Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Сунгуров Дмитрий Сегреевич

Многомодальный метод релевантных векторов в задаче распознавания вторичной структуры белка

511656 - Математические и информационные технологии

Магистерская диссертация

Научный руководитель:
в.н.с. ВЦ РАН, д.т.н.
Моттль Вадим Вячеславович

Москва
2013

Содержание

Аннотация

В работе рассматривается применение мультимодального метода релевантных векторов к задаче предсказания вторичной структуры белка. Мы ограничили задачу распознаванием стрэндов, как самой сложной частью общей задачи. Используется широко известный принцип локального окна, заключающийся в сравнении поданного на вход белка с фрагментами аминокислотных последовательностей белков обучающей совокупности. В отличие от стандартного метода релевантных векторов, предлагаемый в данной работе алгоритм позволяет управлять отбором различных способов сравнения классифицируемого белка и фрагментов обучающей совокупности. Эксперименты, проведенные на стандартном наборе белков RS126, показали существенное сокращение опорных объектов в решающем правиле и способность алгоритма отбирать подмножество наилучших функций сравнения из заданного набора.

1 Введение

В настоящее время основной подход при исследовании белков предполагает что первичная структура белка определяет его пространственную структуру, которая в свою очередь определяет биологическую роль белка. В результате одной из основных задач биоинформатики является установление связи между первичной и пространственной структурой белков.

Вторичная структура представляет собой проекцию локальной геометрии пространственной структуры белка в последовательность символов трехбуквенного алфавита: H - спираль, S - стрэнд, C - спираль. Задача предсказания вторичной структуры все чаще является проверкой множества методов, нацеленных на решение более сложной проблемы предсказания пространственной структуры [1, 2].

Проблема предсказания вторичной структуры впервые возникла в начале 1960х, когда было определено некоторое количество белковых структур с помощью рентгеновской кристаллографии. С началом применения машинного обучения к данной задаче было достигнуто значительное увеличение точности [3]. Несмотря на увеличение средней точности в последние годы наблюдается простой в данной области. Например, эксперименты в рамках конференции CASP (Critical Assessment of the Protein Structure Prediction) [4], которые проводятся с начала 90х, наглядно показывают отсутствие позитивного тренда в точности предсказания вторичной структуры как минимум в течение 10 лет с 1992 до 2002. Возможно, это является причиной, по которой проблема предсказания вторичной структуры была удалена из списка проблем рассматриваемых в рамках CASP (CASP-5 [5]).

Отсутствие наблюдаемого прогресса скорее всего результат множества биологических вспомогательных предположений лежащих в предсказательных моделях. Представляется целесообразным разрабатывать и тестировать алгоритмы, которые были бы основаны на минимуме предположений и включать адекватные процедуры отбора признаков представляющих аминокислотные фрагменты [6] и адекватные методы обучения для выведения зависимости между первичной и вторичной структурами из больших наборов белков с известными структурами.

Наиболее популярный способ предсказания вторичной структуры белка в позиции t - это ее оценка исходя из локального контекста, т.е на основании фрагмента аминокислотной последовательности фиксированной длины, симметрично расположенного по отношению к целевой позиции t [3]. С учетом обучающего множества белков, чья вторичная структура представлена строками в трехбуквенном алфавите h, s, c, это задача распознавания образов.

Метод опорных векторов (SVM, support vector machine) - наиболее популярный метод машинного обучения в распознавании образов [7]. При применении к задаче предсказания вторичной структуры белка [8] одним из преимуществ метода является то, что в результирующее решающее правило входит относительно небольшое количество опорных объектов, полученных из обучающей выборки в результате обучения. Однако, метод накладывает ограничения на функцию сравнения - она должна быть кернелом, т.е быть положительно определенной симметричной функцией.

Целью данной работы является, во-первых, избавиться от ограничений на функции сравнения двух белковых последовательностей, в отличии от чрезмерно требовательных ограничений обычного SVM, и, во-вторых, существенно уменьшить количество опорных векторов в решающем правиле. С этой целью, мы использовали не

традиционный SVM, а RVM Бишопа и Типпинга [9]. Два основных преимущества RVM – это, во-первых, отсутствие ограничений на функции сравнения, и, во-вторых, использование вместо опорных объектов, как это сделано в SVM, небольшого числа так называемых опорных векторов. В задаче предсказания вторичной структуры это означает, что в решающее правило войдет относительно небольшое число фрагментов белков из обучающей совокупности.

Для проблемы предсказания вторичной структуры это означает, что структура, находящаяся на последовательных позициях в полипептидной цепи нового белка будет предсказана на основании сравнения с небольшим числом опорных последовательностей, взятых из обучающей совокупности, и являющихся эталонными объектами, представляющими классы вторичной структуры.

Для предсказания вторичной структуры белка на основе его фрагментов мы применили мультимодальную модификацию RVM, описанную в [10], которая дополнительно позволяет выбирать подмножество наиболее подходящих функций сравнения окон среди заданных.

В качестве обучающей выборки использовался стандартный набор белков RS126.

Для проверки способности мультимодальной RVM выбирать наиболее подходящие функции сравнения, мы исследовали вместе два принципа сравнения - принцип, основанный на позициях аминокислот во фрагменте [11, 12], и новый подход, основанный на разложении Фурье первичной и вторичной структуры белка.

Мы ограничились рассмотрением задачи определения strand'ов во вторичной структуре белков, которая, как показывает практика, представляет собой наиболее проблематичную часть вторичной структуры. Целью данной работы является скорее исследование эффективности RVM к проблеме оконного предсказания вторичной структуры, чем установление нового рекорда точности. Тем не менее, эксперименты на RS126 показали точность около 75% в определении strand'ов.

1.1 Определения и обозначения

Каждый белок определяется двумя последовательностями равной длины:

- Первичная структура: $\omega = (\alpha_1, \alpha_2, \cdots, \alpha_L) \in A, \alpha_i \in A = (\alpha^1, \cdots, \alpha^{20})$ Строка на алфавите из 20 символов — аминокислот.
- Вторичная структура: $\mathbf{y} = (y_1, y_2, \cdots, y_L) \in Y, y_i \in Y = (H, L, S)$ Строка на алфавите из 3 символов — типов вторичной структуры.

Пример:
Первичная структура: $\cdots EMLRIDEGLRLKIYKDTE\cdots$
Вторичная структура: $\cdots HHHHHHHLLSSSSSSLLL\cdots$

Локальное окно:

Определение 1.1 *Окно $w_{t,n}$ длины n со смещением t вырезает из белка n символов, начиная с символа номер t.*

Класс y_t соответствующий каждому окну — тип вторичной структуры в центре окна.

M_i — длина аминокислотной последовательности i-го белка.

Таким образом, для каждого белка существует $M_j - n + 1$ различных окон. Каждое окно вырезает из первичной и вторичной структуры белка подслово длины n. Объектами обучающей совокупности являются элементы первичной структуры, вырезанные окном длины n, классом объекта является соответствующий элемент вторичной структуры.

W_i — признаковое описание окна до применения преобразования Фурье, матрица $n \times 20$, где столбцы — вектора признаковых описаний аминокислот.

x_i — признаковое описание окна после применения преобразования Фурье, отбора гармоник и разворачивания матрицы в строку.

$\{\mathbf{x}_i, y_i\}$ — пары признаковых описаний объектов и их классов.

1.2 Постановка задачи

Дан набор из пар $\{\mathbf{x}_i, y_i\}$, где $\mathbf{x}_i$ —- признаковое описание объекта (окна), y_i —- класс объекта (вторичной структуры). Требуется научиться предсказывать класс вторичной структуры по соответствующему окну.

1.3 Локальный подход к предсказанию вторичной структуры белка – распознавание образов в скользящем окне

Пусть $\omega = (\alpha_t, t = 1, \ldots, M)$ конечная аминокислотная последовательность, которая представляет первичную структуру белка длины $M = M_\omega$, где $\alpha_i \in A = (\alpha^1, \cdots, \alpha^m), m = 20$ символы алфавита на множестве аминокислот. Скрытая вторичная структура будет полностью представлена символьной последовательностью $\mathbf{y} = (y_t, t = 1, \ldots, M)$ той же длины $M = M_\omega$, элементы которой $y_t \in Y = h, s, c$ соответствуют трем классам вторичной структуры: h – спираль, s – стрэнд, c – неопределенная структура, обычно называемая петлей.

Далее, пусть, наблюдатель предоставил обучающую выборку, на которой для аминокислотных последовательностей известны правильные значения вторичной структуры:

$$\{(\omega_l, \mathbf{y}_l), l = 1, \ldots, N^0\}, \omega_l = (\alpha_{lt}, t = 1, \ldots, M_l), \mathbf{y}_l = (y_{lt}, t = 1, \ldots, M_l) \tag{1.1}$$

Новые аминокислотные последовательности $\omega = (\alpha_t, t = 1, \ldots, M_\omega)$ не представлены в обучающей выборке, нам требуется установить вторичную структуру соответствующего белка $\hat{\mathbf{y}(\omega)} = (\hat{y}_t(\omega), t = 1, \ldots, M_\omega)$.

Следуя (??), в данной работе используется принцип скользящего окна. Это означает что решение о классе вторичной структуры на позиции t делается на основании симметричного интервала $\omega_t = (\alpha_t, t - T \leqslant t \leqslant t + T)$ на всей аминокислотной последовательности $\omega = (\alpha_t, t = 1, \ldots, M)$. Нечетная ширина $\mathbf{T} = 2T + 1$ скользящего окна, таким образом определяется параметром полу-шириной T. Оценка вторичной структуры белка основывается только на обрезанной с обоих сторон аминокислотной последовательности на $\hat{\mathbf{y}(\omega)} = (\hat{y}_t(\omega), T+1 \leqslant t \leqslant M-T) = (\hat{y}_t(\omega_t), T+1 \leqslant t \leqslant M-T)$.

Таким образом, оригинальная задача предсказания всей вторичной структуры белка $\hat{\mathbf{y}(\omega)}$ сократилась до ряда независимых задач $\hat{y}_t(\omega_t) = \hat{y}_t(\alpha_{t-T}, \ldots, \alpha_t, \ldots, \alpha_{t+T})$. определения класса вторичной структуры $y_t \in h, s, c$ для центральной аминокислоты α_t в заданном окне.

Оконный подход предполагает рассмотрение обучающей выборки как неупорядоченный набор фрагментов аминокислотных последовательностей $\{(\omega_j, y_j), j = 1, \ldots, N\}$ вырезанных из аминокислотных последовательностей заданных белков $\omega_t = (\alpha_{jt}, t-T \leqslant t \leqslant t+T), y_j \in \{h, s, c\}$. Так как мы рассматриваем задачу различия strand не-strand, мы обучали двух-классовый классификатор: $y_j \in \{1, -1\} = \{s, \bar{s}\} = \{s, \{h, c\}\}$.

2 Методы попарного сравнения аминокислотных фрагментов для предсказания вторичной структуры

В этой работе применялась мультимодальная RVM [10] к предсказанию вторичной структуры белков, с различными подходами к сравнению аминокислотных последовательностей. Рассматривались два метода сравнения – принцип, основанный на позициях аминокислот во фрагменте [11] и новый подход, основанный на разложении Фурье первичной и вторичной структуры белка.

Каждая функция сравнения $S_i(\omega', \omega'')$ должна быть применима к любым двум аминокислотным фрагментам $\omega' = (\alpha_\tau, -T \leqslant \tau \leqslant T)$ и $\omega'' = (\alpha''_\tau, -T \leqslant \tau \leqslant T)$ длины $2T + 1$. В наших экспериментах мы рассматривали $T = 6$ и $T = 17$. В экспериментах использовались два значения:

- – $T = 6$, то есть окно длины $T = 13$, для сравнения со стороны позиций аминоислот.
- – $T = 17$, то есть окно длины $T = 35$, для метода на основе Фурье разложения. Такая длина окна позволяет учитывать зависимости между далеко расположенными аминокислотами.

Для каждого из двух методов рассматривались по 3 различных функций сравнения. Таким образом, всего $n = 6$ функций попарного сравнения белков.

2.1 Позиционный метод сравнения

Этот метод сравнения – обобщение метода, представленного в [11]. Пусть $A = \{\alpha^1, \cdots, \alpha^{20}\}$ алфавит на множестве аминокислот. Для каждой позиции $-T \leqslant \tau \leqslant T$ в окне $\omega = (\alpha_\tau, -T \leqslant \tau \leqslant T)$ и каждой из 20 аминокислот определен бинарный признак $z_{\tau k}(\omega) = 1$, если $\alpha_\tau = \alpha^k$, и $z_{\tau k}(\omega) = 0$, если $\alpha_\tau \neq \alpha^k$.

$$
\begin{aligned}
S_1(\omega', \omega'') &= \sum_{\tau=-T}^{T} \sum_{k=1}^{20} z_{\tau k}(\omega') z_{\tau k}(\omega''), \\
S_2(\omega', \omega'') &= \exp\left\{-\gamma \left[z_{\tau k}(\omega') - z_{\tau k}(\omega'')\right]^2\right\}, \\
S_3(\omega', \omega'') &= \sum_{\tau=-T}^{T} \sum_{k=1}^{20} |z_{\tau k}(\omega') - z_{\tau k}(\omega'')|.
\end{aligned}
\tag{2.1}
$$

Все признаки образуют бинарный вектор $z(w) = z_{\tau k}(\omega), -T \leqslant \tau \leqslant T, k = 1, \ldots, 20)$. Мы исследовали три вышеозначенные функции сравнения, основанные на таком признаковом описании объектов.

В [11] показано, что позиционный принцип сравнения работает лучше, когда применяется к сравнительно небольшим длинам окон, поэтому использовалась рекомендованная длина окна $2T + 1 = 13$.

2.2 Сравнение на базе Фурье-спектра

Метод основанный на преобразовании фурье каждого локального окна. Такой подход представлен впервые. Он основан на том, что подстановочные матрицы аминокислот PAM и BLOSUM являются результатом одной и той же эволюционной модели PAM [13]. Кроме того, в [13] показано, что все матрицы PAM и BLOSUM выражают вероятность существования общего предка для каждой пары аминокислот, и соответственно, должны быть положительно определенными. Отсутствие положительной определенности в опубликованных матрицах – результат традиционного логарифмического представления и округления чисел.

2.2.1 Марковская цепь склонности аминокислот к взаимному мутационному превращению

Принято различать двадцать аминокислот $A = \{\alpha^{(i)}, i = 1, \cdots, m\}, m = 20$, и обозначать их буквами латинского алфавита (табл. 1).

Табл. 1. Двадцать аминокислот и их буквенные обозначения.

Alanine	Ala	A	Methionine	Met	M
Cysteine	Cys	C	Asparagine	Asn	N
Aspartic Acid	Asp	D	Proline	Pro	P
Glutamic Acid	Glu	E	Glutamine	Gln	Q
Phenylalanine	Phe	F	Arginine	Arg	R
Glycine	Gly	G	Serine	Ser	S
Histidine	His	H	Threonine	Thr	T
Isoleucine	Ile	I	Valine	Val	V
Lysine	Lys	K	Tryptophan	Trp	W
Leucine	Leu	L	Tyrosine	Tyr	Y

В современной биоинформатике широкое применение нашла вероятностная модель эволюции Маргарет Дэйхофф [8], получившая название PAM (Point Accepted Mutation), и играющая роль основной теоретической концепции сравнения аминокислот, а затем и белков, по их сходству. Центральная гипотеза этой модели заключается в том, что эволюционные изменения в аминокислотной последовательности белка складываются из случайных независимых изменений (Mutation) отдельных аминокислот цепи (Point), причем таких изменений, которые закрепились в ходе дальнейшего естественного отбора (Accepted).

Эволюционная модель PAM предполагает, что склонность аминокислот к взаимному мутационному превращению количественно выражается квадратной матрицей условных вероятностей $\psi_{j|i} = \psi(\alpha^{(j)}|\alpha^{(i)}$, интерпретируемых как вероятность того, что аминокислота $\alpha^{(i)}$ на очередном шаге эволюции превратится в аминокислоту $\alpha^{(j)}$.

Основным математическим понятием модели Дэйхофф является понятие марковской цепи истории эволюционного изменения аминокислоты в отдельно взятой позиции (точке) цепи $h_t, t = 1, 2, 3 \cdots$ при некоторой достаточно сложной интерпретации понятия "величины шага"$(\cdots, t-1, t, t+1, \cdots)$, которую мы здесь не рассматриваем. В любом случае предполагается, что марковская цепь эволюции, определяемая стохастической матрицей условных вероятностей переходов:

$$\Psi = \begin{pmatrix} \psi_{1|1} & \cdots & \psi_{1|m} \\ \cdots & \ddots & \cdots \\ \psi_{m|1} & \cdots & \psi_{m|m} \end{pmatrix} (m \times m), \sum_{j=1}^{m} \psi_{j|i} = 1, \psi_{j|i} = \psi(\alpha^{(j)}|\alpha^{(i)}) = P(h_t = j|h_{t-1} = i), \tag{2.2}$$

представляет собой эргодический случайный процесс, характеризующийся финальным распределением вероятностей:

$$\begin{array}{l} \xi_j = \xi(\alpha^{(j)}) = P(h_t = j) = \sum_{i=1}^{m} P(h_{t-1} = i) P(h_t = j|h_{t-1} = i) = \\ = \sum_{i=1}^{m} \xi_i \psi_{j|i} = \sum_{i=1}^{m} \xi(\alpha^{(i)}) \psi(\alpha^{(j)}|\alpha^{(i)}). \end{array} \tag{2.3}$$

Удобно представить финальные вероятности марковской цепи эволюционного чередования аминокислот в виде вектора:

$$\xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_m \end{pmatrix} \in R^m, 0 \leqslant \xi_i \leqslant 1, \sum_{i=1}^{m} \xi_i = 1, \tag{2.4}$$

тогда условие эргодичности (2.3) будет означать, что вектор финальных вероятностей является собственным вектором транспонированной переходной матрицы, соответствующим единичному собственному числу:

$$\mathbf{\Psi}^T \xi = \xi. \tag{2.5}$$

Предположение об эргодичности марковской цепи эволюции аминокислот является формализацией предположения, что эволюционный процесс начался "очень давно случайный процесс мутаций успел установиться, и уже не зависит от неизвестных начальных вероятностей состояний. Финальное распределение вероятностей $\xi_i = \xi(\alpha^{(i)}), i = 1 \cdots, m$ интерпретируется как совокупность частот встречаемости аминокислот в составе существующих в природе белков.

Эквивалентные равенства (2.3) и (2.5) дают однородную систему линейных алгебраических уравнений относительно вектора финальных вероятностей

$$(\mathbf{\Psi}^T - \mathbf{I})\xi = 0,$$

в которой одно из уравнений следует заменить условием $\mathbf{1}^T \xi = \sum_{i=1}^{m} \xi_i = 1$.

Другим фундаментальным предположением модели Дэйхофф является предположение об обратимости марковской цепи процесса эволюции:

$$P(h_{t-1} = j|h_s = i) = P(h_t = j|h_{s-1} = i),$$

то есть

$$\xi(\alpha^{(j)})\psi(\alpha^{(i)}|\alpha^{(j)})/\xi(\alpha^{(i)}) = \psi(\alpha^{(j)}|\alpha^{(i)}), \xi(\alpha^{(i)})\psi(\alpha^{(j)}|\alpha^{(i)}) = \xi(\alpha^{(j)})\psi(\alpha^{(i)}|\alpha^{(j)}) \quad (2.6)$$

Такое предположение означает принципиальную невозможность установить в процессе наблюдения, какая из двух аминокислот является прародителем, а какая потомком.

2.2.2 Потенциальная функция на алфавите аминокислот

Сходство пары аминокислот $\alpha^{(i)}, \alpha^{(j)} \in A$ часто оценивают, вычисляя вероятность их происхождения в результате двух независимых ветвей эволюции от одной и той же неизвестной аминокислоты $\alpha^* \in A$ [14]:

$$K(\alpha^{(i)}, \alpha^{(j)}) = \sum_{k=1}^{m} \xi(\alpha^{(k)})\psi(\alpha^{(i)}|\alpha^{(k)})\psi(\alpha^{(j)}|\alpha^{(k)}) \quad (2.7)$$

Нетрудно убедиться, что симметрическая матрица на множестве аминокислот (2.7)

$$\mathbf{K} = \begin{pmatrix} K(\alpha^{(1)}, \alpha^{(1)}) & \cdots & K(\alpha^{(1)}, \alpha^{(m)}) \\ \cdots & \ddots & \cdots \\ K(\alpha^{(m)}, \alpha^{(1)}) & \cdots & K(\alpha^{(m)}, \alpha^{(m)}) \end{pmatrix} (m \times m), \quad (2.8)$$

неотрицательно определена. Свяжем с каждой аминокислотой $\alpha^{(i)} \in A$ вектор условных вероятностей ее появления вместо каждой другой аминокислоты на очередном шаге эволюции, так, что переходная матрица (2.2) будет образована из этих векторов как из столбцов

$$\mathbf{\Psi} = (\psi^{(1)} \cdots \psi^{(m)})(m \times m), \psi^{(i)} = \begin{pmatrix} \psi_{i|1} \\ \vdots \\ \psi_{i|m} \end{pmatrix} \in R^m, i = 1 \cdots, m \quad (2.9)$$

Поскольку в модели эволюции Дэйхофф этот вектор является единственным носителем информации об эволюционном различии аминокислот, естественно называть его вектором эволюционных признаков аминокислоты $\alpha^{(i)}$. Определим диагональную матрицу

$$\mathbf{\Xi} = Diag(\xi)(m \times m),$$

тогда матрица $\mathbf{M}$ может быть представлена в виде

$$\mathbf{M} = \begin{pmatrix} \psi^{(1)T}\mathbf{\Xi}\psi^{(1)} & \cdots & \psi^{(1)T}\mathbf{\Xi}\psi^{(m)} \\ \cdots & \ddots & \cdots \\ \psi^{(m)T}\mathbf{\Xi}\psi^{(1)} & \cdots & \psi^{(m)T}\mathbf{\Xi}\psi^{(m)} \end{pmatrix}$$

как матрица скалярных произведений векторов эволюционных признаков с общим

для всех пар вектором весов ξ . Очевидно, что матрица $\mathbf{M}$ неотрицательно определена как матрица Грамма.

На самом же деле можно утверждать даже большее - матрица $\mathbf{M}$ оказывается положительно определенной для всех принятых в биоинформатике "масштабов"элементарного шага эволюционной цепи.

Таким образом, двухместная функция $K(\alpha^{(i)}, \alpha^{(j)}) : A \times A \to R$ (2.7) является потенциальной функцией на алфавите аминокислот. Эта потенциальная функция погружает 20 реально существующих в природе аминокислот в некоторое линейное пространство $\tilde{A} \subset A$, в данном случае, 20-ти мерное линейное пространство, в котором исходный алфавит аминокислот играет роль базиса, линейно независимого в силу строгой положительной определенности и, следовательно, невырожденности матрицы $\mathbf{M}$.

Потенциальную функцию на алфавите аминокислот (2.7) можно выразить в более удобной форме, если ввести понятие двухшаговой матрицы условных вероятностей переходов в дополнение к одношаговой матрице (2.2):

$$\mathbf{\Psi}_{[2]} = \begin{pmatrix} \Psi_{[2]1|1} & \cdots & \Psi_{[2]1|m} \\ \cdots & \ddots & \cdots \\ \Psi_{[2]m|1} & \cdots & \Psi_{[2]m|m} \end{pmatrix} (m \times m), \Psi_{[2]j|i} = P(h_t = j | h_{t-2} = i) = \sum_{i=1}^{m} \psi_{k|i}\psi_{j|k} \tag{2.10}$$

Очевидно, что

$$\mathbf{\Psi}_{[2]}^T = \mathbf{\Psi}^T \mathbf{\Psi}^T. \tag{2.11}$$

Тогда, принимая во внимание обратимость марковской цепи эволюции по Дэйхофф (2.6), для потенциальной функции (2.7) справедливо равенство

$$K(\alpha^{(i)}, \alpha^{(j)}) = \sum_{k=1}^{m} \xi(\alpha^{(i)})\psi(\alpha^{(k)}|\alpha^{(i)})\psi(\alpha^{(j)}|\alpha^{(k)}) = \xi(\alpha^{(i)}) \sum_{k=1}^{m} \psi(\alpha^{(k)}|\alpha^{(i)})\psi(\alpha^{(j)}|\alpha^{(k)}),$$

и ее можно представить в виде

$$K(\alpha^{(i)}, \alpha^{(j)}) = \xi(\alpha^{(i)})\psi_{[2]}(\alpha^{(j)}|\alpha^{(i)}). \tag{2.12}$$

Легко показать, что мера эволюционного сходства аминокислот $K(\alpha^{(i)}, \alpha^{(j)})$ остается потенциальной функцией, если ее нормировать относительно финальных вероятностей $\xi(\alpha^{(i)})$ и $\xi(\alpha^{(j)})$, т.е. частот встречаемости аминокислот в природе:

Теорема 2.1 *Двухместная функция*

$$\bar{K}(\alpha^{(i)}, \alpha^{(j)}) = \frac{K(\alpha^{(i)}, \alpha^{(j)})}{\xi(\alpha^{(i)})\xi(\alpha^{(j)})} = \frac{\psi_{[2]}(\alpha^{(j)}|\alpha^{(i)})}{\xi(\alpha^{(j)})}. \tag{2.13}$$

является потенциальной функцией.

Доказательство.

Чтобы убедиться в этом, достаточно представить $\bar{K}(\alpha^{(i)}, \alpha^{(j)})$ как скалярное произведение векторов:

$$\bar{K}(\alpha^{(i)}, \alpha^{(j)}) = \sum_{k=1}^{m} \left(\frac{\sqrt{\xi(\alpha^{(k)})}}{\xi(\alpha^{(i)})}\psi(\alpha^{(i)}|\alpha^{(k)})\right)\left(\frac{\sqrt{\xi(\alpha^{(k)})}}{\xi(\alpha^{(j)})}\psi(\alpha^{(j)}|\alpha^{(k)})\right) \tag{2.14}$$

■

2.2.3 Признаковое описание аминокислот в собственном базисе пространства образованного потенциальной функцией

Потенциальная функция $K(\alpha^{(i)}, \alpha^{(j)})$ допускает погружение множества аминокислот в линейное пространство R^{20} со скалярным произведением. Роль скалярного произведения играет сама исходная потенциальная функция.

Теорема 2.2 *Для всякого скалярного произведения(потенциальной функции) одноместная функция* $\| \alpha \| = \sqrt{K(\alpha, \alpha)}$ *является нормой.*

Такая норма называется евклидовой нормой. Метрика, которую она порождает, называется евклидовой метрикой:

$$\begin{aligned} \rho(\alpha', \alpha'') &= \rho(\alpha'', \alpha') = \| \alpha' - \alpha'' \| = \sqrt{K(\alpha' - \alpha'', \alpha' - \alpha'')} = \\ &= K(\alpha', \alpha') + K(\alpha'', \alpha'') - 2K(\alpha', \alpha''). \end{aligned} \tag{2.15}$$

Таким образом, всякое евклидово линейное пространство имеет определенную в нем евклидову метрику.

Рассмотрим общий случай, когда требуется найти собственный базис $\{\theta_i, i = 1, \cdots, n\}$, при заданной совокупности N элементов $A = \{\alpha_j, j = 1, \cdots, N\}$ в евклидовом пространстве $\tilde{A}$ со скалярным произведением $K(\alpha', \alpha'') : \tilde{A} \times \tilde{A} \to R$. Вообще говоря, линейное пространство A' размерности n, натянутое на базис $\{\theta_i, i = 1, \cdots, n\}$, не содержит совокупность $A = \{\alpha_j, j = 1, \cdots, N\}$, так что ее элементы не могут быть представлены как линейные комбинации по этому базису, но всегда можно найти для них наилучшие приближения:

$$\alpha_j \neq \sum_{i=1}^{n} a_i \Theta_i, \hat{\alpha}_j = \sum_{i=1}^{n} \hat{a}_i \Theta_i. \tag{2.16}$$

Поставим сначала задачу найти элемент $\hat{\alpha}_j$ наиболее близкий к α_j в смысле евклидовой метрики:

$$\hat{\alpha}_j = argmin_{\phi \in A'} \rho^2(\alpha_j, \phi) = argmin_{\phi \in A'} [K(\alpha_j, \alpha_j) + K(\phi, \phi) - 2K(\alpha_j, \phi)] =$$
$$= argmin_{\phi \in A'} [K(\phi, \phi) - 2K(\alpha_j, \phi)].$$

Такой элемент $\hat{\alpha}_j \in A'$ называется проекцией элемента $\alpha_j \in \tilde{A}$ на линейное подпространство A', натянутое на базис $\{\theta_i, i = 1, \cdots, n\} \in \tilde{A}$.

Поскольку для всякого элемента $\phi \in A'$ существует представление $\phi = \sum_{i=1}^{n} a_i \theta_i$, то в силу линейности скалярного произведения надо решить задачу:

$$\hat{\mathbf{a}} = (\hat{a}_i, i = 1, \cdots, n) = argmin_{a \in R^n} (\sum_{i=1}^{n} \sum_{l=1}^{n} K(\theta_i, \theta_l) a_i a_l - 2 \sum_{l=1}^{n} K(\alpha_j, \theta_l) a_l).$$

Минимизации подлежит квадратичная функция, строго выпуклая в силу положительной определенности матрицы скалярных произведений элементов линейно независимого базиса. Точка минимума

$$\frac{\partial}{\partial a_i}[\sum_{k=1}^{n}\sum_{l=1}^{n}K(\theta_k,\theta_l)a_k a_l - 2\sum_{l=1}^{n}K(\alpha_j,\theta_l)a_l] = 2[\sum_{l=1}^{n}K(\theta_i,\theta_l)a_l - K(\alpha_j,\theta_i)] = 0$$

является единственной, откуда следует, что следует решить строго определенную систему из n линейных алгебраических уравнений с n неизвестными:

$$\sum_{l=1}^{n}K(\theta_i,\theta_l)a_l = K(\alpha_j,\theta_i), i = 1,...,n.$$

Если базис является ортогональным $K(\theta_i,\theta_l) = 0$ при $i \neq l$, то система уравнений распадается на n отдельных равенств

$$\hat{a}_i = \frac{K(\alpha_j,\theta_i)}{K(\theta_i,\theta_i)},$$

а если к тому же базис ортонормирован $K(\theta_i,\theta_i) = 1$, то равенства еще проще:

$$\hat{a}_i = K(\alpha_j,\theta_i).$$

Итак, наилучшее приближение элемента $\alpha_j \in \tilde{A}$ по базису $\{\theta_i, i = 1,\cdots,n\} \in \tilde{A}$ есть линейная комбинация

$$\hat{\alpha}_j = \sum_{i=1}^{n}\frac{K(\alpha_j,\theta_i)}{K(\theta_i,\theta_i)}\theta_i, \tag{2.17}$$

а в случае ортонормированного базиса

$$\hat{\alpha}_j = \sum_{i=1}^{n}K(\alpha_j,\theta_i)\theta_i, \tag{2.18}$$

Вернемся к задаче нахождения представления конечной совокупности элементов евклидова линейного пространства в собственном базисе.

Неточность наилучшего приближения естественно количественно измерять как сумму квадратов евклидовых норм остатков для всех элементов заданной совокупности $A = \{\alpha_j, j = 1,\cdots,N\}$:

$$J(\theta_i, i = 1,\cdots,n|A) = \sum_{j=1}^{N}\rho^2(\alpha_j,\hat{\alpha}_j) = \sum_{j=1}^{N}\rho^2(\alpha_j,\sum_{i=1}^{n}\hat{a}_i\Theta_i.$$

Поставим задачу найти наилучший базис $\{\theta_i, i = 1,\cdots,n\}$, для которого суммарная ошибка представления элементов заданной совокупности $A = \{\alpha_j, j = 1,\cdots,N\}$ будет наименьшей. Такой базис будем называть собственным базисом этой совокупности.

Достаточно искать собственный базис только среди ортогональных базисов

$$K(\theta_i,\theta_k) = \begin{cases} b_i, i = k, \\ 0, i \neq k, \end{cases}$$

где $b_i > 0$ — заданные нормы элементов базиса.

Пусть ортогональный базис $\{\theta_i, i = 1, \cdots, n\}$ фиксирован. Тогда согласно (2.18):

$$
\begin{aligned}
J \quad & (\theta_i, i = 1, \cdots, n|A) = \sum_{j=1}^{N} \rho^2(\alpha_j, \hat{\alpha}_j) = \sum_{j=1}^{N} K(\alpha_j - \hat{\alpha}_j, \alpha_j - \hat{\alpha}_j) = \\
= \quad & \sum_{j=1}^{N} K(\alpha_j - \sum_{k=1}^{n} K(\alpha_j, \theta_k)\theta_k, \alpha_j - \sum_{i=1}^{n} K(\alpha_j, \theta_i)\theta_i) = \\
= \quad & \sum_{j=1}^{N} [K(\alpha_j, \alpha_j - \sum_{i=1}^{n} K(\alpha_j, \theta_i)\theta_i) - K(\sum_{k=1}^{n} K(\alpha_j, \theta_k)\theta_k, \alpha_j - \sum_{i=1}^{n} K(\alpha_j, \theta_i)\theta_i)] = \\
= \quad & \sum_{j=1}^{N} [K(\alpha_j, \alpha_j - \sum_{i=1}^{n} K(\alpha_j, \theta_i)\theta_i) - \sum_{k=1}^{n} K(\alpha_j, \theta_k) K(\theta_k, \alpha_j - \sum_{i=1}^{n} K(\alpha_j, \theta_i)\theta_i)].
\end{aligned}
$$

Здесь $K(\theta_k, \alpha_j - \sum_{i=1}^{n} K(\alpha_j, \theta_i)\theta_i) = 0$ для всех $k = 1, \cdots, n$, поскольку остаточный член многочлена наилучшего приближения ортогонален ко всем элементам базиса, следовательно,

$$
\begin{aligned}
J \quad & (\theta_i, i = 1, \cdots, n|A) = \sum_{j=1}^{N} \rho^2(\alpha_j, \hat{\alpha}_j) == \sum_{j=1}^{N} K(\alpha_j, \alpha_j - \sum_{i=1}^{n} K(\alpha_j, \theta_i)\theta_i) = \\
= \quad & \sum_{j=1}^{N} K(\alpha_j, \alpha_j) - \sum_{i=1}^{n} \sum_{j=1}^{N} (K(\alpha_j, \theta_i))^2.
\end{aligned}
$$

Таким образом, для минимизации неточности приближения заданной совокупности $A = \{\alpha_j, j = 1, \cdots, N\}$ надо максимизировать сумму квадратов скалярных произведений элементов базиса со всеми элементами совокупности:

$$
\begin{cases}
F(\theta_i, i = 1, \cdots, n|A) = \sum_{i=1}^{n} \sum_{j=1}^{N} (K(\alpha_j, \theta_i))^2 \to max(\theta_i \in \tilde{A}, i = 1, \cdots, n), \\
K(\theta_i, \theta_i) = b_i, i = 1, \cdots, n.
\end{cases}
\tag{2.19}
$$

Функция Лагранжа для критерия (4):

$$
L(\theta_i, \lambda_i, i = 1, \cdots, n) = \sum_{i=1}^{n} \sum_{j=1}^{N} (K(\alpha_j, \theta_i))^2 - \sum_{i=1}^{n} \lambda_i (K(\theta_i, \theta_i) - b_i).
$$

Ее градиент (дифференциал Фреше) по θ_i:

$$
\nabla_{\theta_i} L(\theta_k, \lambda_k, k = 1, \cdots, n) = 2 \sum_{j=1}^{N} K(\alpha_j, \theta_i)\alpha_j - 2\lambda_i \theta_i = \phi \in \tilde{A}.
$$

Мы пришли к условию

$$
\sum_{j=1}^{N} K(\alpha_j, \theta_i)\alpha_j = \lambda_i \theta_i. \tag{2.20}
$$

Заметим, что значения $K(\alpha_j, \theta_i)$ в (2.20) являются действительными числами, следовательно, элементы собственного базиса представляют собой линейные комбинации элементов совокупности $A = \{\alpha_j, j = 1, \cdots, N\}$:

$$\theta_i = \sum_{j=1}^{N} c_{ij}\alpha_j, i = 1, \cdots, n. \tag{2.21}$$

Тогда согласно (2.17) $K(\alpha_j, \theta_i) = K(\alpha_j, \sum_{k=1}^{N} c_{ik}\alpha_k) = \sum_{k=1}^{N} K(\alpha_j, \alpha_k)c_{ik}, c_{ik} = (1/\lambda_i)K(\alpha_j, \theta_i)$, откуда следует условие для определения коэффициентов c_{ij}:

$$\sum_{k=1}^{N} K(\alpha_j, \alpha_k)c_{ik} = \lambda_i c_{ij}, i = 1, \cdots, n. \tag{2.22}$$

Таким образом, величины $\{\lambda_i, i = 1, \cdots, n\}$ являются собственными числами матрицы $[K(\alpha_j, \alpha_k), j, k = 1, \cdots, N](N \times N)$, а коэффициенты $\{(c_{i1}, \cdots, c_{iN}, i = 1, \cdots, n\}$ — соответствующими им собственными векторами этой матрицы $(c_{i1}, \cdots, c_{iN})^T \in R^N$. В нашем случае число элементов собственного базиса $\{\theta_i \in \tilde{A}, i = 1, \cdots, n\}$ равно числу объектов $n = N$. Тогда каждый объект совокупности $A = \{\alpha_j, j = 1, \cdots, N\}$ в точности выражается в виде линейной комбинации элементов ортонормированного базиса (2), (3):

$$\alpha_j = \sum_{i=1}^{N} a_{ij}\Theta_i = \sum_{i=1}^{N} K(\alpha_j, \theta_i)\theta_i. \tag{2.23}$$

Каждому объекту α_j соответствует вектор его представления по собственному базису

$$\mathbf{a}_j = (a_{1j} \ldots a_{20j})^T \in R^N, a_{ij} = K(\alpha_j, \theta_i) = K(\alpha_j, \sum_{k=1}^{N} c_{ik}\alpha_k) = \sum_{k=1}^{N} K(\alpha_j, \alpha_k)c_{ik} = \lambda_i c_{ij} \tag{2.24}$$

Поскольку каждый объект обучающей совокупности – это фрагмент длины n первичной структуры белка (где n – длина окна), то при переходе к описанию аминокислот как векторов из R^{20}, мы получим матрицу $20 \times n$.

2.2.4 Дискретное преобразование Фурье

Для последовательности $w_n, n = 0, 1, \cdots, N-1$ состоящей из N действительных или комплексных чисел определяется дискретное преобразование Фурье (ДПФ):

$$X_k = \sum_{n=0}^{N-1} w_n e^{-\frac{2\pi i}{N}kn}, k = 0, \cdots, N-1, \tag{2.25}$$

где X_k - комплексная амплитуда k-й гармоники, $e^{-\frac{2\pi i}{N}kn}$ - ортогональные дискретные экспоненциальные функции.

Рассмотрим коэффициенты Фурье действительного сигнала:

$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn}$,
$X_{N-k} = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}(N-K)n}$,
$e^{-\frac{2\pi i}{N}kn} = \cos(\frac{2\pi i}{N}kn) - i\sin(\frac{2\pi i}{N}kn)$,
$e^{-\frac{2\pi i}{N}(N-k)n} = \cos(\frac{2\pi i}{N}(N-k)n) - i\sin(\frac{2\pi i}{N}(N-k)n)$,
$\cos(\frac{2\pi i}{N}(N-k)n) = \cos(\frac{2\pi i}{N}kn)$,
$\sin(\frac{2\pi i}{N}kn) = -\sin(\frac{2\pi i}{N}(N-k)n)$.

Таким образом

$$Re(X_k) = Re(X_{N-k}), Im(X_k) = -Im(X_{N-k}). \tag{2.26}$$

Т.е вторая половина коэффициентов является зеркальным отражением первой, и может игнорироваться.

Таким образом, для действительного сигнала всего $[N/2]$ (для нечетного N) комплексных и 1 действительный коэффициент Фурье (0 гармоника).

2.2.5 Выделение основных гармоник

Так как при сдвиге окна аминокислотная последовательность меняется мало (при достаточно большом размере окна), то желательно сделать так, чтобы вектор признаков этого окна так же менялся слабо.

Утверждение 2.1 *Чем больше номер гармоники, тем сильнее в среднем изменяются при сдвиге окна значения комплексных амплитуд соответствующие этой гармонике.*

Это утверждение основано на теореме о циклическом сдвиге дискретного преобразования Фурье:

Теорема 2.3 *Если даны последовательности* $\{w_k\} = \{w_k, k = 0, 1, 2, \cdots, N-1\}, \{w'_k\} = \{w'_k, k = 0, 1, 2, \cdots, N-1\}$, *такие что*

$$w_{k+h} = w'_k, k = 0, 1, 2, \cdots, N-1, \tag{2.27}$$

и $\{X_k\} = \{X_k, k = 0, 1, 2, \cdots, N-1\}, \{X'_k\} = \{X'_k, k = 0, 1, 2, \cdots, N-1\}$ — *дискретное преобразование Фурье соответственно последовательностей* $\{w_k\}, \{w'_k\}$,*то*

$$X'_k = e^{\frac{2\pi i}{N}kh} X_k, k = 0, 1, 2, \cdots, N-1 \tag{2.28}$$

Так как $|e^{-\frac{2\pi i}{N}k}| = 1$ (рассматриваем сдвиг на -1), то $X_k * e^{-\frac{2\pi i}{N}k}$ — поворот радиус-вектора X_k на аргумент $e^{-\frac{2\pi i}{N}k}$ в комплексной плоскости. Т.к $k = 0, \cdots, [N/2]+1$, то аргумент $e^{-\frac{2\pi i}{N}k}$ изменяется от 0 до П для максимального $k = [N/2]+1$. Следовательно, для случая, когда при сдвиге окна первый элемент, который отсутствует в уже сдвинутом окне, равен новому элементу (добавляемого при сдвиге) это утверждение выполняется. Если считать размер окна достаточно большим, а разницу первого и нового элемента небольшой, то утверждение справедливо.

Таким образом, для того, чтобы соседние окна имели схожее признаковое описание, можно исключать из рассмотрения признаки соответствующие высокочастотным составляющим.

2.2.6 Применение дискретного преобразования Фурье к признаковому пространству локальных окон

Каждый объект описывается матрицейw размерами $20 \times N$ где N — длина окна.

Каждую строчку этой матрицы мы раскладываем по дискретному преобразованию Фурье. Так как исходная матрица состоит из действительных чисел, то информацию несут только $[N/2] + 1$ коэффициентов Фурье из которых 1 действительный (нулевая гармоника), остальные комплексные. Действительная и мнимая часть каждого коэффициента рассматриваются как отдельные признаки. Таким образом, для каждой строки матрицы получаем вектор длины N. Полученная матрица $\mathbf{F}$ размера $20 \times N$ разворачивается в вектор: $\mathbf{f} = (f_1^T, \ldots, f_n^T)$, где f_i — i-й столбец матрицы $\mathbf{F}$. Таким образом, в $\mathbf{f}$ коэффициенты Фурье расположены от соответствующим низкочастотным составляющим в начале к высокочастотным в конце. Количество используемых коэффициентов Фурье будет настраиваться по скользящему контролю вместе с остальными параметрами алгоритма.

2.2.7 Функции сравнения аминокислотных фрагментов на базе разложения Фурье

Суть сравнения двух аминокислотных фрагментов (ω', ω'') на базе разложения Фурье заключается в использовании вектора признаков (**??**) в рамках одного метода сравнения $S(\omega', \omega'') = S(\mathbf{f}(\omega'), \mathbf{f}(\omega''))$. Мы исследовали три функции:

$$\begin{aligned} S_4(\omega', \omega'') &= \mathbf{f}^T(\omega')\mathbf{f}(\omega''), \\ S_5(\omega', \omega'') &= \exp\left\{-\gamma||\mathbf{f}(\omega') - \mathbf{f}(\omega'')||^2\right\}, \\ S_6(\omega', \omega'') &= \sum_{k=1}^{20}\sum_{l=0}^{4}|f_{kl}(\omega') - f_{kl}(\omega'')|. \end{aligned} \tag{2.29}$$

Такие функции сравнения подходят для сравнения дальних зависимостей в предсказании вторичной структуры, длина окна $2T + 1 = 35$.

3 Многомодальный метод релевантных векторов

Мы использовали мультимодальную RVM для предсказания вторичной структуры методом скользящего окна [10].

Критерий обучения предложенной RVM:

$$\begin{cases} \sum\limits_{i=1}^{n}\sum\limits_{l=1}^{N}\left[(1-\mu)a_{il}^2 + \mu|a_{il}|\right] + C\sum\limits_{j=1}^{N}\delta_j \to \min\limits_{a_{il}, b, \delta_j}, \\ y_j\left(\sum\limits_{i=1}^{n}\sum\limits_{l=1}^{N} a_{il}S_i(\omega_l, \omega_j) + b\right) \geqslant 1 - \delta_j, \\ \delta_j \geqslant 0, j = \overline{1, N}. \end{cases} \tag{3.1}$$

Такой критерий обучения отличается от обычного SVM более сложным регуляризирующим членом, являющимся смесью L_2 и L_1 норм с весовым коэффициентом $0 \leqslant \mu \leqslant 1$. Критерий позволяет управлять селективностью по отношению к модальностям и опорным объектам.

Будем использовать следующие обозначения для, соответственно, способов сравнения объектов, объектов обучения и вторичных признаков:

Двойственная задача выпуклого программирования:

$$\begin{cases} W(\lambda_1,\ldots,\lambda_N|\mu) = \sum_{j=1}^{N}\lambda_j - \\ -\dfrac{1}{4(1-\mu)}\sum\limits_{i=1}^{n}\sum\limits_{l=1}^{N}\left\{\min\begin{bmatrix}\mu+\sum\limits_{j=1}^{N} y_j\lambda_j x_{il,j}\\ 0\\ \mu-\sum\limits_{j=1}^{N} y_j\lambda_j x_{il,j}\end{bmatrix}\right\}^2 \to \max_{\lambda_1,\ldots,\lambda_N}, \\ \sum_{j=1}^{N} y_j\lambda_j = 0,\ 0\leqslant\lambda_j\leqslant C,\ j=\overline{1,N} \end{cases}$$

Число переменных равно числу аминокислотных окон в обучающей совокупности.

Разделяющая гиперплоскость полностью определяется опорными модальностями и опорными объектами:

$$d(\omega) = \underbrace{\sum_{ij\in\hat{F}} a_{ij} S_i(\omega_j,\omega)}_{\text{только релевантные окна и функции сравнения}} + b \begin{cases} \text{strand} > 0 \\ \text{не strand} < 0 \end{cases} \tag{3.2}$$

При $\mu = 0$, метод эквивалентен классическому SVM, сохраняя все вторичные признаки $x_{ij}(\omega) = S_i(\omega_j,\omega)$, а именно, всю обучающую совокупность как набор эталонных объектов и все способы попарного сравнения (модальности) выражающиеся функциями $S_i(\omega_j,\omega)$. При росте параметра $0 \to \mu \to 1$ подмножество опорных признаков $\hat{F}$ сохраняется, и наборы опорных объектов $\hat{J}$ и модальностей $\hat{I}$ сокращаются вместе с ним. При $\mu \to 1$ критерий становится чрезвычайно селективным.

4 Вычислительный эксперимент

Чтобы определить эффективность RVM для предсказания вторичной структуры при разных значениях селективности, использовался набор, который содержит 126 белков, среди которых схожестью менее 25% обладают все белки длиной более 80 аминокислот

Из белков RS126 получается Ω из $|\Omega| = 19075$ аминокислотных фрагментов $\omega \in \Omega$ длиной $2T+1 = 35$. Каждый фрагмент отнесен к классу $y = \pm 1$, согласно тому, является ли его центральный элемент стрендом или не является. Было выполнилнено 4 эксперимента на данном наборе данных.

В каждом эксперименте множество всех окон Ω было случайным образом разбито на обучение $\Omega_{tr} \in \Omega$ размера $N = |\Omega_{tr}| = 1600$ и контроль $\Omega_{test} = \Omega \setminus \Omega_{tr}$ размера $|\Omega_{test}| = 17475$.

Множество функций срванения формируется из (2.1), (2.29), т.е. их $n = 6$. Функции сравнения, основанные на методе Фурье (2.29), учитывают $2T+1 = 35$ аминокислот каждого фрагмента, а сравнение по позициям (2.1), применяемое к более коротким фрагментам длины $2T+1 = 13$, игнорирует 11 аминокислот с обеих сторон фрагментов длины 35.

Каждый из четырех экспериментов был устроен следующим образом. Мультимодальный RVM [10] обучался на совокупности $\Omega_{tr}, N = 1600$ при семи различных

значениях селективности: $\mu \in \{0, 0.3, 0.5, 0.6, 0.8, 0.9999, 0.99999(\mu \to 1)\}$. Таким образом, обучение мультимодальной RVM запускалось $4 \times 7 = 28$ раз. Это серьезный объем вычислений для настольного компьютера.

Результат каждого обучения мультимодального RVM с конкретным значением μ есть подмножество вторичных релевантных признаков $\hat{F} = \{i, j : \hat{a}_{ij}(\mu) \neq 0\} \subseteq F$ и значений параметров $\left(a_{ij}(\mu) \in \hat{F}, b(\mu)\right)$ разделяющей гиперплоскости (3.2). Наиболее важное значение имеют множества релевантных объектов(аминокислотных фрагментов обучающей совокупности)$\hat{J}(\mu) = \{j : \exists i(a_{ij} \neq 0)\} \subseteq \{\overline{1, N}\}$ и релевантных модальностей $\hat{I}(\mu) = \{i : \exists j(a_{ij} \neq 0)\} \subseteq I = \{\overline{1, n}\}$. Обозначим мощности этих двух множеств $\hat{N}(\mu) = |\hat{J}(\mu)| \leqslant N$ и $\hat{n}(\mu) = |\hat{I}(\mu)| \leqslant n$.

Оставшееся множество $\Omega_{test} = \Omega \backslash \Omega_{tr}$ аминокислотных фрагментов было случано разбито на 10 подмножеств для контроля, после чего мы посчитали точность распознавания вторичной структуры {стренд} против {нестренд} для каждого из них. Окончательная точность для каждого из значений селективности характеризуется двумя числами: матожиданием $Acc(\mu)$ и дисперсией $\sigma(\mu)$. Доверительный интервал для точности классификации мы оценили как $Acc(\mu) \pm 2\sigma(\mu)$.

Рисунок 2 демонстрирует зависимость точности предсказания $Acc(\mu)$ и количества релевантных аминокислотных фрагментов, участвующих в решающем правиле $\hat{N}(\mu)$, от уровня селективности μ. Все результаты показаны в таблице 1.

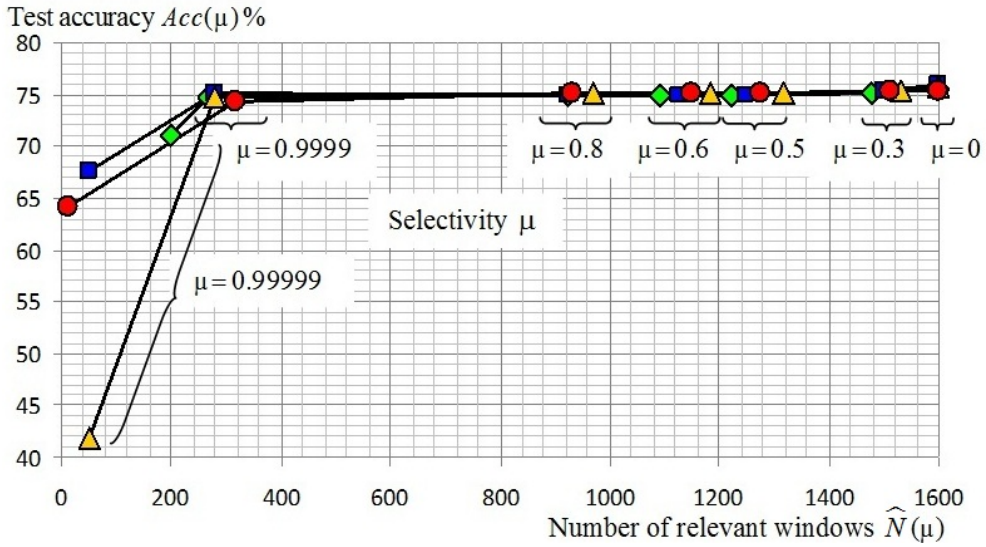

Рис. 1: Экспериментальная зависимость количества релевантных аминокислотных фрагментов $\hat{N}$ и точности распознвания стрендов на контроле Acc от уровня селективности μ.

Из рисунка видно, что во всех экспериментах наилучшая точность классификации 75.5% достигается при $\mu = 0$, т.е. когда все 1600 аминокислотных фрагментов, образующих обучающую совокупность, и все 6 функций сравнения участвуют в разделяющей гиперплоскости (3.2). Последняя соответствует пространству вторичных признаков размерностью $nN = 9600$, получаемых из фрагмента $\omega = (\alpha_\tau \in \mathbb{A}, -T \leqslant \tau \leqslant T)$. Особенно интересно, что не наблюдается эффекта переобучения после построения разделяющей гиперплоскости в линейном пространстве векторов вторичных признаков $\mathbf{x}(\omega) = \left(x_{ij}(\omega), i = \overline{1, 6}, j = \overline{1, 1600}\right) \in \mathbb{R}^{9600}$, чья размерность в 6 раз превышает количество объектов обучающей совокупности.

С ростом μ уменьшается количество релевантных фрагментов $\hat{N}(\mu)$ и количество релевантных функций сравнения $\hat{n}(\mu)$, формирующих вторичные признаки аминокислотных фрагментов. Точность остается практически на одном уровне во всех независимых экспериментах вплоть до уровня селективности $\mu = 0.9999$, когда осталось порядка 300 релевантных аминокислотных фрагментов из 1600 и только $\hat{n} = 3$ функции сравнения. Уменьшение точности по отношению к ее уровню для $\mu = 0$ не превосходит 1%.

При дальнейшем увеличении $\mu > 0.9999$ происходит значительное снижение точности распознавания на всех обучающих совокупностях.

5 Заключение

Применение методов машинного обучения к задачам биоинформатики дает эффективное направление исследований в вычислительной биологии и computer science. Математически строгие алгоритмы генерации и отбора признаков позволяют получать более высокое качество классификации. В данной работе применен метод, основанный на RVM, к предсказанию вторичной структуры белка. Важной особенностью данного метода является то, что он позволяет автоматически отбирать наиболее информативные признаки из общего множества признаков. Средняя точность распознавания стрендов оказалась равной примерно 75, что находится на уровне уже существующих алгоритмов классификации. Особенно интересно то, что не было замечено переобучения, несмотря на то, что размерность векторов вторичных признаков в несколько раз выше размера обучающей совокупности. Разработанные алгоритмы позволяют существенно уменьшить количество аминокислотных фрагментов, необходимых для предсказания вторичной структуры белка с хорошей точностью.

Список литературы

[1] *Brändén C., Tooze J.* Introduction to Protein Structure. Introduction to Protein Structure Series. — Garland Publishing, 1999. http://books.google.ru/books?id=zsWcpqrgG74C.

[2] *Rost B.* Review: Protein secondary structure prediction continues to rise // *J. Struct. Biol.* — 2001. — Vol. 134. — Pp. 204–218.

[3] *Yoo P. Zhou B. Z. A.* Machine learning techniques for protein secondary structure prediction: An overview and evaluation // *Current Bioinformatics.* — 2008. — Vol. 3. — Pp. 74–86.

[4] Critical assessment of the protein structure prediction. protein structure prediction center. sponsored by the us national library of medicine (nih/nlm).

[5] Predictions without templates: new folds, secondary structure, and contacts in CASP5. / P. Aloy, A. Stark, C. Hadley, R. B. Russell // *Proteins.* — 2003. — Vol. 53 Suppl 6. — Pp. 436–456. http://dx.doi.org/10.1002/prot.10546.

[6] *I. Y T.* Bioinformatics in the Post-Genomic Era: The Role of Biophysics. — New York: Nova Biomedical Books, 2007.

[7] *Vapnik V.* Statistical learning theory. — Wiley, 1998. — Pp. I–XXIV, 1–736.

[8] *Ward J. McGuffin L. B. B. J. D.* Secondary structure prediction with support vector machines // *Bioinformatics.* — 2003. — Vol. 19. — Pp. 1650–1655.

[9] *Bishop C. M., Tipping M. E.* Variational relevance vector machines // *CoRR.* — 2013. — Vol. abs/1301.3838.

[10] Convex support and relevance vector machines for selective multimodal pattern recognition / O. Seredin, V. Mottl, A. Tatarchuk et al. // ICPR. — 2012. — Pp. 1647–1650.

[11] *Ni Y., Niranjan M.* Exploiting long-range dependencies in protein beta-sheet secondary structure prediction. // PRIB / Ed. by T. Dijkstra, E. Tsivtsivadze, E. Marchiori, T. Heskes. — Vol. 6282 of *Lecture Notes in Computer Science.* — Springer, 2010. — Pp. 349–357.

[12] *Engel DE D. W.* Amino acid propensities are position-dependent throughout the length of a-helices // *J. Mol. Biol.* — 2004. — Vol. 337. — Pp. 1195–1205.

[13] *Dayhoff M. O., Schwartz R. M.* Chapter 22: A model of evolutionary change in proteins // in Atlas of Protein Sequence and Structure. — 1978.

[14] Probabilistic evolutionary model for substitution matrices of pam and blosum families.: Tech. rep. / V. Sulimova, V. Mottl, I. Muchnik, C. Kulikowski: Center for Discrete Mathematics and Theoretical Computer Science, 2008.

			Точность определения стрендов $Acc(\mu)$	Количество релевантных окон $\hat{N}(\mu)$	Количество $\hat{n}(\mu)$ и список релевантных функций
Эксперимент 1 ◆	Селективность μ	0	75.63 ± 1.78%	1600	$6, \hat{I} = \{1,2,3,4,5,6\}$
		0.3	75.04 ± 1.75%	1476	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.5	74.95 ± 1.74%	1222	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.6	74.96 ± 1.74%	1094	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.8	74.96 ± 1.72%	924	$4, \hat{I} = \{1,2,\not{3},4,5,\not{6}\}$
		0.9999	74.63 ± 1.76%	267	$3, \hat{I} = \{1,2,\not{3},\not{4},5,\not{6}\}$
		0.99999	71.04 ± 1.69%	200	$2, \hat{I} = \{1,2,\not{3},\not{4},\not{5},\not{6}\}$
Эксперимент 2 ■	Селективность μ	0	75.85 ± 1.52%	1600	$6, \hat{I} = \{1,2,3,4,5,6\}$
		0.3	75.23 ± 1.72%	1501	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.5	75.01 ± 1.63%	1247	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.6	75.01 ± 1.65%	1127	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.8	75.01 ± 1.65%	924	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.9999	75.10 ± 1.73%	278	$3, \hat{I} = \{1,2,\not{3},\not{4},5,\not{6}\}$
		0.99999	67.60 ± 0.80%	49	$3, \hat{I} = \{1,2,\not{3},\not{4},5,\not{6}\}$
Эксперимент 3 ▲	Селективность μ	0	75.70 ± 1.22%	1600	$6, \hat{I} = \{1,2,3,4,5,6\}$
		0.3	75.30 ± 0.79%	1531	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.5	75.10 ± 0.94%	1317	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.6	75.08 ± 0.99%	1183	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.8	75.08 ± 0.99%	971	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.9999	74.74 ± 0.79%	280	$3, \hat{I} = \{1,2,\not{3},\not{4},5,\not{6}\}$
		0.99999	41.84 ± 2.33%	51	$3, \hat{I} = \{1,2,\not{3},\not{4},5,\not{6}\}$
Эксперимент 4 ●	Селективность μ	0	75.33 ± 0.99%	1600	$6, \hat{I} = \{1,2,3,4,5,6\}$
		0.3	75.30 ± 0.95%	1514	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.5	75.07 ± 0.97%	1275	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.6	75.03 ± 0.99%	1150	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.8	75.03 ± 0.99%	933	$5, \hat{I} = \{1,2,\not{3},4,5,6\}$
		0.9999	74.27 ± 1.53%	318	$3, \hat{I} = \{1,2,\not{3},\not{4},5,\not{6}\}$
		0.99999	64.16 ± 1.81%	12	$3, \hat{I} = \{1,2,\not{3},\not{4},5,\not{6}\}$

Рис. 2: Таблица 1. Результаты четырех независимых экспериментов (обозначения как на Рис. 1).