11 показано изменение ошибки II–го рода в зависимости от ошибки I–го рода: кривая 1 — учитывались только физико-химические свойства аминокислот, кривая 2 — дополнительно использовалась информация о частоте встречаемости аминокислот.

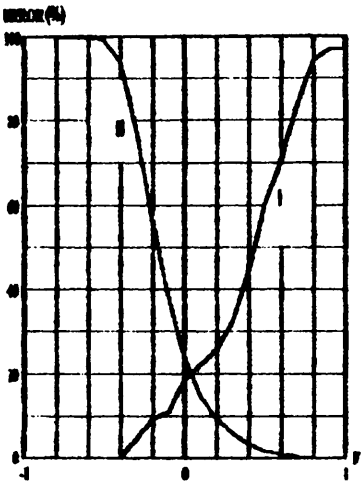

Рис. 10,а

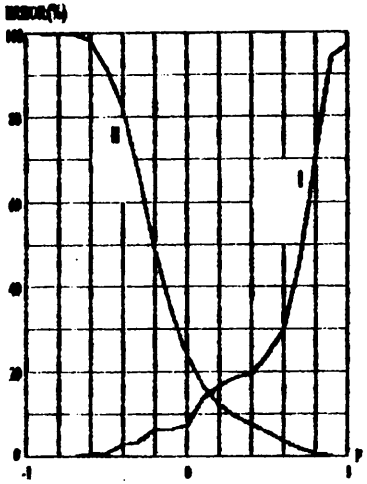

Рис. 10,б

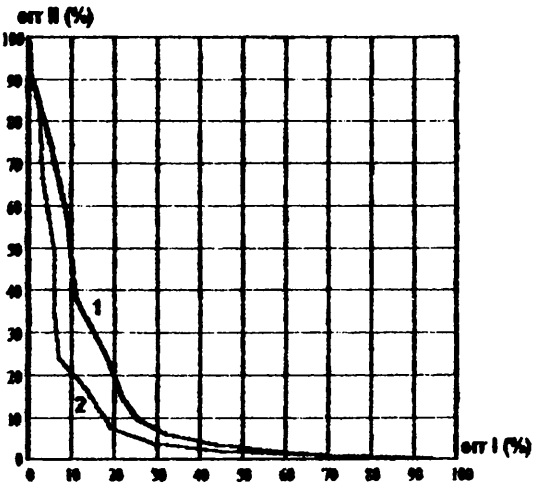

Рис. 11

На рис. 12–13 представлены результаты экспериментов по распознаванию сигнальных пептидов в контрольных фрагментах белков грамм–отрицательных бактерий, как содержащих сайты разрезания (172 фрагмента), так и без сайтов разрезания (93 фрагмента). Было выделено 13611 объектов ширины $L = 8$: 172 объекта образа «Сигнальный пептид» и 13439 объектов образа «Негатив». На рис. 12 приведены графики ошибок I–го и II–го рода в зависимости от порогового значения функции принадлежности F: с учетом только физико–химических свойств аминокислот (а), и с дополнительным использованием информации о частоте встречаемости аминокислот (б). На рис. 13 показано изменение ошибки II–го рода в зависимости от ошибки I–го рода: кривая 1 — учитывались только физико–химические свойства аминокислот, кривая 2 — дополнительно использовалась информация о частоте встречаемости аминокислот.

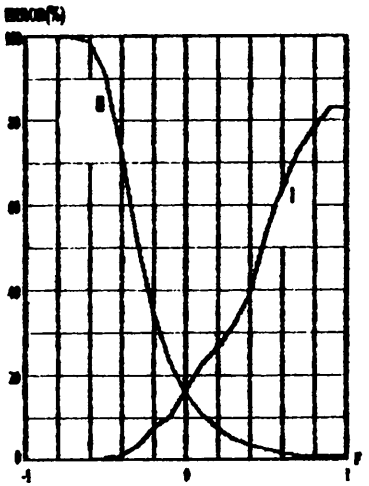

Рис. 12,а

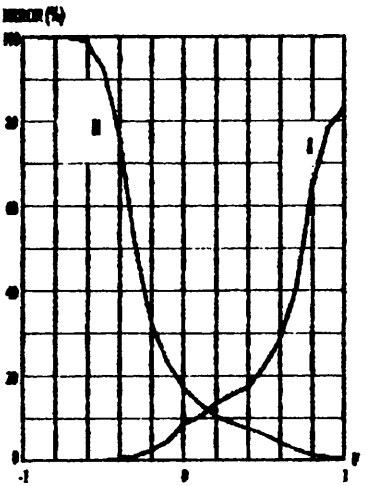

Рис. 12,б

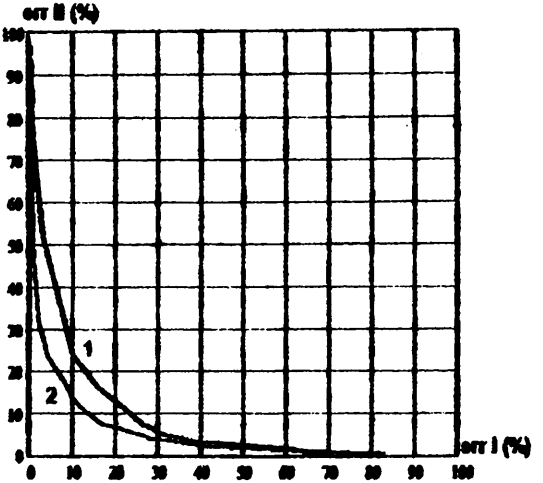

Рис. 13

На рис. 14–15 представлены результаты экспериментов по разделению сигнальных пептидов, содержащих сайты разрезания, и сигнальных якорей в фрагментах белков эукариот. На контроль было предъявлено 1002 объекта ширины $L = 8$: 955 объектов образа «Сигнальный пептид» и 47 объектов образа «Якорь». На рис. 14 приведены графики ошибок I–го и II–го рода в зависимости от порогового значения функции принадлежности F: с учетом только физико–химических свойств аминокислот (а), и с дополнительным использованием информации о частоте встречаемости аминокислот (б). На рис. 15 показано изменение ошибки II–го рода в зависимости от ошибки I–го рода: кривая 1 — учитывались только физико-химические свойства аминокислот, кривая 2 — дополнительно использовалась информация о частоте встречаемости аминокислот.

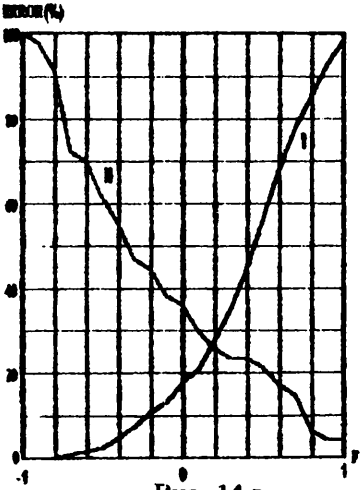

Рис. 14,а

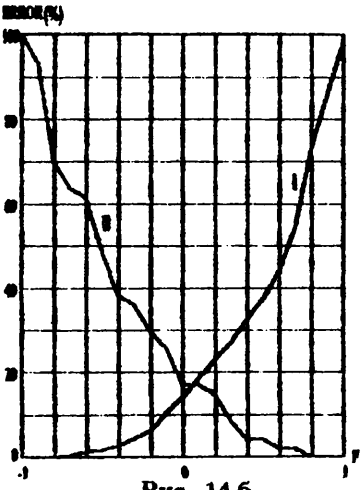

Рис. 14,б

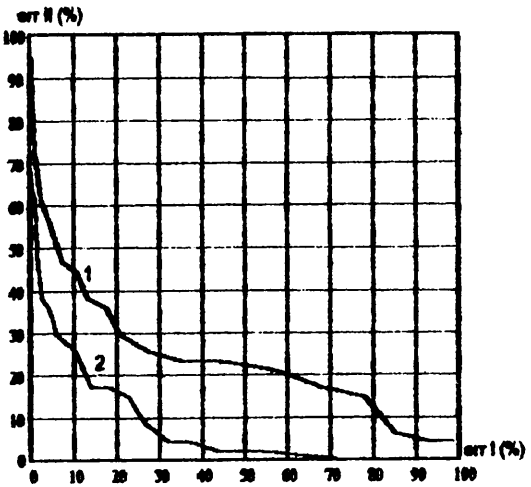

Рис. 15

На рис. 16–17 представлены результаты экспериментов по распознаванию сигнальных якорей в фрагментах белков эукариот. На контроль было предъявлено 50158 объектов ширины $L = 8$: 47 объектов образа «Якорь» и 50111 объектов образа «Негатив». На рис. 16 приведены графики ошибок I–го и II–го рода в зависимости от порогового значения функции принадлежности F: с учетом только физико–химических свойств аминокислот (а), и с дополнительным использованием информации о частоте встречаемости аминокислот (б). На рис. 17 показано изменение ошибки II–го рода в зависимости от ошибки I–го рода: кривая 1 — учитывались только физико–химические свойства аминокислот, кривая 2 — дополнительно использовалась информация о частоте встречаемости аминокислот.

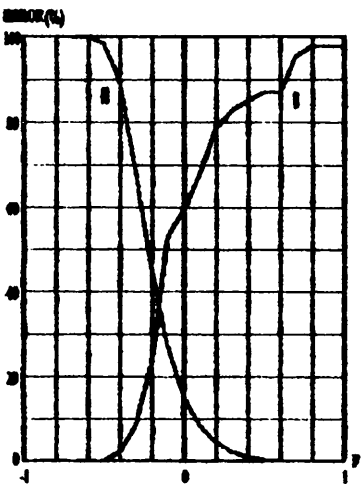

Рис. 16,а

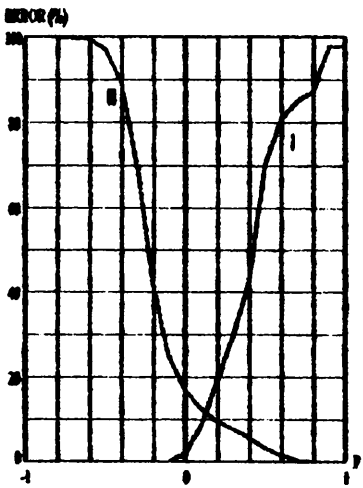

Рис. 16,б

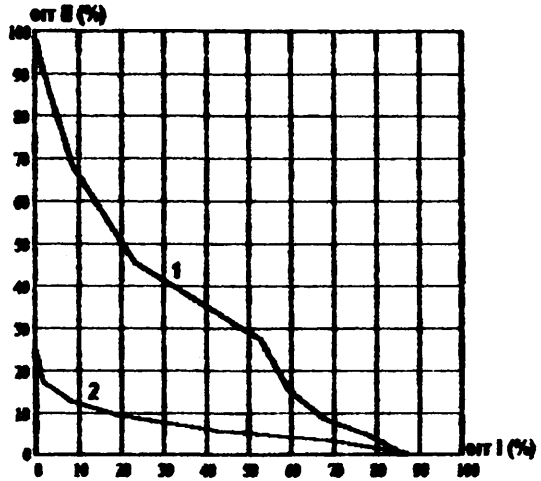

Рис. 17

На рис. 18–19 представлены результаты экспериментов по распознаванию сигнальных пептидов в фрагментах белков эукариот. На контроль было предъявлено 62358 объектов ширины $L = 8$: 705 объектов образа «Сигнальный пептид» и 61653 объекта образа «Негатив». На рис. 18 приведены графики ошибок I–го и II–го рода в зависимости от порогового значения функции принадлежности F: с учетом только физико–химических свойств аминокислот (а), и с дополнительным использованием информации о частоте встречаемости аминокислот (б). На рис. 19 показано изменение ошибки II–го рода в зависимости от ошибки I–го рода: кривая 1 — учитывались только физико-химические свойства аминокислот, кривая 2 — дополнительно использовалась информация о частоте встречаемости аминокислот.

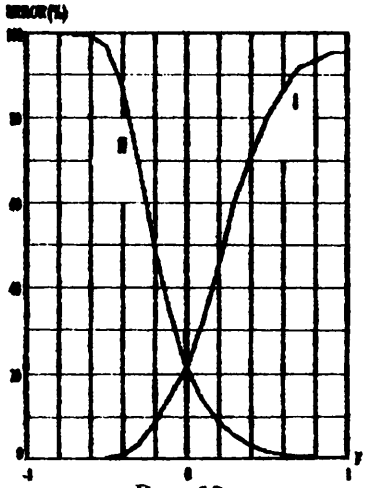

Рис. 18,а

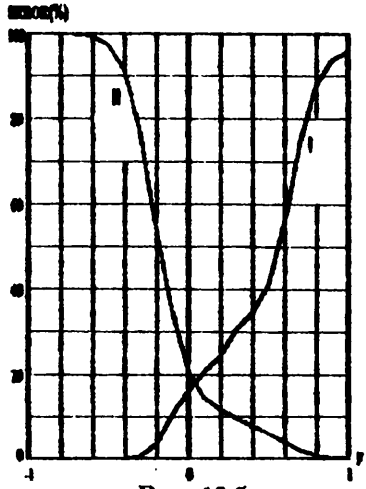

Рис. 18,б

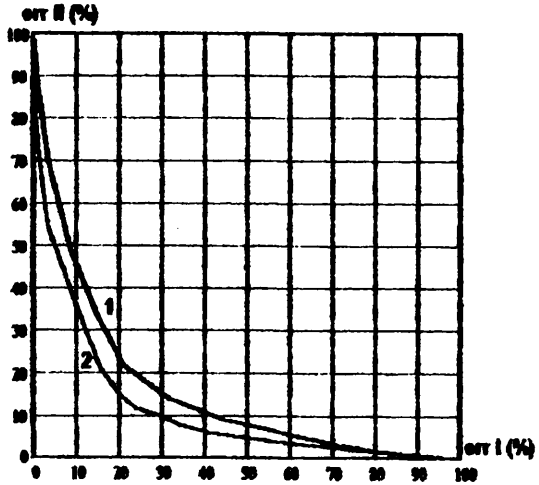

Рис. 19

На рис. 20–21 представлены результаты экспериментов по распознаванию сигнальных пептидов в фрагментах белков человека. На контроль было предъявлено 15735 объектов ширины $L = 8$: 290 объектов образа «Сигнальный пептид» и 15445 объектов образа «Негатив». На рис. 20 приведены графики ошибок I–го и II–го рода в зависимости от порогового значения функции принадлежности F: с учетом только физико-химических свойств аминокислот (а), и с дополнительным использованием информации о частоте встречаемости аминокислот (б). На рис. 21 показано изменение ошибки II–го рода в зависимости от ошибки I–го рода: кривая 1 — учитывались только физико–химические свойства аминокислот, кривая 2 — дополнительно использовалась информация о частоте встречаемости аминокислот.

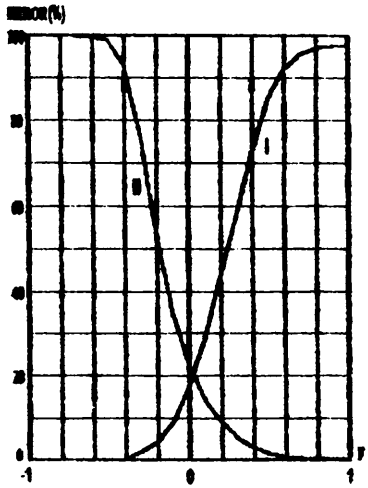

Рис. 20,а

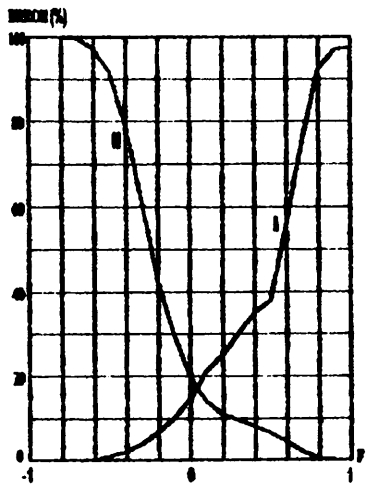

Рис. 20,б

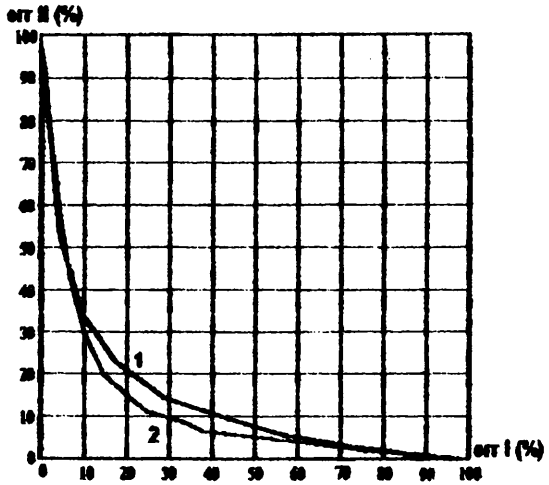

Рис. 21.

На рис. 22–23 представлены результаты экспериментов по распознаванию сигнальных якорей в фрагментах белков человека. На контроль было предъявлено 15291 объектов ширины $L = 8$: 14 объектов образа «Якорь» и 15277 объектов образа «Негатив». На рис. 22 приведены графики ошибок I-го и II-го рода в зависимости от порогового значения функции принадлежности F: с учетом только физико-химических свойств аминокислот (а), и с дополнительным использованием информации о частоте встречаемости аминокислот (б). На рис. 23 показано изменение ошибки II-го рода в зависимости от ошибки I-го рода: кривая 1 — учитывались только физико-химические свойства аминокислот, кривая 2 — дополнительно использовалась информация о частоте встречаемости аминокислот.

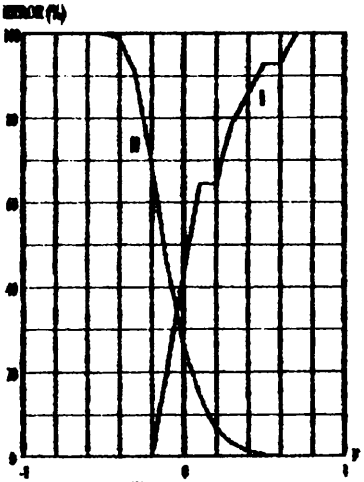

Рис. 22,а

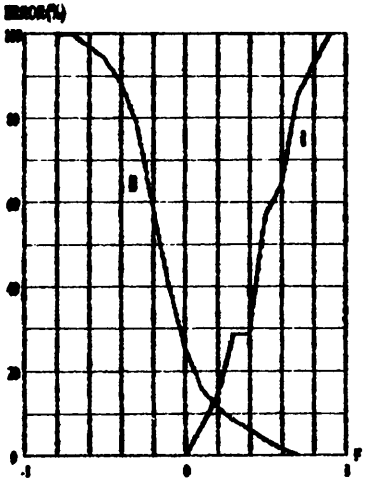

Рис. 22,б

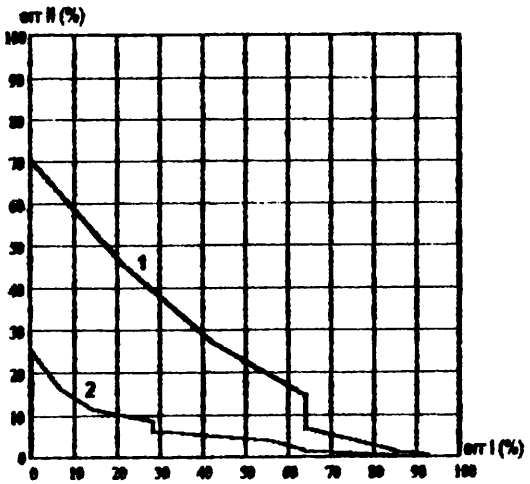

Рис. 23.

На рис. 24–25 представлены результаты экспериментов по разделению сигнальных пептидов, содержащих сайта разрезания, и сигнальных якорей в фрагментах белков человека. На контроль было предъявлено 416 объектов ширины $L = 8$: 402 объекта образа «Сигнальный пептид» и 14 объектов образа «Якорь». На рис. 24 приведены графики ошибок I-го и II-го рода в зависимости от порогового значения функции принадлежности F: с учетом только физико-химических свойств аминокислот (а), и с дополнительным использованием информации о частоте встречаемости аминокислот (б). На рис. 25 показано изменение ошибки II-го рода в зависимости от ошибки I-го рода: кривая 1 — учитывались только физико-химические свойства аминокислот, кривая 2 — дополнительно использовалась информация о частоте встречаемости аминокислот.

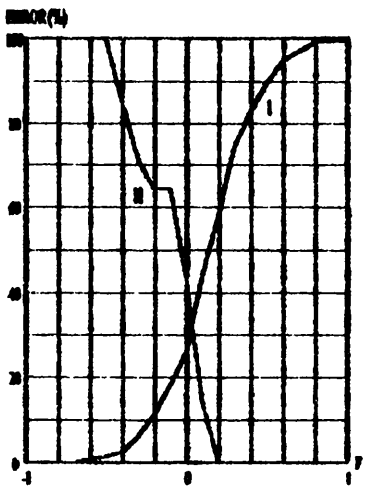

Рис. 24,а

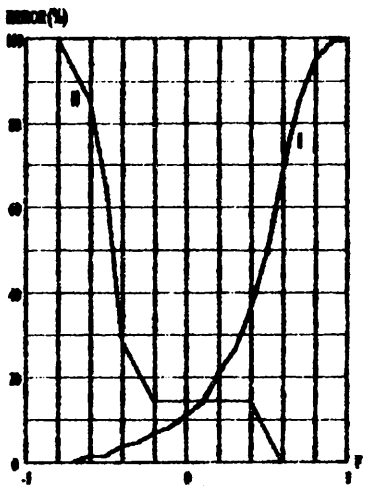

Рис. 24,б

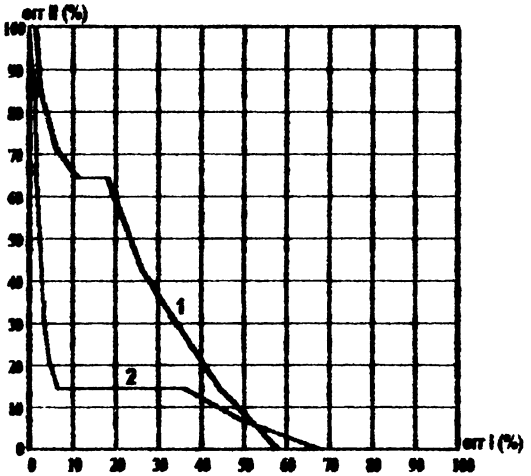
отт II (%)
100
90
80
70
60
50
40
30
20
10
0
1
2
0
10
20
30
40
50
60
70
80
90
100
отт I (%)

Рис. 25

ПРИЛОЖЕНИЕ 2

В табл. 3–4 приведены результаты экспериментов по распознаванию трех образов в контрольных фрагментах белков эукариот. К распознаванию было предъявлено 65607 объектов ширины $L = 8$: 705 образа «Сигнальный пептид», 47 — «Якорь» и 64855 объектов образа «Негатив». В табл. 3 представлены результаты распознавания с учетом только физико–химических свойств аминокислот, в табл. 4 — результаты распознавания с учетом как физико–химических свойств аминокислот, так и информации о частоте встречаемости аминокислот.

Таблица 3

Предъявлены	Распознаны		
	Сигнальный пептид	Якорь	Негатив
Сигнальный пептид	508	60	137
Якорь	14	13	20
Негатив	11743	9236	43876

Таблица 4

Предъявлены	Распознаны		
	Сигнальный пептид	Якорь	Негатив
Сигнальный пептид	583	12	110
Якорь	17	9	21
Негатив	12267	4827	47761

В табл. 5–6 представлены результаты экспериментов по распознаванию трех образов в контрольных фрагментах белков человека. К распознаванию было предъявлено 21266 объектов ширины $L = 8$: 290 образа «Сигнальный пептид», 14 — «Якорь» и 20962 объекта образа «Негатив». В табл. 5 представлены результаты распознавания с учетом только физико–химических свойств аминокислот, в табл. 6 — результаты распознавания с учетом

как физико-химических свойств аминокислот, так и информации о частоте встречаемости аминокислот.

Таблица 5

Предъявлены	Распознаны		
	Сигнальный пептид	Якорь	Негатив
Сигнальный пептид	199	43	48
Якорь	3	6	5
Негатив	3580	5047	12335

Таблица 6

Предъявлены	Распознаны		
	Сигнальный пептид	Якорь	Негатив
Сигнальный пептид	238	31	21
Якорь	2	12	0
Негатив	3681	5429	11852