На правах рукописи

Сулимова Валентина Вячеславовна

ПОТЕНЦИАЛЬНЫЕ ФУНКЦИИ ДЛЯ АНАЛИЗА СИГНАЛОВ И СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ РАЗНОЙ ДЛИНЫ

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва 2009

Работа выполнена в Тульском государственном университете
на кафедре автоматики и телемеханики

Научный руководитель

доктор технических наук, профессор Вадим Вячеславович Моттль

Официальные оппоненты

доктор физико-математических наук Михаил Николаевич Устинин,

доктор технических наук, профессор Леонид Моисеевич Местецкий

Ведущая организация
Институт проблем управления РАН

Защита диссертации состоится « __ » _____ 2009 г. в __ ч. на заседании диссертационного совета Д 002.017.02 в учреждении Российской академии наук Вычислительный центр им. АА. Дородницына РАН по адресу: 119333, г. Москва, ул. Вавилова, 40.

С диссертацией можно ознакомиться в библиотеке ВЦ РАН.

Автореферат разослан « ___» _____ 2009 г.

Ученый секретарь
диссертационного совета
доктор физико-математических наук В.В. Рязанов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Сигналы и символьные последовательности являются наиболее распространенными видами организации данных. Широко известны такие задачи, как задача распознавания подписей по их динамическим характеристикам (on-line signatures), речевых команд и слитной речи, биологических свойств и пространственной структуры полимерных молекул белков по составляющим их последовательностям над алфавитом двадцати существующих в природе аминокислот.

Все перечисленные задачи имеют важную характерную особенность – в них сигналы или символьные последовательности, представляющие реальные объекты, в общем случае имеют различную длину. В результате, оказывается трудным заранее указать фиксированное число признаков, которые смогли бы сформировать пространство, удовлетворяющее гипотезе компактности, лежащей в основе классических методов классификации объектов.

Вообще говоря, проблеме представления векторных и символьных последовательностей разной длины в алгоритмах анализа данных посвящена обширная литература. Наиболее популярным является беспризнаковый подход, основанный на измерении парного сходства последовательностей путем вычисления потенциальной функции, т.е. двухместной симметрической действительной функции, образующей неотрицательно определенную матрицу для любой конечной совокупности объектов. В результате множество всех последовательностей разной длины оказывается погруженным в гипотетическое линейное пространство со скалярным произведением, роль которого играет сама потенциальная функция. Такая математическая конструкция позволяет применять хорошо разработанные линейные методы анализа данных к совокупностям объектов произвольной природы.

В то же время, методология формирования потенциальных функций над множествами последовательностей разной длины еще далека от завершения и требует дальнейшего развития.

Первая проблемная ситуация заключается в том, что для большинства прикладных задач потенциальная функция на множестве последовательностей отвечает практическим целям анализа данных только в том случае, если она основана на некоторой элементарной потенциальной функции на множестве примитивов. Такое требование естественным образом выполняется для векторных сигналов, в качестве примера которых в данной диссертации рассматриваются динамические подписи.

Что же касается символьных последовательностей, то подавляющее большинство публикаций на эту тему ориентировано на анализ биологических полимеров, в частности, аминокислотных последовательностей белков. Именно этот вид символьных последовательностей находится в центре внимания в данной диссертации. В современной биохимии общепринятым способом измерения сходства аминокислот являются подстановочные матрицы PAM (Point Accepted Mutation) и BLOSUM (BLock SUbstitution Matrix), которые в традиционной форме не являются потенциальными функциями. С этой точки зрения соответствующие способы построения потенциальных функций на множествах символьных последовательностей разной длины являются эвристическими.

Вторая проблемная ситуация порождена тем обстоятельством, что наличие формальных свойств скалярного произведения у формируемой меры сходства аминокислотных последовательностей, как правило, оказывается недостаточным для ее эффективного использования при решении задач классификации белков на семейства, обладающие сходными биологическими функциями. Центральной гипотезой биоинформатики, многократно подтвержденной на практике, является предположение, что эволюционно близкие белки выполняют похожие биологические функции в организме. В связи с этим специалисты в области молекулярной биологии крайне недоверчиво относятся к мерам сходства белков, значения которых не могут быть интерпретированы как меры их эволюционной близости. Однако ни один из известных способов формирования потенциальных функций на множестве аминокислотных последовательностей не является одновременно математически корректным и обоснованным с точки зрения вероятностной модели эволюции белков.

Кроме того, несмотря на то, что дискретные сигналы и символьные последовательности имеют похожую структуру, в настоящее время не существует единого математического аппарата для их сравнения. Единственный корректный способ построения потенциальных функций на множестве сигналов разной длительности, предложенный французским ученым Ж.-Ф. Вертом, имеет ряд неестественных ограничений, дополнительно введенных по сравнению с аналогичным подходом, разработанным для символьных последовательностей.

Наконец, **третья проблемная ситуация**, выбранная для исследования в данной диссертации, порождена необходимостью интерпретации результатов классификации последовательностей разной длины, полученной тем или иным алгоритмом, основанным на их погружении в линейное пространство путем введения потенциальной функции. Всякая конечная совокупность последовательностей, выделенная в качестве класса, имеет естественную модель в виде его центра, т.е. гипотетического среднего объекта в линейном замыкании всех последовательностей. Можно доказать, что результат усреднения конечного множества последовательностей разной длины, в частности, аминокислотных последовательностей белков, в смысле линейных операций, определяемых некоторой потенциальной функцией, в общем случае не будет являться последовательностью конечной длины.

Для разрешения первой проблемной ситуации относительно аминокислотных последовательностей белков в диссертации используется тот факт, что обе общепринятые меры сходства аминокислот, как PAM, так и BLOSUM, численно выражают правдоподобие гипотезы об общем происхождении двух указанных аминокислот от одной неизвестной аминокислоты в результате двух независимых шагов эволюции. Такая двухместная функция на алфавите аминокислот всегда является потенциальной функцией по своей структуре. Практически используемые подстановочные матрицы PAM и BLOSUM не являются положительно определенными только в силу логарифмического представления результата, к тому же округленного до целого значения.

Разрешение второй проблемной ситуации основано на идее прямого переноса принципа измерения эволюционного сходства аминокислот на аминокислотные последовательности в целом. Потенциальные функции предлагается строить как функции правдоподобия гипотезы, что две заданные последовательности получены из общего неизвестного прародителя в результате двух независимых ветвей эволюции. Разные потенциальные функции на множестве символьных последовательностей, рассматриваемые в диссертации, отличаются друг от друга только разными предположениями о множестве допустимых прародителей и априорном распределении вероятностей на нем, а также разными вероятностными моделями эволюционных преобразований, сводящихся к случайным вставкам, удалениям и заменам символов в исходной последовательности.

Что касается сигналов, то, хотя прикладные задачи их анализа не требуют измерения именно эволюционного сходства, предложенная вероятностная концепция не противоречит их природе. Потенциальные функции на множестве сигналов строятся по тому же принципу и отличаются только спецификой случайных преобразований, в которых вместо вставок и удалений элементов фигурируют локальные сжатия и растяжения оси времени.

В качестве теоретической основы **разрешения третьей проблемной ситуации** предлагается постановка задачи поиска общего прародителя фиксированной длины n для группы последовательностей путем максимизации правдоподобия гипотезы об их случайном независимом происхождении из скрытого общего прародителя известной длины n в результате предложенных в данной работе случайных преобразований. При этом последовательность-прародитель предлагается искать в виде совокупности независимых распределений его элементов, что соответствует общепринятому в биоинформатике понятию профиля.

Цель работы. Целью диссертационной работы является разработка методов построения и алгоритмов вычисления потенциальных функций на множествах сигналов и символьных последовательностей разной длины, позволяющих погрузить исходное множество объектов в соответствующее гипотетическое линейное пространство со скалярным произведением, адекватное решаемой типовой задаче классификации сигналов либо символьных последовательностей.

Задачи исследования. Для достижения поставленной цели в диссертации сформулированы и решены следующие задачи:

1. Разработка вероятностного принципа построения потенциальных функций на конечном алфавите элементов последовательностей на основе марковской модели их случайных преобразований.
2. Построение потенциальных функций на множестве аминокислот на основе модели эволюции М. Дэйхофф.
3. Построение моделей случайного преобразования на множествах сигналов и символьных последовательностей разной длительности.
4. Разработка вероятностного принципа формирования потенциальных функций на множествах сигналов и символьных последовательностей разной длительности.
5. Разработка алгоритмов вычисления потенциальных функций на множествах сигналов и символьных последовательностей.
6. Разработка методов наглядного представления об общем прародителе для заданной совокупности последовательностей разной длины.
7. Использование потенциальных функций для автоматической классификации белков на функциональные семейства и для верификации личности по динамике подписи.

Методы исследования. Исследование базируется на использовании теории распознавания образов, теории линейных пространств со скалярным произведением, теории марковских случайных процессов.

Научная новизна. В работе предложены новые вероятностные модели случайных преобразований сигналов и символьных последовательностей, в частности, модели эволюционных изменений аминокислотных последовательностей белков. На основе этих моделей впервые построен класс корректных потенциальных функций, выражающих правдоподобие гипотезы о наличии общего прародителя у пары сравниваемых сигналов либо символьных последовательностей разной длины. Впервые доказано, что меры сходства аминокислот PAM и BLOSUM, общепринятые в современной биоинформатике, основаны на одной и той же модели эволюции аминокислот, разработанной Маргарет Дэйхофф, и по своей структуре являются потенциальными функциями. Впервые поставлена и решена задача поиска общего прародителя заданной длины для группы последовательностей в терминах введенного случайного преобразования.

Положения, выносимые на защиту

1. Семейство потенциальных функций на алфавите аминокислот, выражающее смысл общепринятых подстановочных матриц PAM и BLOSUM.
2. Комплекс вероятностных моделей случайных преобразований сигналов и символьных последовательностей.
3. Класс корректных потенциальных функций, выражающих правдоподобие гипотезы о наличии общего прародителя у пары сравниваемых сигналов либо символьных последовательностей разной длины.
4. Алгоритмы вычисления потенциальных функций на множествах сигналов и символьных последовательностей разной длины.
5. Задача поиска общего прародителя заданной длины для группы последовательностей.

Достоверность полученных результатов подтверждается доказательствами сформулированных в диссертации теорем и результатами решения прикладных задач.

Практическая значимость. Разработанные принципы и алгоритмы позволяют корректно использовать методы анализа данных, разработанные для линейных признаковых пространств, для решения задач классификации сигналов и символьных последовательностей разной длины, в которых трудно заранее сформировать пространство достаточно информативных числовых характеристик объектов.

Связь с плановыми научными исследованиями. Работа выполнена при поддержке грантов Российского фонда фундаментальных исследований №№ 05-01-00679-а, 06-01-08042-офи, 08-01-00695-а и 08-01-12023-офи, а также грантов INTAS № 04-77-7347 и Young Scientist PhD Fellowship № 06-1000014-6563.

Реализация и внедрение результатов работы. Результаты исследования применены для решения задачи автоматической классификации белков по составляющим их последовательностям аминокислот на классы белков, выполняющих сходные биологические функции в организме, и задачи верификации личности по динамике подписи.

Апробация работы. Основные положения и результаты диссертации докладывались на конференциях: «Математические методы распознавания образов» (Пущино, 2003 г., Звенигород, 2005 г., Зеленогорск 2007 г.), «Распознавание образов и анализ изображений» (Санкт-Петербург, 2004), «Интеллектуализация обработки информации» (Алушта, Крым, 2004, 2006, 2008 гг.), «Обработка сигналов и изображений» (IASTED SIP-2006, Гонолулу, Гавайи, 2006 г.), «Международная конференция по распознаванию образов» (ICPR-2008, Флорида, США, 2008 г.), на семинарах партнеров по гранту INTAS «Принципы распознавания сигналов, символьных последовательностей и изображений на основе измерения их несходства» в Москве (2005 г.), в Гилдфорде, Великобритания (2005 г.), в Праге, Чехия (2006 г.) и Киеве, Украина (2007 г.), на семинаре по анализу данных, Биркбек колледж, Лондон, Великобритания, 2008 г., на Международном симпозиуме по исследованиям в области биоинформатики и ее приложениям ISBRA-2009, Флорида, США, 2009 г.

Публикации. По тематике исследований опубликовано 11 статей, в том числе 2 статьи в журналах, рекомендованных ВАК.

Структура и объем работы. Диссертация состоит из введения, 5 глав, основных выводов и списка литературы. Материал изложен на 121 страницах, содержит 22 рисунка, 6 таблиц и список литературы из 118 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность исследований по построению потенциальных функций для решения задач анализа сигналов и символьных последовательностей разной длины, сформулированы цели и задачи проводимых исследований, положения, выносимые на защиту, приведены сведения о структуре диссертации, ее апробации и практическом использовании полученных результатов.

В первой главе приведены примеры типичных прикладных задач анализа сигналов и символьных последовательностей разной длины, такие как задача верификации личности по динамике подписи и задача агрегации белков в соответствии с биологическими функциями по составляющим их последовательностям аминокислотных остатков. Анализ приведенных задач показывает, что наиболее подходящим путем их решения является беспризнаковый подход, в котором принятие решения о классификации объекта принимается на основе информации о его сходстве с другими объектами рассматриваемого множества, минуя явное вычисление векторов их признаков.

Обоснован выбор метода потенциальных функций в качестве основной методологической концепции для беспризнакового измерения парного сходства динамических подписей и аминокислотных последовательностей белков.

Показана недостаточность существующих методов измерения парного сходства сигналов и символьных последовательностей разной длины и сформулирован общий для них принцип формирования потенциальных функций.

Первая глава завершается изложением основных задач исследования, сформулированных на основе проведенного анализа выявленных проблем.

Во второй главе рассмотрены пути введения потенциальных функций на множестве примитивов $\omega \in A$, из которых формируются последовательности $\boldsymbol{\omega} = (\omega_t, t = 1,...,N)$.

Для случая сигналов, элементами которых являются, как правило, векторы действительных чисел $\omega_t = \mathbf{x}_t = (x_t^1 \cdots x_t^m)^T \in A = R^m$, простейший путь введения потенциальной функции в R^m основан на использовании естественных линейных операций в конечномерном линейном пространстве примитивов и полностью исчерпывается множеством скалярных произведений $\mu(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \mu(\mathbf{x}', \mathbf{x}'') = \mathbf{x}'^T \mathbf{Q} \mathbf{x}''$ в R^m, где $\mathbf{Q}$ - любая невырожденная матрица.

Однако так определенная потенциальная функция не является мерой сходства векторов $\mathbf{x}', \mathbf{x}'' \in R^m$, под которой естественно понимать некоторую величину, монотонно убывающую с увеличением евклидова расстояния $\rho(\mathbf{x}', \mathbf{x}'') = \mathbf{x}'^T\mathbf{Q}\mathbf{x}' + \mathbf{x}''^T\mathbf{Q}\mathbf{x}'' - 2\mathbf{x}'^T\mathbf{Q}\mathbf{x}''$. Таким свойством обладают так называемые радиальные потенциальные функции[1] $\mu(\mathbf{x}', \mathbf{x}'') = \exp\left(-\gamma\rho^2(\mathbf{x}', \mathbf{x}'')\right)$.

Пути введения потенциальных функций на множестве примитивов символьных последовательностей рассматриваются в данной диссертации только применительно к аминокислотным последовательностям белков, т.е. последовательностям над алфавитом аминокислот $A = \{a^1, ..., a^m\}$, $m = 20$.

В качестве теоретической концепции сравнения аминокислот в диссертации принята вероятностная модель эволюции Маргарет Дэйхофф, называемая PAM[2]. Ее основным математическим понятием является понятие марковской цепи эволюции аминокислот в отдельно взятой точке цепи, определяемой матрицей переходных вероятностей $\mathbf{\Psi} = \left(\psi(\alpha^j \mid \alpha^i)\right)$ превращения аминокислоты α^i в аминокислоту α^j на следующем шаге эволюции. При этом предполагается, что эта марковская цепь представляет собой эргодический и обратимый случайный процесс, т.е. процесс, характеризующийся финальным распределением вероятностей $\sum_{\alpha^i \in A} \xi(\alpha^i)\psi(\alpha^j \mid \alpha^i) = \xi(\alpha^j)$ и удовлетворяющий условию обратимости $\xi(\alpha^i)\psi(\alpha^j \mid \alpha^i) = \xi(\alpha^j)\psi(\alpha^i \mid \alpha^j)$.

Марковский процесс эволюции, наблюдаемый с любым шагом s, также является марковским процессом. Этот процесс определяется многошаговой матрицей $\mathbf{\Psi}_{[s]} = \underbrace{\mathbf{\Psi} \times ... \times \mathbf{\Psi}}_{s}$ переходных вероятностей.

Теорема 1. *Двухместная функция над алфавитом аминокислот*

$$\mu_{[s]}(\alpha^i, \alpha^j) = \psi_{[s]}(\alpha^j \mid \alpha^i)\xi(\alpha^i) = \psi_{[s]}(\alpha^i \mid \alpha^j)\xi(\alpha^j) \tag{1}$$

и нормированная на финальные вероятности функция

$$\tilde{\mu}_{[s]}(\alpha^i, \alpha^j) = \mu_{[s]}(\alpha^i, \alpha^j) / \xi(\alpha^i)\xi(\alpha^j) = \psi_{[s]}(\alpha^j \mid \alpha^i) / \xi(\alpha^j) = \psi_{[s]}(\alpha^i \mid \alpha^j) / \xi(\alpha^i) \tag{2}$$

являются потенциальными функциями для любой степени s.

Утверждение этой теоремы используется в диссертации для математической интерпретации широко используемой в биоинформатике серии подстановочных матриц PAM с разным значением эволюционного шага s, которые изначально имеют ту же структуру, что и функция $\tilde{\mu}_{[s]}(\alpha^i, \alpha^j)$, но используются традиционно в логарифмической форме $d^{ij}_{[s]} = 10 \log_{10} \tilde{\mu}_{[s]}(\alpha^i, \alpha^j)$ с последующим округлением до целых, что приводит к неизбежной потере свойств потенциальной функции.

В последующих главах диссертации для измерения сходства аминокислот используются именно исходные значения $\mu_{[s]}(\alpha^i, \alpha^j)$ и $\tilde{\mu}_{[s]}(\alpha^i, \alpha^j)$.

Обладая свойствами потенциальной функции на алфавите аминокислот $A = \{\alpha^1, ..., \alpha^m\}$, $m = 20$, каждая из мер сходства $\mu_{[s]}(\alpha^i, \alpha^j)$ или $\tilde{\mu}_{[s]}(\alpha^i, \alpha^j)$ погружает его в гипотетическое двадцатимерное линейное пространство $\tilde{A}_s \supset A$. В качестве полного линейно независимого базиса удобно принять сам исходный алфавит, интерпретируемый как конечное подмножество точек. Совокупность, вообще говоря, воображаемых элементов линейного пространства $\tilde{\alpha} = \sum_{i=1}^{m} c^i \alpha^i \in \tilde{A}_s$ естественно интерпретировать как некоторые «обобщенные» аминокислоты. В частности, если $c^i \geq 0$ и $\sum_{i=1}^{m} c^i = 1$, то обобщенная аминокислота $\tilde{\alpha}$ имеет смысл вероятностной смеси реальных аминокислот с вектором вероятностей $(c^1 \cdots c^m)$.

---

[1] Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. М.: Наука, 1970, 384 с.

[2] Dayhoff M.O., Schwartz R.M., Orcutt B.C. A model for evolutionary change in proteins. In: Atlas for Protein Sequence and Structure (M.O. Dayhoff, ed.), 1978, Vol. 5, pp. 345-352.

Еще один принципиально важный для современной биоинформатики результат, представленный во второй главе, касается соотношения двух наиболее популярных мер сходства аминокислот, выраженных семействами подстановочных матриц PAM и BLOSUM.

Авторы меры сходства BLOSUM Джорджия и Стивен Хеникофф получили ее с использованием исключительно статистического подхода, примененного к результатам множественного выравнивания групп аминокислотных последовательностей, в отличие от меры сходства PAM, полученной на основе анализа филогенетических деревьев над группами эволюционно близких белков и базирующейся на марковской модели эволюции аминокислот.

В данной главе доказано, что мера сходства BLOSUM также имеет эволюционное обоснование, более того, она может быть выражена в терминах той же модели эволюции PAM, имеет ту же структуру, и, соответственно, является потенциальной функцией на множестве аминокислот. Единственное различие между данными мерами сходствами определяется исключительно разными исходными данными, на основе которых производится оценка параметров эволюционной модели.

Третья глава посвящена изложению достаточно универсального вероятностного принципа формирования потенциальных функций на множестве сигналов и символьных последовательностей разной длины, предлагаемого в диссертации.

Основная идея построения потенциальной функции

Пусть Ω есть множество всех конечных последовательностей $\boldsymbol{\omega} = (\omega_t, t = 1,...,N)$ над конечномерным линейным пространством примитивов $\omega_t \in \tilde{A}$. В частности, роль элементов последовательностей играют конечномерные векторы $\omega_t = \mathbf{x}_t = (x_t^1 \cdots x_t^m)^T \in A = \tilde{A} = R^m$ в случае сигналов, и обобщенные аминокислоты $\omega_t = \tilde{\alpha}_t = (c_t^1 \cdots c_t^m)^T \in \tilde{A} \supset A$ в случае аминокислотных последовательностей.

Введем также специальные обозначения $\Omega_n = \{\boldsymbol{\omega} = (\omega_t, t = 1,...,N), \omega_t \in \tilde{A}, N = n\}$ для множества всех последовательностей длины n и $\Omega_{\geq n} = \{\boldsymbol{\omega} = (\omega_t, t = 1,...,N), \omega_t \in \tilde{A}, N \geq n\}$ для множества всех последовательностей длин не менее n.

Сходство двух последовательностей $\boldsymbol{\omega}', \boldsymbol{\omega}'' \in \Omega$ будем оценивать значением правдоподобия гипотезы об их общем происхождении в результате двух независимых реализаций фиксированного случайного преобразования $\left(\varphi_n(\boldsymbol{\omega} \mid \vartheta), \boldsymbol{\omega} \in \Omega\right)$ одной и той же неизвестной последовательности $\vartheta = (\vartheta_i \in A, i = 1,...,n) \in \Omega_n \subseteq \Omega$ случайной конечной длины n с распределением $r(n)$, играющей роль общего прототипа и случайно выбранной из конечномерного линейного пространства Ω_n с некоторым распределением $(p_n(\vartheta), \vartheta \in \Omega_n)$:

$$\mathcal{K}(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum\nolimits_{n=0}^{\infty} r(n) \int_{\vartheta \in \Omega_n} p_n(\vartheta) \varphi_n(\boldsymbol{\omega}' \mid \vartheta) \varphi_n(\boldsymbol{\omega}'' \mid \vartheta) d\vartheta . \qquad (3)$$

Теорема 2. *При любом выборе распределения случайной длины общего прародителя $r(n)$, семейства распределений $\left\{\left(p_n(\vartheta), \vartheta \in \Omega_n\right), n = 1,2,...\right\}$ и семейства условных распределений $\left\{\left(\varphi_n(\boldsymbol{\omega} \mid \vartheta), \boldsymbol{\omega} \in \Omega\right), \vartheta \in \Omega_n\right\}$ функция $\mathcal{K}(\boldsymbol{\omega}', \boldsymbol{\omega}'')$ является потенциальной функцией на Ω.*

Доказательство теоремы сводится к доказательству выполнения условий Мерсера[1] и почти очевидно.

Модель случайного преобразования последовательностей

В качестве семейства случайных преобразований $\{(\varphi_n(\boldsymbol{\omega} \mid \vartheta), \boldsymbol{\omega} \in \Omega), \vartheta \in \Omega_n\}$ в диссертации рассматриваются только двухэтапные преобразования следующей структуры.

Этап 1. Всякая конечная последовательность непересекающихся интервалов

$$\mathbf{v} = \left(v_1,...,v_n\right),\ v_i = \left\{\dot{v}_i \leq t \leq \ddot{v}_i\right\},\ 1 \leq \underbrace{\dot{v}_1 \leq \ddot{v}_1}_{v_1} < \underbrace{\dot{v}_2 \leq \ddot{v}_2}_{v_2} < ... < \underbrace{\dot{v}_n \leq \ddot{v}_n}_{v_n}, \qquad (4)$$

определяет структуру случайного преобразования последовательности-прародителя $\vartheta = (\vartheta_1,...,\vartheta_n)$. Эту структуру будем называть односторонним выравниванием совокупности ее позиций $(1,...,n)$ и позиций гипотетической формируемой последовательности $(1,2,3,...)$. Счет-

[1] Mercer T. Functions of positive and negative type and their connection with the theory of integral equations. Trans. London. Philos. Soc., 1999, A, 209, 415-416.

ное множество всех односторонних выравниваний будем обозначать $V_n = \{\mathbf{v} = (v_1,...,v_n)\}$, а распределение вероятностей на нем – $q_n(\mathbf{v}) \geq 0$, $\mathbf{v} \in V_n$.

Этап 2. Для всякой структуры $\mathbf{v} \in V_n$ определено случайное преобразование, зависящее от структуры $\eta_n(\boldsymbol{\omega} \mid \vartheta, \mathbf{v}) \geq 0$, причем $\eta_n(\boldsymbol{\omega} \mid \vartheta, \mathbf{v}) = 0$ для всех последовательностей $\boldsymbol{\omega} \notin \Omega_{\geq \ddot{v}_n}$ длины меньше $\ddot{v}_n$, т.е $\int_{\boldsymbol{\omega} \in \Omega_{\geq \ddot{v}_n}} \eta_n(\boldsymbol{\omega} \mid \vartheta, \mathbf{v}) = 1$. Таким образом, длина формируемой последовательности всегда оказывается больше или равной длине последовательности-прародителя, т.е. $\boldsymbol{\omega} \in \Omega_{\geq n}$.

В итоге, случайное преобразование $\{(\varphi_n(\boldsymbol{\omega} \mid \vartheta), \boldsymbol{\omega} \in \Omega_{\geq n}), \vartheta \in \Omega_n\}$ есть смесь

$$\varphi_n(\boldsymbol{\omega} \mid \vartheta) = \sum\nolimits_{\mathbf{v} \in V_n} q_n(\mathbf{v}) \eta_n(\boldsymbol{\omega} \mid \vartheta, \mathbf{v}). \qquad (5)$$

Ключевая и дополнительная подпоследовательности

Одностороннее выравнивание (4), определяющее структуру случайного преобразования последовательности-прародителя $\vartheta = (\vartheta_i, i = 1,...,n) \in \Omega_n$ в последовательность не меньшей длины $\boldsymbol{\omega} = (\omega_t, t = 1,...,N) \in \Omega_{\geq n}$, понимается как прямое перечисление интервалов $\mathbf{v} = (v_1,...,v_n)$, в которые будут отображаться элементы $(\vartheta_1,...,\vartheta_n)$ исходной последовательности (рис. 1). Совокупность соответствующих фрагментов $\boldsymbol{\omega}_{v_i} = (\omega_{\dot{v}_i},...,\omega_{\ddot{v}_i})$ в формируемой последовательности $\boldsymbol{\omega} = (\omega_t, t = 1,...,N \geq t_n)$ будем называть *ключевой подпоследовательностью* $\bar{\boldsymbol{\omega}}_v = (\boldsymbol{\omega}_{v_i}, i = 1,...,n)$ в составе результирующей последовательности. Ключевая подпоследовательность может иметь любую длину, поскольку интервалы v_i могут быть, вообще говоря, сколь угодно длинными, т.е. $\bar{\boldsymbol{\omega}}_{\mathbf{v}} \in \Omega$. Совокупность остальных позиций будем называть *дополнительной подпоследовательностью* $\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}} = (\omega_t, t = 1,...,N, t \notin v_i, i = 1,...,n)$, которая также может иметь любую длину $\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}} \in \Omega$. Последовательность в целом есть объединение ключевой и дополнительной подпоследовательностей $\boldsymbol{\omega} = \bar{\boldsymbol{\omega}}_{\mathbf{v}} \bigcup \bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}}$.

● – *ключевая подпоследовательность* ($\bar{\boldsymbol{\omega}}$)

○ – *дополнительная подпоследовательность* ($\bar{\bar{\boldsymbol{\omega}}}$)

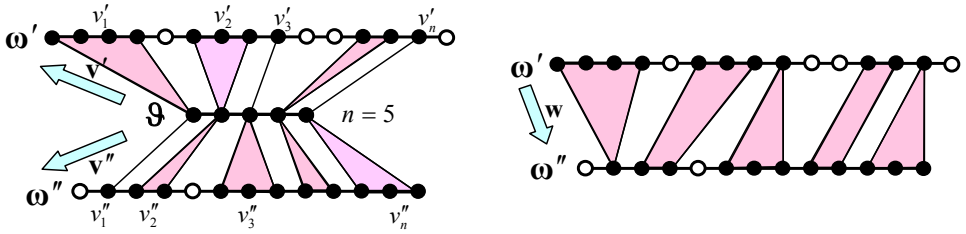

Рисунок 1 – Структура преобразования последовательностей

Независимость ключевой подпоследовательности от дополнительной

Случайное преобразование $\eta_n(\boldsymbol{\omega} \mid \vartheta, \mathbf{v})$, зависящее от структуры, построим как комбинацию двух независимых распределений вероятностей:

$$\eta_n(\boldsymbol{\omega} \mid \vartheta, \mathbf{v}) = \bar{\eta}_n(\bar{\boldsymbol{\omega}}_{\mathbf{v}} \mid \vartheta, \mathbf{v}) \bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}}), \qquad (6)$$

причем распределение на множестве дополнительных подпоследовательностей $\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}} \in \Omega$ определим как не зависящее ни от исходной последовательности-прародителя, ни от структуры преобразования $\mathbf{v} \in V_n$, и, таким образом, не привязанное к предполагаемой длине n последовательности-прародителя.

Независимость элементов последовательности-прародителя

В данной работе распределение $(p_n(\vartheta), \vartheta \in \Omega_n)$ на множестве вариантов последовательности-прототипа $\vartheta = (\vartheta_i \in \tilde{A}, i = 1,...,n) \in \Omega_n \subseteq \Omega$ принятой длины n рассматривается как совокупность одинаковых независимых распределений ее элементов $\xi(\vartheta)$:

$$p_n(\vartheta) = p_n(\vartheta_1,...,\vartheta_n) = \prod\nolimits_{i=1}^{n} \xi(\vartheta_i). \qquad (7)$$

Аналогично, будем полагать, что для всех интервалов T_i в составе выравнивания $v = (v_1,...,v_n)$ определены независимые элементарные случайные преобразования $(\eta_{v_i}(\boldsymbol{\omega}_{v_i} \mid \vartheta_i), \boldsymbol{\omega}_{v_i} \in \Omega_{|v_i|})$ исходного элемента ϑ_i в соответствующий фрагмент формируемой последовательности $\boldsymbol{\omega}_{v_i} = (\omega_{\dot{v}_i},...,\omega_{\ddot{v}_i})$ длины $|v_i|$. Соответственно, ключевая подпоследовательность фрагментов фор-

мируемой последовательности $\bar{\boldsymbol{\omega}}_{\mathbf{v}} = (\boldsymbol{\omega}_{v_i}, i = 1,...,n)$ образована совокупностью таких независимых распределений

$$\bar{\eta}_n(\bar{\boldsymbol{\omega}}_{\mathbf{v}} \mid \vartheta, \mathbf{v}) = \prod_{i=1}^{n} \eta_{v_i}(\boldsymbol{\omega}_{v_i} \mid \vartheta_i). \tag{8}$$

В свою очередь, каждое из этих преобразований представляет собой совокупность независимых одинаковых условных распределений $\psi(\omega_t \mid \vartheta_i)$:

$$\eta_{v_i}(\boldsymbol{\omega}_{v_i} \mid \vartheta_i) = \prod_{t=\dot{v}_i}^{\ddot{v}_i} \psi(\omega_t \mid \vartheta_i). \tag{9}$$

Общая структура потенциальной функции

Каждая пара односторонних выравниваний $\mathbf{v}' = (v'_i, i = 1,...,n) \in V_n$ и $\mathbf{v}'' = (v''_i, i = 1,...,n) \in V_n$ порядка n, определяющих структуры преобразований $\vartheta \to \boldsymbol{\omega}'$ и $\vartheta \to \boldsymbol{\omega}''$, $\vartheta \in \Omega_n$, $\boldsymbol{\omega}', \boldsymbol{\omega}'' \in \Omega_{\geq n}$ определяет сквозное парное выравнивание того же порядка:

$$\mathbf{w} = (\mathbf{v}', \mathbf{v}'') = \left[\begin{pmatrix} v'_1 \\ v''_1 \end{pmatrix}, ..., \begin{pmatrix} v'_n \\ v''_n \end{pmatrix} \right].$$

Наоборот, выравнивание $\mathbf{w}$ определяет пару односторонних выравниваний $(\mathbf{v}'_{\mathbf{w}}, \mathbf{v}''_{\mathbf{w}})$. Множество парных выравниваний порядка n есть декартово произведение $W_n = V_n \times V_n$.

Распределение вероятностей $q_n(\mathbf{v})$ на V_n образует распределение на W_n:

$$q_n(\mathbf{w}) = q_n(\mathbf{v}'_{\mathbf{w}}) q_n(\mathbf{v}''_{\mathbf{w}}). \tag{10}$$

Необходимо отметить, что не любые парные выравнивания могут определить пару последовательностей $\boldsymbol{\omega}'$ и $\boldsymbol{\omega}''$ длин N' и N'', а только такие, которые удовлетворяют условиям $v'_{n,\mathbf{w}} \leq N'$ и $v''_{n,\mathbf{w}} \leq N''$. Множество таких допустимых парных выравниваний будем обозначать $W_{nN'N''} \subset W_n$. Распределение вероятностей на множестве допустимых выравниваний $W_{nN'N''}$ определим как

$$q_{nN'N''}(\mathbf{w}) = q_n(\mathbf{w}) z(N' - v'_{n,\mathbf{w}}) z(N'' - v''_{n,\mathbf{w}}) = q_n(\mathbf{v}'_{\mathbf{w}}) q_n(\mathbf{v}''_{\mathbf{w}}) z(N' - v'_{n,\mathbf{w}}) z(N'' - v''_{n,\mathbf{w}}),$$

где, $z(\tau)$ – распределение вероятностей длины заключительной части генерируемой последовательности, удовлетворяющее обычным требованиям $z(\tau) \geq 0$, $\sum_{\tau=0}^{\infty} z(\tau) = 1$ и определяющее длину формируемой последовательности через длину ее заключительной части $(N_{\boldsymbol{\omega}} - v_n)$.

В терминах множества парных выравниваний и с учетом предположений (7)-(9) общая структура потенциальной функции (3) запишется в виде:

$$\mathcal{K}(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum_{n=0}^{\infty} r(n) \sum_{\mathbf{w} \in W_{nN'N''}} q_{nN'N''}(\mathbf{w}) \Phi(\boldsymbol{\omega}', \boldsymbol{\omega}'' \mid \mathbf{w}) \tag{11}$$

где $\Phi(\boldsymbol{\omega}', \boldsymbol{\omega}'' \mid \mathbf{w}) = \bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}'_{\mathbf{v}'_{\mathbf{w}}}) \bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}''_{\mathbf{v}''_{\mathbf{w}}}) \int_{\vartheta \in \Omega_n} \prod_{i=1}^{n} \xi(\vartheta_i) \left(\prod_{t=\dot{v}'_{i,\mathbf{w}}}^{\ddot{v}'_{i,\mathbf{w}}} \psi(\omega'_t \mid \vartheta_i) \right) \left(\prod_{t=\dot{v}''_{i,\mathbf{w}}}^{\ddot{v}''_{i,\mathbf{w}}} \psi(\omega''_t \mid \vartheta_i) \right) d\vartheta$ – совместная плотность вероятности появления пары сравниваемых последовательностей, условная относительно парного выравнивания $\mathbf{w}$.

Конкретный вид потенциальной функции определяют, во-первых, распределение вероятностей $r(n)$ на множестве значений длины последовательности-прародителя, во-вторых, распределение $q_n(\mathbf{v})$ на множестве односторонних выравниваний V_n, определяющее распределение $q_n(\mathbf{w})$ и, в-третьих, элементарные распределения $\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}})$, $\psi(\omega \mid \vartheta)$ и $\xi(\vartheta)$. В дальнейшем, данные распределения $\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{v})$, $\psi(\omega \mid \vartheta)$ и $\xi(\vartheta)$ будут отличаться для сигналов и символьных последовательностей. Этим и будет определяться различие между соответствующими двумя видами потенциальных функций.

Может оказаться целесообразным нормировать функции $\Phi(\bar{\bar{\boldsymbol{\omega}}}'_{\mathbf{v}'_{\mathbf{w}}}, \bar{\bar{\boldsymbol{\omega}}}''_{\mathbf{v}''_{\mathbf{w}}} \mid \mathbf{w})$ на произведение полных условных плотностей, получая нормированную потенциальную функцию:

$$\bar{\mathcal{K}}(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum_{n=0}^{\infty} r(n) \sum_{\mathbf{w} \in W_{nN'N''}} q_{nN'N''}(\mathbf{w}) \frac{\Phi(\bar{\bar{\boldsymbol{\omega}}}'_{\mathbf{v}'_{\mathbf{w}}}, \bar{\bar{\boldsymbol{\omega}}}''_{\mathbf{v}''_{\mathbf{w}}} \mid \mathbf{w})}{h_n(\bar{\bar{\boldsymbol{\omega}}}'_{\mathbf{v}'_{\mathbf{w}}} \mid \mathbf{v}'_{\mathbf{w}}) h_n(\bar{\bar{\boldsymbol{\omega}}}''_{\mathbf{v}''_{\mathbf{w}}} \mid \mathbf{v}''_{\mathbf{w}})}, \tag{12}$$

где $h_n(\bar{\omega}_{\mathbf{v}_\mathbf{w}} \mid \mathbf{v}_\mathbf{w}) = \bar{\bar{\eta}}(\bar{\bar{\omega}}_{\mathbf{v}_\mathbf{w}}) \int_{\vartheta \in \Omega_n} \prod_{i=1}^{n} \xi(\vartheta_i) \left(\prod_{t=\dot{v}_{i,\mathbf{w}}}^{\ddot{v}_{i,\mathbf{w}}} \psi(\omega_t \mid \vartheta_i) \right) d\vartheta$.

Кроме нормированных и ненормированных потенциальных функций будем различать потенциальные функции фиксированного и нефиксированного порядка, а также локальные и глобальные потенциальные функции.

Потенциальные функции фиксированного порядка n являются очень важным частным случаем функций (11) и (12), впервые предложенным в данной работе. Они могут быть получены путем такого выбора распределения $r(n)$, что $r(n)=1$ и $r(k)=0$ для любых $k \neq n$. Такой выбор распределения $r(n)$ означает предположение о том, что последовательность-прародитель ϑ имела длину n. В этом случае потенциальная функция (11) будет иметь вид:

$$\mathcal{K}_n(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum_{\mathbf{w} \in W_{nN'N''}} q_{nN'N''}(\mathbf{w}) \Phi(\bar{\omega}'_{\mathbf{v}'_\mathbf{w}}, \bar{\omega}''_{\mathbf{v}''_\mathbf{w}} \mid \mathbf{w})$$

В случае же, если нет основания для введения каких-либо предположений относительно длины скрытой последовательности-прародителя, то естественно в качестве распределения $r(n)$ выбрать несобственное «равномерное» распределение $r(n) = const$. В этом случае потенциальная функция (11) будет иметь вид:

$$\mathcal{K}(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum_{n=0}^{\infty} \sum_{\mathbf{w} \in W_{nN'N''}} q_{nN'N''}(w) \Phi(\bar{\omega}'_{\mathbf{v}'_\mathbf{w}}, \bar{\omega}''_{\mathbf{v}''_\mathbf{w}} \mid \mathbf{w})$$

Такая функция названа в данной работе ***потенциальной функцией абсолютно нефиксированного порядка***.

Потенциальная функция называется ***локальной***, если в сравнении пары последовательностей участвуют только их центральные части. Это означает, прежде всего, что распределение на множестве выравниваний $q_n(\mathbf{w})$ полностью инвариантно к совместным сдвигам ключевых элементов

$$q_n(\mathbf{w}) = q_n\left[\begin{pmatrix} \dot{v}'_1, \ddot{v}'_1 \\ \dot{v}''_1, \ddot{v}''_1 \end{pmatrix}, \ldots, \begin{pmatrix} \dot{v}'_n, \ddot{v}'_n \\ \dot{v}''_n, \ddot{v}''_n \end{pmatrix}\right] = q_n\left[\begin{pmatrix} \dot{v}'_1 + \Delta', \ddot{v}'_1 + \Delta' \\ \dot{v}''_1 + \Delta'', \ddot{v}''_1 + \Delta'' \end{pmatrix}, \ldots, \begin{pmatrix} \dot{v}'_n + \Delta', \ddot{v}'_n + \Delta' \\ \dot{v}''_n + \Delta'', \ddot{v}''_n + \Delta'' \end{pmatrix}\right],$$

т.е. определяет случайную конфигурацию центральных частей сравниваемых последовательностей, но не их положение. Соответственно, длины и состав дополнительных подпоследовательностей $\bar{\bar{\omega}}'_{\mathbf{v}'_\mathbf{w}}, \bar{\bar{\omega}}''_{\mathbf{v}''_\mathbf{w}}$ с вероятностной точки зрения не определены, т.е. в роли $\bar{\bar{\eta}}(\bar{\bar{\omega}}_\mathbf{v})$ и $z(\tau)$ принимается несобственное «равномерное» распределение $\bar{\bar{\eta}}(\bar{\bar{\omega}}_\mathbf{v}) = const$ и $z(\tau) = const$.

Потенциальная функция называется ***глобальной***, если относительно распределения на множестве выравниваний $q_n(\mathbf{w})$ не делается никаких специальных предположений и распределение $\bar{\bar{\eta}}(\bar{\bar{\omega}}_\mathbf{v})$ не является безразличным:

$$\bar{\bar{\eta}}(\bar{\bar{\omega}}_\mathbf{v}) = \prod_{t=1}^{|\bar{\bar{\omega}}_\mathbf{v}|} \xi(\bar{\bar{\omega}}_{\mathbf{v},t}), \qquad (13)$$

где $\xi(\omega), \omega \in \tilde{A}$ – некоторое распределение на множестве примитивов.

Для символьных последовательностей в данной работе рассматриваются как глобальные, так и локальные потенциальные функции в их нормированной форме, а также потенциальные функции фиксированного порядка. Во всех случаях распределение вероятностей $q_n(\mathbf{v})$ выберем таким образом, чтобы выполнялось условие: $q_n(\mathbf{v}) \neq 0$ только для таких односторонних выравниваний (4), для которых начало и конец каждого интервала v_i, $i = 1, \ldots, n$ совпадают, т.е. $\dot{v}_i = \ddot{v}_i = v_i$. Таким образом, одностороннее выравнивание, определяющее структуру случайного преобразования последовательности-прародителя $\vartheta = (\vartheta_i, i = 1, \ldots, n) \in \Omega_n$ в последовательность не меньшей длины $\boldsymbol{\omega} = (\omega_t, t = 1, \ldots, N) \in \Omega_{\geq n}$, понимается как прямое перечисление позиций $\mathbf{v} = (v_1, \ldots, v_n)$, в которые будут отображаться элементы исходной последовательности (рис. 2).

В главе 2 линейное пространство примитивов $\tilde{A}$, образующих символьные последовательности, рассмотрено как натянутое на исходный конечный алфавит $A \subset \tilde{A}$. Распределение $\left(p_n(\vartheta), \vartheta \in \Omega_n\right)$ (7) на множестве вариантов последовательности-прототипа $\vartheta = (\vartheta_i \in \tilde{A}, \ i = 1, \ldots, n) \in \Omega_n \subseteq \Omega$ приня-

той длины n в случае символьных последовательностей будем рассматривать в пределах конечного алфавита символов $\vartheta \in A \subset \tilde{A}$, т.е. $\xi(\vartheta) = 0$ при $\vartheta \in \tilde{A} \setminus A$.

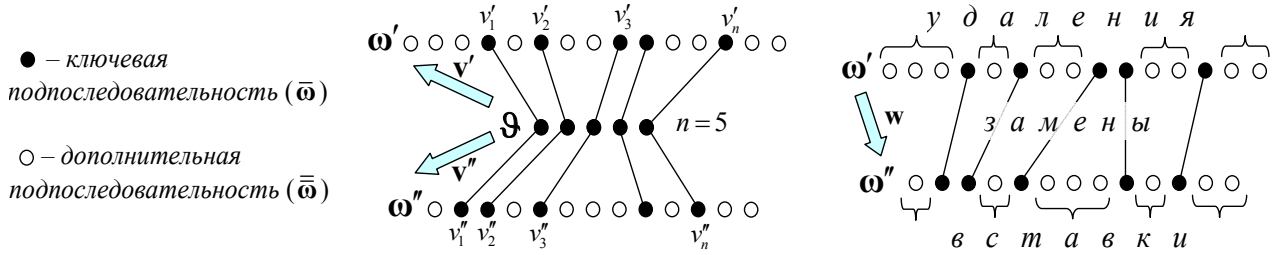

Рисунок 2 – Структура преобразования символьных последовательностей

В частности, в качестве многократно повторяемых в выражении (11) распределений $\xi(\vartheta)$ будем использовать финальное распределение в модели эволюции Дэйхофф, а в качестве условных распределений $\psi(\omega \mid \vartheta)$ – распределения, определяемые матрицей переходных вероятностей М. Дэйхофф превращения аминокислоты $\alpha^i = \vartheta$ в аминокислоту $\alpha^j = \omega$.

Ниже приведены частные структуры наиболее часто используемых в данной диссертации потенциальных функций, полученных с учетом предположений, учитывающих особенности символьных последовательностей, а также предположения об эргодичности и обратимости марковской цепи эволюции аминокислот.

Нормированная глобальная потенциальная функция нефиксированного порядка на множестве символьных последовательностей имеет вид:

$$K^{[s]}(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum\nolimits_{n=0}^{\infty} r(n) \sum\nolimits_{\mathbf{w} \in W_{nN'N''}} q_{nN'N''}(\mathbf{w}) \prod\nolimits_{i=1}^{n} \tilde{\mu}_{[s]}\left(\omega''_{v''_{i,\mathbf{w}}}, \omega'_{v'_{i,\mathbf{w}}}\right), \tag{14}$$

где $\tilde{\mu}_{[s]}(\omega'', \omega')$ - потенциальная функция на множестве аминокислот (2) для шага эволюции s.

Нормированная локальная потенциальная функция нефиксированного порядка на множестве символьных последовательностей имеет вид:

$$K^{[s]}(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum\nolimits_{n=0}^{\infty} r(n) \sum\nolimits_{\mathbf{w} \in W_{nN'N''}} q_{nN'N''}(\mathbf{w}) \prod\nolimits_{i=1}^{n} \tilde{\mu}_{[s]}\left(\omega''_{v''_{i,\mathbf{w}}} \mid \omega'_{v'_{i,\mathbf{w}}}\right) \tag{15}$$

Локальная потенциальная функция фиксированного порядка на множестве символьных последовательностей имеет вид:

$$K_n^{[s]}(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum\nolimits_{\mathbf{w} \in W_{nN'N''}} q_{nN'N''}(\mathbf{w}) \prod\nolimits_{i=1}^{n} \xi\left(\omega'_{v'_{i,\mathbf{w}}}\right) \psi_{[s]}\left(\omega''_{v''_{i,\mathbf{w}}} \mid \omega'_{v'_{i,\mathbf{w}}}\right) \tag{16}$$

Для сигналов в данной работе рассматриваются только глобальные нормированные потенциальные функции. При той же формальной структуре одностороннего выравнивания, что и для символьных последовательностей $\mathbf{v} = (v_1, ..., v_n)$ (4), распределение вероятностей $q_n(\mathbf{v})$ будем задавать так, чтобы с вероятностью 1 оно определяло полную сегментацию оси формируемого сигнала на n интервалов: $\dot{v}_1 = 1$, $\dot{v}_i = \ddot{v}_{i-1} + 1$, $i = 1, ..., n$.

При таком выборе распределения $q_n(\mathbf{v})$ формируемый сигнал $\boldsymbol{\omega} = (\omega_t, t = 1, ..., N) \in \Omega_{\geq n}$ полностью совпадает с ключевой последовательностью, состоящей из n сегментов $\boldsymbol{\omega} = \bar{\boldsymbol{\omega}}_{\mathbf{v}} = (\bar{\omega}_{v_i}, i = 1, ..., n)$, в то время как дополнительная подпоследовательность всегда будет пустой $\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}} \in \Omega_0$. Соответственно, распределение $\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}})$ теряет смысл, т.е. $\bar{\bar{\eta}}(\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}}) = 1$ для $\bar{\bar{\boldsymbol{\omega}}}_{\mathbf{v}} \in \Omega_0$, и в модели не присутствует.

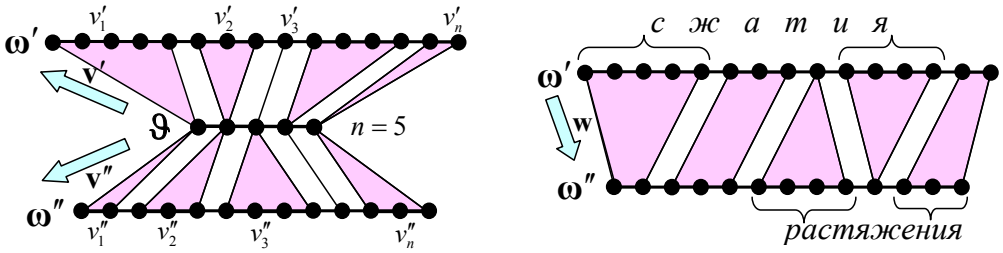

Рисунок 3 – Структура преобразования сигналов

Далее, поскольку сигналы, как правило, образованы последовательностями конечномерных векторов $\omega_t = \mathbf{x}_t = (x_t^1 \cdots x_t^m)^T \in R^m$, то естественно выбрать распределение $\xi(\vartheta)$ в составе (11) и (13) в классе нормальных распределений с нулевым математическим ожиданием и независимыми идентичными компонентами:

$$\xi(\vartheta) = \xi(\mathbf{y}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma^2\mathbf{I}\right). \tag{17}$$

Аналогично, условное распределение $\psi(\omega \mid \vartheta)$ выберем в виде нормальной линейной модели

$$\psi(\omega \mid \vartheta) = \psi(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{x} \mid \mathbf{y}, \delta^2\mathbf{I}\right). \tag{18}$$

С учетом предположений (17) и (18) нормированная глобальная потенциальная функция нефиксированного порядка примет вид:

$$\tilde{\mathcal{K}}(\boldsymbol{\omega}', \boldsymbol{\omega}'') = \sum_{n=0}^{\infty} r(n) \sum_{\mathbf{w} \in W_{nN'N''}} q_n(\mathbf{w}) \prod_{i=1}^{n} \left(1 \Big/ \mathcal{N}\left(\bar{\mathbf{x}}''_{v''_{i,\mathbf{w}}} \mid \mathbf{0}, (\sigma^2 + \delta^2)\mathbf{I}\right)\right) \mathcal{N}\left(\bar{\mathbf{x}}''_{v''_{i,\mathbf{w}}} \mid \bar{\mathbf{m}}_{\bar{\mathbf{x}}''_{v''_{i,\mathbf{w}}} \mid \bar{\mathbf{x}}'_{v'_{i,\mathbf{w}}}}, \gamma^2_{|v'_{i,\mathbf{w}}|}\mathbf{I} + \delta^2\mathbf{I}\right). \tag{19}$$

где $\bar{\mathbf{x}}'_{v'_{i,\mathbf{w}}} = (\mathbf{x}'_{\dot{v}'_{i,\mathbf{w}}}, ..., \mathbf{x}'_{\ddot{v}'_{i,\mathbf{w}}}) \in R^{m|v'_{i,\mathbf{w}}|}$ и $\bar{\mathbf{x}}''_{v''_{i,\mathbf{w}}} = (\mathbf{x}''_{\dot{v}''_{i,\mathbf{w}}}, ..., \mathbf{x}''_{\ddot{v}''_{i,\mathbf{w}}}) \in R^{m|v''_{i,\mathbf{w}}|}$ – соответствующие фрагменты двух сигналов, выделяемые i-й позицией парного выравнивания $\mathbf{w}$, $\bar{\mathbf{m}}_{\bar{\mathbf{x}}''_{v''_{i,\mathbf{w}}} \mid \bar{\mathbf{x}}'_{v'_{i,\mathbf{w}}}} = (\underbrace{\mathbf{m}_{\mathbf{y}_i \mid \bar{\mathbf{x}}'_{v'_{i,\mathbf{w}}}} \cdots \mathbf{m}_{\mathbf{y}_i \mid \bar{\mathbf{x}}'_{v'_{i,\mathbf{w}}}}}_{|v''_{i,\mathbf{w}}| \text{ раз}})$ – математическое ожидание появления фрагмента $\bar{\mathbf{x}}''_{v''_{i,\mathbf{w}}}$, условное относительно исходного фрагмента $\bar{\mathbf{x}}'_{v'_{i,\mathbf{w}}}$, и $\gamma^2_{|v'_{i,\mathbf{w}}|} = 1\Big/\left(1/\sigma^2 + |v'_{i,\mathbf{w}}|/\delta^2\right)$ – дисперсия апостериорного распределения i-го элемента неизвестного прародителя $\mathbf{y}_i$, относительно соответствующего фрагмента исходного сигнала $\bar{\mathbf{x}}'_{v'_{i,\mathbf{w}}}$.

Для сигналов естественно принять $\sigma^2 \to \infty$, причем в этом случае только нормированная потенциальная функция имеет смысл.

Алгоритмы вычисления потенциальных функций рассмотрены в диссертации только для специального класса распределений вероятностей на множестве парных выравниваний $q_n(\mathbf{w})$, который, в то же время, является достаточно широким для практических приложений. Распределение $q_n(\mathbf{w}) = q_n(\mathbf{v}')q_n(\mathbf{v}'')$ естественно выбирать зависящим только от длин локальных трансформаций осей сравниваемых последовательностей (удалений-вставок для символьных последовательностей или растяжений-сжатий для сигналов). Типичным является предположение, что случайные длины этих трансформаций являются независимыми, и имеют распределения вероятностей, монотонно убывающие с увеличением длины:

$$g_i(d_i \mid a, b) \propto \begin{cases} 1, & d_i = 1, \\ \exp\left[-\beta\left(a + bd_i\right)\right], & d_i > 1, \end{cases} \tag{20}$$

где $d_0 = v_1$, $d_i = v_i - v_{i-1}$, $i = 1, ..., n-1, d_n = N_\omega - v_n$ – для символьных последовательностей и, $d_i = \ddot{v}_i - \dot{v}_i + 1$, $i = 1, ..., n$ – для сигналов.

При этом если $a = 0$, то «стоимость» двух локальных трансформаций длин d_i и d_j будет равна «стоимости» одной трансформации суммарной длины $d_i + d_j$. Если же $a > 0$, то один длинная трансформация будет более предпочтительной, чем несколько коротких, имеющих в сумме ту же длину.

Распределение $q_n(\mathbf{v})$, определяющее $q_n(\mathbf{w})$ отличается для локальных и глобальных потенциальных функций. Так, для локальных потенциальных функций учитываются только трансформации в средней части $q_n(\mathbf{v}) \propto \prod_{i=2}^{n-1} g(d_i(\mathbf{v}) \mid a, b)$, тогда как при глобальном сравнении учитываются все трансформации: $q_n(\mathbf{v}) = \prod_{i=1}^{n} g(d_i(\mathbf{v}) \mid a, b)$.

В четвертой главе приводятся методы решения задач анализа данных, адаптированные для потенциальных функций, а также предлагается метод поиска общего прародителя для группы аминокислотных последовательностей.

Поиск общего прародителя группы аминокислотных последовательностей

Пусть A – множество аминокислот и $\bar{\boldsymbol{\omega}} = \{\boldsymbol{\omega}_j = (\omega_{jt} \in A, t = 1,...,N_j), j = 1,...,M\}$ – анализируемая совокупность последовательностей. Основная гипотеза состоит в том, что все белки $\bar{\boldsymbol{\omega}}$ получены независимо из некоторой общей аминокислотной последовательности-прародителя $\boldsymbol{\vartheta} = (\vartheta_i, i = 1,...,n)$ известной длины n путем описанного в главе 3 моделью случайного преобразования. При этом предполагается, что элементы ϑ_i были выбраны из алфавита аминокислот A независимо в соответствии с неизвестными наблюдателю распределениями вероятностей $\boldsymbol{\beta}_i = (\beta_{ik}, k = 1,...,m)$, $0 \le \beta_{ik} \le 1$, $\sum_{k=1}^{m} \beta_{ik} = 1$. Совокупность таких распределений $\bar{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_n)$ соответствует принятому в биоинформатике понятию профиля последовательности. С учетом случайности выбора $\boldsymbol{\vartheta}$ функция правдоподобия для одного белка будет иметь вид:

$$\varphi(\boldsymbol{\omega} \mid \bar{\boldsymbol{\beta}}) \propto \sum_{\mathbf{v} \in V_{nN_{\boldsymbol{\omega}}}} q(\mathbf{v}) \eta_n(\boldsymbol{\omega} \mid \bar{\boldsymbol{\beta}}, \mathbf{v}), \text{ где } \eta_n(\boldsymbol{\omega} \mid \bar{\boldsymbol{\beta}}, \mathbf{v}) = \prod_{i=1}^{n} \sum_{k=1}^{m} \beta_{ik} \psi(\omega_{v_{i,\mathbf{w}}} \mid \alpha^{(k)})$$

Задачу поиска общего прародителя поставим как задачу оценивания последовательности распределений $\bar{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_n)$ по всей совокупности белков $\bar{\boldsymbol{\omega}}$. Оценку будем производить максимизируя соответствующую функцию правдоподобия $F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}})$:

$$\hat{\bar{\boldsymbol{\beta}}} = \arg\max F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}) = \arg\max \prod_{j=1}^{M} \varphi(\boldsymbol{\omega}_j \mid \bar{\boldsymbol{\beta}}) = \arg\max \sum_{j=1}^{M} \ln\left(\sum_{\mathbf{v} \in V_{nN_{\boldsymbol{\omega}}}} q_{N_{\boldsymbol{\omega}}}(\mathbf{v}) \eta_n(\boldsymbol{\omega}_j \mid \bar{\boldsymbol{\beta}}, \mathbf{v}) \right)$$

В основу решения данной задачи положена итерационная EM-процедура, впервые предложенная М.И. Шлезингером в 1965 году, которая применима к широкому классу функций правдоподобия для вероятностных моделей со скрытыми параметрами. В данном случае в качестве такого скрытого параметра выступает выравнивание $\mathbf{v}$.

Пусть на s-м шаге получено приближение к искомому профилю последовательности-прародителя $\bar{\boldsymbol{\beta}}^s = (\boldsymbol{\beta}_1^s,...,\boldsymbol{\beta}_n^s)$ и пусть найдено апостериорное распределение на множестве выравниваний j-го белка $p(\mathbf{v} \mid \boldsymbol{\omega}_j, \bar{\boldsymbol{\beta}}^s)$, $\mathbf{v} \in V_{nN_j}$, в предположении, что $\bar{\boldsymbol{\beta}}^s$ – истинный профиль исходной последовательности. Выберем $\bar{\boldsymbol{\beta}}^{s+1}$ по правилу:

$$\bar{\boldsymbol{\beta}}^{s+1} = \arg\max \sum_{j=1}^{M} \sum_{\mathbf{v} \in V_{nN_j}} p(\mathbf{v} \mid \boldsymbol{\omega}_j, \bar{\boldsymbol{\beta}}^s) \ln \eta(\boldsymbol{\omega}_j \mid \bar{\boldsymbol{\beta}}, \mathbf{v}) \qquad (21)$$

Теорема 3. *При определении $\bar{\boldsymbol{\beta}}^{s+1}$ в соответствии с (21) справедливо неравенство $F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}^{s+1}) \ge F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}^s)$, причем $F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}^{s+1}) = F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}^s)$ тогда и только тогда, когда $\nabla_{\boldsymbol{\beta}_i} F(\bar{\boldsymbol{\omega}} \mid \bar{\boldsymbol{\beta}}^s) = 0$ для всех элементов прародителя $i = 1,...,n$.*

Теорема 4. *Задача (21) эквивалентна совокупности независимых задач для отдельных элементов профиля*

$$\boldsymbol{\beta}_i^{s+1} = \arg\max_{\beta_{i1},...,\beta_{im}} \sum_{l=1}^{m} h_i(\bar{\boldsymbol{\beta}}^s, \bar{\boldsymbol{\omega}}) \ln \sum_{k=1}^{m} \beta_{ik} \psi\left(\alpha^{(l)} \mid \alpha^{(k)}\right), \text{ где } h_i(\bar{\boldsymbol{\beta}}^s, \boldsymbol{\omega}_j) = \sum_{j=1}^{M} \sum_{\substack{t=1,\\ \omega_{jt} = \alpha^{(l)}}}^{N_j} p^{it}(\bar{\boldsymbol{\beta}}^s, \boldsymbol{\omega}_j) \qquad (22)$$

и $p^{it}(\bar{\boldsymbol{\beta}}^s, \boldsymbol{\omega}_j)$ – апостериорная вероятность того, что i-й элемент последовательности-прародителя преобразуется в t-ю позицию j-го белка.

Решение задачи (22) очевидно. Компоненты элемента профиля $\boldsymbol{\beta}_i^{s+1} = (\beta_{i1}^{s+1},...,\beta_{im}^{s+1})$ являются решением системы линейных алгебраических уравнений с матрицей условных вероятностей эволюционного чередований аминокислот:

$$\sum_{k=1}^{m} \left[\psi(\alpha^{(l)} \mid \alpha^{(k)}) \right] \beta_{ik}^{s+1} = u_{il},\ l = 1,...,m,\ u_{il} = h_i(\bar{\boldsymbol{\beta}}^s, \boldsymbol{\omega}_j) / \sum_{r=1}^{m} h_r(\bar{\boldsymbol{\beta}}^s, \boldsymbol{\omega}_j)$$

Пятая глава посвящена экспериментальному исследованию предложенного класса потенциальных функций в задаче установления гомологий белков путем автоматической классификации составляющих их аминокислотных последовательностей, задаче поиска общего прародителя группы аминокислотных последовательностей и задаче верификации личности по динамике подписи.

Также, в данной главе приведены методы анализа данных, применяемые для решения указанных прикладных задач, сформулированные в терминах потенциальных функций и существенно эксплуатирующие их свойства, а именно: метод распознавания на два класса по одной

потенциальной функции (метод опорных объектов), метод распознавания на два класса по конечному множеству потенциальных функций и метод автоматической классификации k-средних.

Распознавание по одной потенциальной функции. Метод опорных объектов

Пусть Ω – множество всех объектов некоторой природы. В терминах потенциальной функции $K(\omega', \omega'')$, определенной на множестве объектов $\Omega^* = \{\omega_j, j = 1,...M\} \subset \Omega$ и погружающей его в линейное пространство $\tilde{\Omega} \supset \Omega \supset \Omega^*$, решающее правило метода опорных векторов В.Н. Вапника для классификации объектов на два класса $g = +1$ и $g = -1$ может быть представлено в виде разделяющей гиперплоскости

$$y(\omega) = K(\vartheta, \omega) + b > 0 \to g = 1, \;\; y(\omega) \le 0 \to g = -1, \tag{23}$$

полностью определяемой направляющем элементом $\vartheta \in \tilde{\Omega}$ и смещением $b \in R$.

Направляющий элемент оптимальной разделяющей гиперплоскости может быть найден как линейная комбинация $\vartheta = \sum_{j:\lambda_j > 0} g_j \lambda_j \omega_j$ объектов обучающей совокупности с коэффициентами $\lambda_j \ge 0$, которые являются решением двойственной задачи обучения по методу опорных векторов

$$\begin{cases} \sum_{j=1}^{N} \lambda_j - (1/2) \sum_{j=1}^{N} \sum_{l=1}^{N} \left[g_j g_l K(\omega_j, \omega_l) \right] \lambda_j \lambda_l \to \max, \\ \sum_{j=1}^{N} g_j \lambda_j = 0, \;\; 0 \le \lambda_j \le C/2, \; j = 1,..., N. \end{cases} \tag{24}$$

Выбор смещения $b \in R$ может быть осуществлен по-разному для обеспечения симметричного или несимметричного положения гиперплоскости и не представляет проблемы.

Распознавание по нескольким потенциальным функциям

В данной работе применяется принцип комбинирования потенциальных функций, идея которого впервые была предложена Моттлем В.В., Серединым О.С. и Красоткиной О.В[1], а затем существенно развита А.И. Татарчуком и В.В. Моттлем. Данный подход позволяет автоматически в процессе адаптивного обучения выбирать потенциальную функцию, наиболее адекватную данным учителя.

Пусть $K_i(\omega', \omega'')$, $i = 1,..., n$ – потенциальные функции, определенные на множестве объектов $\omega \in \Omega$. Каждая из них погружает множество Ω в гипотетическое линейное пространство $\Omega \subset \tilde{\Omega}_i$, $i = 1,..., n$. Удобно рассматривать их совместно как декартово произведение $\tilde{\tilde{\Omega}} = \tilde{\Omega}_1 \times ... \times \tilde{\Omega}_n = \{\bar{\omega} = < \omega_1, ..., \omega_n > : \omega_i \in \tilde{\Omega}_i\}$. При этом решающее правило в комбинированном линейном пространстве $\tilde{\tilde{\Omega}}$ удобно искать в виде:

$$y(\omega) = \sum_{i=1}^{n} r_i K_i(\vartheta_i, \omega) + b > 0 \to g = 1, \;\; y(\omega) \le 0 \to g = -1,$$

где $r_i \ge 0$ – неотрицательные веса при потенциальных функциях. Идея адаптивного обучения в комбинированном пространстве $\tilde{\tilde{\Omega}}$ состоит в одновременном нахождении направляющих элементов ϑ_i в отдельных линейных пространствах $\tilde{\Omega}_i$ и неотрицательных весов r_i и реализуется в виде итерационной процедуры:

$$\vartheta_i^k = r_i^{k-1} \sum_{j:\lambda_j^k > 0} g_j \lambda_j^k \omega_j, \;\; r_i^k = (r_i^{k-1})^2 \sum_{j:\lambda_j^k > 0} \sum_{l:\lambda_l^k > 0} K_i(\omega_j, \omega_l) \lambda_j^k \lambda_l^k \tag{25}$$

На каждой итерации k коэффициенты $\lambda_1^k \ge 0, ..., \lambda_N^k \ge 0$ находятся как решение задачи обучения, имеющей структуру, аналогичную (24):

$$\begin{cases} \sum_{j=1}^{N} \lambda_j - \frac{1}{2} \sum_{j=1}^{N} \sum_{l=1}^{N} \left[g_j g_l \sum_{i=1}^{n} r_i^k K(\omega_j, \omega_l) \right] \lambda_j \lambda_l \to \max, \\ \sum_{j=1}^{N} g_j \lambda_j = 0, \;\; 0 \le \lambda_j \le C/2, \; j = 1,..., N. \end{cases} \tag{26}$$

[1] Моттль В.В., Середин О.С., Красоткина О.В. Комбинирование потенциальных функций при восстановлении зависимостей по эмпирическим данным //Искусственный интеллект, 2004, №2, стр. 134—139.

Определение константы b^k на каждой итерации не представляет сложности. Как правило, процесс сходится за 10-15 шагов.

Метод k-средних для автоматической классификации объектов на k классов

Пусть $\Omega^*=\{\boldsymbol{\omega}_j, j=1,...,M\}$ – множество объектов, которые необходимо разбить на k непересекающихся подмножеств $\Omega^* = \Omega_1^* \bigcup \Omega_2^* \bigcup ... \bigcup \Omega_k^*$, $\Omega_i^* \bigcap \Omega_l^* = \varnothing$, $i,l=1,...,k, i \neq l$.

Любая потенциальная функция $K(\boldsymbol{\omega}', \boldsymbol{\omega}'')$, определенная на Ω^*, погружает его в линейное пространство $\tilde{\Omega} \subseteq \Omega \subseteq \Omega^*$ с евклидовой метрикой

$$\rho^2(\boldsymbol{\omega}', \boldsymbol{\omega}'') = K(\boldsymbol{\omega}', \boldsymbol{\omega}') + K(\boldsymbol{\omega}'', \boldsymbol{\omega}'') - 2K(\boldsymbol{\omega}', \boldsymbol{\omega}''),\ \boldsymbol{\omega}', \boldsymbol{\omega}'' \in \tilde{\Omega}^*. \qquad (27)$$

Итерационная процедура метода k-средних заключается в поочередном выполнении на каждой $(s+1)$-й итерации двух шагов:

1) нахождение k фиксированных абстрактных центров $\vartheta_1^{s+1}, ..., \vartheta_k^{s+1} \in \tilde{\Omega}$ по имеющемуся разбиению $\{\Omega_i^{*(s)}, i=1,...,k\}$ по правилу $\vartheta_i^{s+1} = \arg\min\limits_{\vartheta \in \tilde{\Omega}} \sum_{\boldsymbol{\omega}_l \in \Omega_i^{*(s)}} \rho^2(\boldsymbol{\omega}_l, \vartheta)$, что, с учетом (27) и свойств дифференциала Фреше, позволяет найти явное выражение для центров $\vartheta_i^{s+1} = \left(1/|\Omega_i^{*(s)}|\right) \sum_{\boldsymbol{\omega}_l \in \Omega_i^{*(s)}} \boldsymbol{\omega}_l$.

2) нахождение нового разбиения по найденным на данной итерации центрам:

$$\Omega_i^{*(s+1)} = \left\{ \boldsymbol{\omega}_j \in \Omega^* : \rho^2(\boldsymbol{\omega}_j, \vartheta_i^{s+1}) = \min_{l=1,...,k} \rho^2(\boldsymbol{\omega}_j, \vartheta_l^{s+1}) \right\}, i=1,...,k. \qquad (28)$$

С учетом (27) и линейности $K(\boldsymbol{\omega}', \boldsymbol{\omega}'')$ относительно операции сложения, можно записать:

$$\rho^2(\boldsymbol{\omega}_j, \vartheta_l^{s+1}) = K(\boldsymbol{\omega}_j, \boldsymbol{\omega}_j) + \frac{1}{|\Omega_i^{*(s)}|^2} \sum_{\boldsymbol{\omega}_j \in \Omega_i^{*(s)}} \sum_{\boldsymbol{\omega}_l \in \Omega_i^{*(s)}} K(\boldsymbol{\omega}_j, \boldsymbol{\omega}_l) - 2\frac{1}{|\Omega_i^{*(s)}|} \sum_{\boldsymbol{\omega}_j \in \Omega_i^{*(s)}} K(\boldsymbol{\omega}_j, \boldsymbol{\omega}_l), \qquad (29)$$

что, при подстановке в (28), позволяет избежать непосредственного вычисления абстрактных центров, минуя таким образом шаг 1. Критерием окончания итерационного процесса может служить стабилизация классификации.

Для определения количества кластеров в данной работе используется итерационный алгоритм, предложенный профессором Б.Г.Миркиным, основная идея которого заключается в следующем.

На каждой i-й итерации найдем реальный объект, максимально удаленный от центра данных $\phi = \left(1/|\Omega^*|\right) \sum_{j=1}^{M} \boldsymbol{\omega}_j$, $\phi \in \tilde{\Omega}$: $\mathbf{c}_i = \arg\max\limits_{\boldsymbol{\omega}_j, j=1,...,M} (\rho^2(\boldsymbol{\omega}_j, \phi))$, причем с учетом (29) $\mathbf{c}_i$ может быть найден без непосредственного вычисления абстрактного центра ϕ. К i-му классу будем относить объекты по правилу: $\Omega_i^* = \{\boldsymbol{\omega}_j \in \Omega^* : \rho^2(\boldsymbol{\omega}_j, \mathbf{c}_s) < \rho^2(\boldsymbol{\omega}_j, \phi), j=1,...,|\Omega^*|\}$. Выделенные объекты временно удаляются из исходного множества $\Omega^* = \Omega^* \setminus \Omega_i^*$ и процесс повторяется без смены центра ϕ. В результате получаем разбиение $\{\Omega_i^*, i=1,..,k\}$, которое принимается в качестве начального приближения для метода k-средних.

Анализ аминокислотных последовательностей

Современные методы секвенирования белков позволяют достаточно быстро получать новые аминокислотные последовательности, но не дают информации о роли, которую играют соответствующие белки в живых организмах. В то же время, известно, что белки, имеющие общего прародителя (гомологичные белки), как правило, обладают близкими свойствами и выполняют похожие функции. В связи с этим, задача выявления гомологий является важной задачами современной биоинформатики.

Природа предложенных в данной работе потенциальных функций, значения которых имеют смысл вероятностей совместного происхождения пары последовательностей из одного общего прародителя, очень адекватна понятию гомологичных белков, что дает основание предполагать полезность данных потенциальных функций(ПФ) для установления гомологии белков. Действительно, преимущество локальных потенциальных функций нефиксированного порядка перед другими известными мерами сходства было продемонстрировано в ряде экспериментов фран-

цузским ученым Ж. Ф. Вертом, предложившим алгебраическую структуру такой потенциальной функции, правда, без ее теоретического обоснования.

В данной работе приводится пример, когда другой частный случай, а именно, глобальная потенциальная функция нефиксированного порядка (14) дает лучшие результаты по сравнению с локальной потенциальной функцией (15).

Для решения задачи автоматической классификации в данной работе используется метод k-средних с предварительным определением числа классов.

В качестве базы для экспериментального исследования использовались аминокислотные последовательности вирусов простого герпеса из базы данных VIDA[1], разделенные на три класса на основе анализа эволюции вирусов герпеса[2]. Данная классификация и результаты, полученные с использованием четырех различных мер сходства приведены на рисунке 4.

Рисунок 4 – Классификации белков на основе анализа эволюции вирусов герпеса и результаты автоматической классификации

Из рисунка 4 видно, что только глобальная вероятностная потенциальная функция позволяет объединить в один класс белки, выполняющие одну и ту же функцию.

Поиск общего прародителя группы последовательностей

Эксперименты по исследованию процедуры поиска общего прародителя проводились на модельных и реальных данных. В приведенном на рис. 5 примере данные генерировались по следующей схеме. 1) Случайным образом была сгенерирована последовательность-прародитель длиной $n = 20$: $\vartheta = (\text{WFCNLPKALIVWPCNQIMAG})$. 2) На основе ϑ были сгенерированы анализируемые последовательности $\omega_j, j = 1,...,30$ путем добавления слева, в центр и справа случайных последовательностей случайной длины.

Согласно предложенному подходу, был найден вероятностный профиль общего прародителя длины $n = 20$, обеспечивающего максимум правдоподобия гипотезы, что все последовательности $\omega_j, j = 1,...,30$ получены из него путем независимых случайных преобразований (5). Для полученного в данном случае вероятностного профиля характерно, что в каждой позиции $i = 1,...,n$ одна из аминокислот имеет вероятность, равную единице или близкую к ней. В связи с этим результат поиска общего прародителя может быть выражен буквально в виде последовательности $\hat{\vartheta} = (\text{WFCNLPKALIVWPCNQIMAG})$, которая в данном случае абсолютно точно совпала с истинной последовательностью-прародителем ϑ.

---

[1] Virus Database at University College London. http://www.biochem.ucl.ac.uk/ bsm/virus_database/VIDA3/VIDA.html

[2] McGeoch, DJ, Rixon, FJ, and Davison, AJ: Topics in herpesvirus genomics and evolution, *Virus Research* 2006, 117, 90-104.

[3] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 215 (3), 1990, pp. 403–410.

[4] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48 (3), 1970, pp. 443–530.

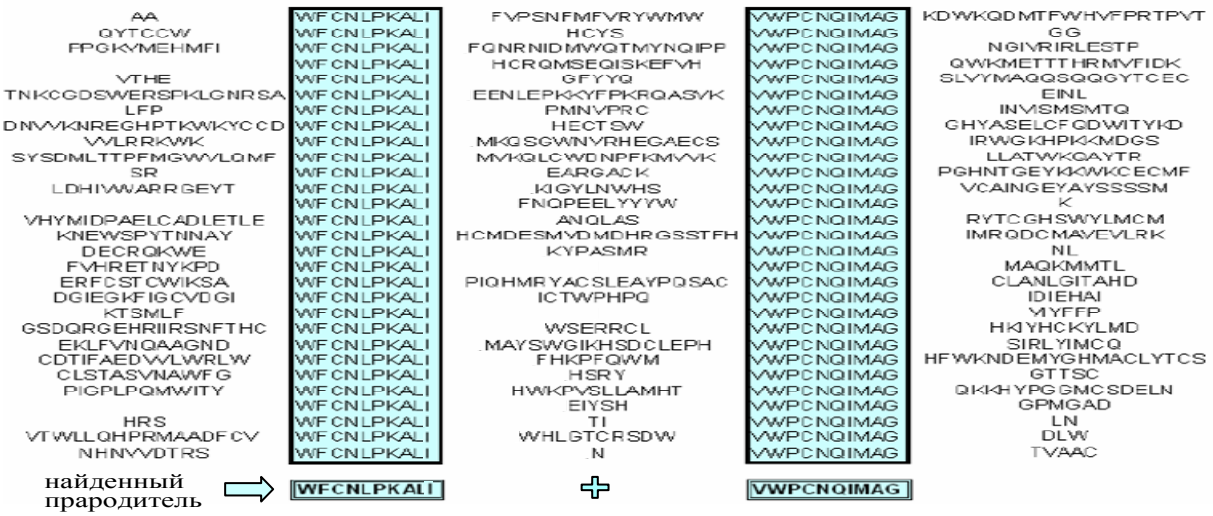

Рисунок 5 – Анализируемые последовательности и найденный общий прародитель

В диссертации также приведен пример поиска прародителя для реальных последовательностей, пространственная структура которых содержит известный домен (элемент пространственной структуры белка, представляющий собой достаточно стабильную и независимую подструктуру, укладка которой происходит независимо от остальных частей). Найденный вероятностный профиль общего прародителя позволил выделить консервативные регионы в каждой из 20 анализируемых последовательностей. На рисунке 6 представлены примеры пространственных структур анализируемых белков с выделенными на них областями, соответствующими найденным консервативным регионам. Очевидно, что эти области соответствуют одному и тому же домену, что также подтверждает эффективность использования предложенной процедуры поиска общего прародителя.

Рисунок 6 – Пространственные структуры белков 1LCK, 2H8H и 1GRI с выделенными на них найденными консервативными регионами

Задача верификации личности по динамике подписи – это двухклассовая задача распознавания, заключающаяся в проверке гипотезы, что анализируемая подпись принадлежит некоторому назвавшему себя автору.

Каждая подпись вводится в компьютер непосредственно в процессе ее написания и оказывается представленной семикомпонентным дискретным сигналом $\boldsymbol{\omega} = (\mathbf{x}_s, s = 1,...,N)$, включающим такие компоненты, как координаты (X и Y), угол наклона пера, угол поворота, усилие нажатия, скорость и ускорение (рис. 7).

Рисунок 7 – Примеры подписей и их представления в виде многокомпонентных сигналов

Следует отметить, что формула (19) определяет целый класс потенциальных функций (ПФ) на множестве сигналов и априори не представляется возможным определить какая из них будет более адекватна данным учителя. В данном эксперименте используется 12 потенциальных функций (табл. 1), отличающихся друг от друга разными наборами компонент сигналов, учитываемых при сравнении и разными значениями параметра $a = b$ (20), имеющего смысл штрафа на локальные растяжения и сжатия осей сравниваемых сигналов.

Таблица 1 – Множество используемых потенциальных функций (ПФ)

ПФ		**Подмножество компонент**	**ПФ**		**Подмножество компонент**
$a=2$	$a=4$		$a=2$	$a=4$	
K_1	K_2	координаты	K_7	K_8	координаты, скорость, ускорение
K_3	K_4	углы наклона	K_9	K_{10}	координаты, углы наклона, усилие нажатия
K_5	K_6	усилие нажатия	K_{11}	K_{12}	все компоненты

В качестве базы для экспериментального исследования использовалась база Международного соревнования по верификации личности (SVC2004), содержащая 1600 подписей 40 человек, по 20 оригинальных подписей и 20 умышленных подделок для каждого человека. Данные для обучения и контроля для каждого автора приведены в таблице 2.

Таблица 2 – Данные для одного решающего правила

Обучающее множество		**Тестовое множество**	
5 оригинальных подписей	5 умышленных подделок 10x39 случайных подделок	15 оригинальных подписей	15 умышленных подделок 39 случайных подделок

Для каждого автора обучение и распознавание проводилось 13 раз: для каждой из потенциальных функций $K_1 - K_{12}$ в соответствии с (24) и для комбинирования (25-26). В таблице 3 представлены результаты верификации. Из таблицы видно, что процент ошибок, полученный с использованием комбинирования потенциальных функций, оказался меньше, чем для любой из потенциальных функций отдельно.

Таблица 3 – Результаты верификации личности по подписи

ПФ	**Ошибка, %**	**Адекватна для***	**ПФ**	**Ошибка,%**	**Адекватна для***	**ПФ**	**Ошибка,%**	**Адекватна для***
K_1	0.65	6 человек	K_5	2.75	0 человек	K_9	0.58	6 человек
K_2	1.01	9 человек	K_6	2.50	4 человек	K_{10}	0.76	3 человек
K_3	5.58	0 человек	K_7	0.98	4 человек	K_{11}	0.47	5 человек
K_4	7.50	0 человек	K_8	1.41	2 человек	K_{12}	1.01	1 человек
Ошибка при комбинировании потенциальных функций K_1 - K_{12}: **0.36%**								

*Количество человек, для которых соответствующая ПФ была выбрана в результате комбинирования

На рисунке 7 представлен интересный пример, демонстрирующий результат комбинирования потенциальных функций. Все компоненты изображенных на рисунке оригинальных подписей и умышленных подделок некоторого автора очень похожи, за исключением усилия нажатия. Единственной потенциальной функцией, имеющей ненулевой вес в результате комбинирования для данного автора стала потенциальная функция K_6, учитывающая только усилие нажатия.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

1. В данной работе доказано, что меры сходства аминокислот PAM и BLOSUM, общепринятые в современной биоинформатике, основаны на одной и той же модели эволюции аминокислот, разработанной Маргарет Дэйхофф, и по своей структуре являются потенциальными функциями.

2. Предложены новые вероятностные модели случайных преобразований сигналов и символьных последовательностей, в частности, модели эволюционных изменений аминокислотных последовательностей белков.

3. На основе этих моделей построен класс корректных потенциальных функций, выражающих правдоподобие гипотезы о наличии общего прародителя у пары сравниваемых сигналов либо символьных последовательностей разной длины.

4. Поставлена и решена задача поиска общего прародителя заданной длины для группы последовательностей в терминах введенного случайного преобразования.

5. Разработаны алгоритмы, реализующие предложенные схемы построения потенциальных функций на множествах сигналов и символьных последовательностей разной длины.

6. Построенные потенциальные функции применены для решения задачи верификации личности по динамике подписи, для агрегации аминокислотных последовательностей белков в функциональные семейства.

СПИСОК ОСНОВНЫХ ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Моттль В.В., Дмитриев Д.А., Сулимова В.В. Измерение попарного несходства подписей для идентификации личности. Сборник трудов международной конференции ММТТ-16, 2003г, том 4, с.216-218.
2. Моттль В.В., Сулимова В.В. Идентификация личности по динамике подписи методом опорных векторов. Сборник трудов международной конференции ММТТ-16, 2003г, том 4, с.219-222.
3. **Моттль В.В., Середин О.С., Сулимова В.В. Формирование потенциальных функций для беспризнакового распознавания сигналов и символьных последовательностей. Известия ТулГУ. Серия. Вычислительная техника. Информационные технологии. Системы управления. Т1. Вып.2. Информационные технологии. – Тула: ТулГУ, 2004, стр. 11-16.**
4. Моттль В.В., Середин О.С., Сулимова В.В. Потенциальные функции для беспризнакового восстановления зависимостей на множествах сигналов и символьных последовательностей. Искусственный интеллект, 2’2004, стр. 140-144.
5. **V. Mottl, O. Seredin, V. Sulimova. Mathematically correct methods of similarity measurement on sets of signals and symbolic sequences of different length. Pattern Recognition and Image Analysis, Vol.15, No. 1, 2005, pp. 87-89 .**
6. Моттль В.В., Сулимова В.В., Татарчук А.И. Автоматический выбор наиболее информативных фрагментов в задачах распознавания сигналов разной длительности. Таврический вестник математики и информатики – № 1, 2006, стр. 109-115.
7. Сулимова В.В., Разин Н.А., Моттль В.В., Мучник И.Б. Множественное выравнивание совокупности аминокислотных последовательностей на основе вероятностной модели эволюции. Таврический вестник математики и информатики N 2, 2008, pp. 202-210.
8. V. Mottl, M. Lange, V. Sulimova, A. Yermakov. Signature verification based on fusion of on-line and off-line kernels. Proc. of 19-th International conference on Pattern Recognition (ICPR 2008), Florida, USA, December 2008.
9. Sulimova V., Mottl V., Kulikowski C., Muchnik I. Probabilistic evolutionary model for substitution matrices of PAM and BLOSUM families. DIMACS Technical Report 2008-16, Rutgers University, 17 p., 2008. ftp://dimacs.rutgers.edu/pub/dimacs/TechnicalReports/TechReports/2008/2008-16.pdf
10. Mirkin B., Sulimova V. Mottl V. Is protein sequence data sufficient for deriving homology groups? Extending Dayhoff’s model to sequences. Technical Report. School of Computer Science and Information Systems, Birkbeck, University of London, February 2009, 23 p. http://www.dcs.bbk.ac.uk//research/techreps/2009/bbkcs-09-01.pdf .
11. Sulimova V., Mottl V., Mirkin B., Muchnik I., Kulikowski C. A Class of Evolution-Based Kernels for Protein Homology Analysis: A Generalization of the PAM Model. Proc. of 5th International Symposium on Bioinformatics Research and Applications, Nova Southeastern University, Ft. Lauderdale, Florida, USA, May 13-16, 2009.

Жирным шрифтом выделены публикации в журналах, рекомендованных ВАК.

Все приведенные в статьях результаты исследований, кроме результатов экспериментов в (7), получены лично автором. Все статьи, кроме (10), написаны лично автором с последующим участием соавторов в редактировании текста.