Математическая биология и биоинформатика
2021. Т. 16. № 2. С. 299–316. doi: 10.17537/2021.16.299

========== **БИОИНФОРМАТИКА** ==========

Главные компоненты генетических последовательностей: корреляции и достоверность

Ефимов В.М.*[1,2,3,4], Ефимов К.В.[5], Ковалева В.Ю.[2], Матушкин Ю.Г.[1]

[1]*Институт цитологии и генетики СО РАН, Новосибирск, Россия*
[2]*Институт систематики и экологии животных СО РАН, Новосибирск, Россия*
[3]*Новосибирский государственный университет, Новосибирск, Россия*
[4]*Томский государственный университет, Томск, Россия*
[5]*Высшая школа экономики, Москва, Россия*

Аннотация. Известно, что любой числовой ряд можно разложить на главные компоненты с помощью сингулярного спектрального анализа. Недавно мы предложили новый метод анализа PCA-Seq, который позволяет вычислять числовые главные компоненты для последовательности элементов любой природы. В частности, последовательность может быть символьной, в том числе, нуклеотидной или аминокислотной. При этом неизбежно встают два вопроса: об интерпретации полученных главных компонент и об оценке их достоверности. Для интерпретации главных компонент разумно вычислять их корреляции с любыми числовыми характеристиками элементов изучаемой последовательности, используемыми в данной предметной области – внешними факторами. При оценке достоверности корреляций между последовательностями необходимо учитывать, что стандартные критерии значимости опираются на предположение независимости наблюдений, которое для реальных последовательностей, как правило, не выполняется. В статье рассматривается применение для этих целей якорного бутстрепа, также ранее разработанного авторами статьи. В этом методе предполагается, что объекты могут быть представлены точками метрического пространства и в совокупности составляют в нем некоторую фиксированную структуру, в частности, последовательность. Объектам приписываются те же случайные целочисленные веса, что и при классическом бутстрепе. Этого достаточно для получения бутстреп-распределения коэффициентов корреляции и оценки их достоверности. При исследовании гена *SLC9A1* (синонимы *APNH, NHE1, PPP1R143*) выявились достоверные корреляции первой главной компоненты кодирующей последовательности с гидрофобностью / “трансмембранностью” соответствующих фрагментов аминокислотной последовательности, содержанием в них фенилаланина, а также разностью содержания тимина и аденина в нуклеотидных фрагментах. Похожая закономерность была найдена другими авторами для других генов, весьма вероятно, что она имеет более общий характер.

Ключевые слова: *SSA, PCA-Seq, ген SLC9A1(NHE1), CDS, вторичная структура белка, внешние факторы, якорный бутстреп.*

Сокращения:

БД – база данных
ВС – вторичная структура белка
ВФ – внешний фактор, внешние факторы
ГК – главные компоненты, PC
МГК – метод главных компонент, PCA

*efimov@bionet.nsc.ru

ЭФР – эмпирическая функция распределения
Nboot – число бутстреп-испытаний
PCo – главные координаты
SS – α-спираль
SSA – сингулярный спектральный анализ
TM – мембрана

ВВЕДЕНИЕ

Ранее нами был предложен метод PCA-Seq, позволяющий рассчитывать главные компоненты (ГК) для последовательности элементов любой природы, включая нечисловые [1]. При этом неизбежно встает вопрос об интерпретации и достоверности ГК. При оценке достоверности ГК обычных числовых матриц «объект – признак» традиционно рассматриваются два подхода: вероятностный и геометрический [2]. Вероятностный подход опирается на дополнительное предположение, что строки матрицы, характеризующие объекты, случайным образом выбраны из некоторого генерального распределения, т.е. сама матрица является случайной выборкой. По сути, предполагается отсутствие любой структурированности данных. Геометрический подход предполагает, что объекты могут быть представлены точками метрического пространства, между которыми можно вычислить матрицу расстояний. Множество объектов можно рассматривать как некоторую структуру, например, как последовательность, граф или набор кластеров.

Классическая математическая статистика предполагает не только наличие генерального распределения, но и известность его типа (предпочитается нормальное). По выборке определяются только его неизвестные параметры. На практике тип генерального распределения, как правило, неизвестен, поскольку нет другой информации, кроме самой выборки. Но из теории известно, что эмпирическая функция распределения (ЭФР), рассчитываемая по выборке, сходится к генеральной с ростом объема выборки. Поэтому имеет смысл использовать именно ЭФР в качестве модели генерального распределения, и тогда ни тип генерального распределения, ни оно само уже не нужны. Правда, нужны компьютеры, но, когда они появились, появился и бутстреп [3].

По отношению к классической математической статистике бутстреп, безусловно, был революцией. Но сам по себе бутстреп является строго вероятностным подходом. Используемый в нем случайный выбор с возвращением – это классический вероятностный прием. Как следствие, некоторые объекты выбираются несколько раз, да еще и по построению перемешиваются, а некоторые другие – ни разу. Поэтому классический бутстреп тоже плохо применим к структурированным данным, особенно, если требуется сохранение всех объектов [4].

Как сейчас становится ясным, можно обойтись не только без генерального распределения, но и без ЭФР, и без случайного выбора объектов с возвращением. Вместо этого можно считать, что в исходной матрице все строки-объекты имеют единичные веса. Тогда можно оставить все объекты на своих местах («заякорить»), а самим объектам приписать те же случайные целочисленные веса, что и при классическом бутстрепе [5]. Это и есть якорный бутстреп [6]. Как следствие, в якорном бутстрепе некоторые объекты получат большие, а некоторые – нулевые веса. Сумма всех весов равна числу объектов. Но это не означает разрушения структуры. Если исходные предположения классического бутстрепа выполняются, и исследуемую матрицу можно рассматривать, как случайную бесструктурную выборку, то в якорном бутстрепе результаты получатся ровно те же, что и в классическом. Если же объекты составляют некоторую геометрическую структуру, то якорный бутстреп, в отличие от

Математическая биология и биоинформатика. 2021. Т. 16. № 2. doi: 10.17537/2021.16.299

классического, позволяет учесть ее особенности, что делает возможной оценку устойчивости структуры, сохраняя все достоинства и саму идею бутстрепа.

Структурой для любой последовательности является сам порядок следования элементов друг за другом. Желательно сохранить этот порядок, насколько это возможно, поскольку нас интересуют корреляции с упорядоченными внешними факторами (ВФ). При обычном бутстрепе объекты выбираются случайным образом, поэтому порядок объектов тоже становится случайным. В якорном бутстрепе объекты сначала сохраняют упорядочивание по своим собственным номерам, а потом нумеруются заново (подробнее см. Материал и методы).

Методом PCA-Seq любую генетическую последовательность (нуклеотидов, кодонов, аминокислот, элементов вторичной структуры (ВС) белков и т. д.) можно представить в виде числовой матрицы «фрагменты последовательности – ГК», упорядоченной по фрагментам [1]. В ГК сохраняется порядок следования фрагментов. Поэтому достоверность корреляций ГК любой последовательности с ВФ можно оценить с помощью якорного бутстрепа. Демонстрация этой возможности является целью настоящей статьи.

В качестве примера взята кодирующая последовательность гена *SLC9A1* (синонимы *APNH, NHE1, PPP1R143*). Кодируемый ею белок состоит из двух доменов длиной ~500 и ~300 остатков. N-терминальный трансмембранный домен много раз пересекает мембрану в обоих направлениях, поэтому характерным элементом его ВС является большое количество α-спиралей. Кроме того, несколько α-спиралей входят в состав С-терминального цитозольного домена, находящегося вне мембраны [7].

Известно, что SLC9A1 – мембранный белок, переносящий Na^+ в клетку и H^+ из клетки – играет центральную роль в регуляции гомеостаза pH и является одним из основных источников кислотности в поверхностных слоях кожи [8]. Он также активируется наиболее частой мутацией при меланоме, BRAFV600E, тем самым вызывая нарушение регуляции pH во время инициации меланомы. Заболеваемость меланомой растет и в настоящее время является причиной большинства смертей, связанных с раком кожи [9, 10].

МАТЕРИАЛ И МЕТОДЫ

Расчет якорного бутстрепа для последовательности элементов любого типа

Рассмотрим для примера короткую искусственную последовательность $X = \{x_1, x_2, \ldots, x_{15}\}$. Пять шагов процесса построения для нее якорной бутстреп-копии проиллюстрированы таблицами 1a–1e. В таблице 1a в колонке N идут порядковые номера элементов последовательности X, в колонке W – веса элементов. Для начальной последовательности все веса равны единице. X в общем случае может быть многомерной. В якорном бутстрепе [6] элементам равновероятно приписываются случайные веса W_b с той же суммой (табл. 1b). Удаляем элементы с нулевыми весами (табл. 1c). Размножаем оставшиеся элементы в соответствии с их весами, сохраняя порядок в колонке N_b. Веса элементов W снова становятся единичными (табл. 1d). Нумеруем элементы заново (табл. 1e), то есть, возвращаем начальные номера N из табл. 1a.

Таблица 1e отличается от классической бутстреп-копии только порядком элементов. В соответствии с идеологией бутстрепа новые значения равновероятно выбраны из старых. Но порядок элементов, в отличие от классического, максимально сохранен. Это позволяет вычислять бутстреп-корреляции с любыми ВФ.

Для получения бутстреп-копий исходных данных и для классического, и для якорного бутстрепа достаточно датчика случайных чисел и сортировки. Реальная проблема заключается в другом. Кроме многократного расчета самого бутстрепа

необходимы многократные повторения исходного алгоритма, который должен применяться к каждой бутстреп-копии исходных данных. Для метода PCA-Seq – это расчет корреляций с ВФ, но на практике встречаются и гораздо более сложные случаи, которые требуют соответствующего программного обеспечения.

Таблицы 1a–1e. Схема расчета якорного бутстрепа для последовательности элементов любого типа

a				b				c				d				e		
N	**X**	**W**		**N**	**X**	**Wb**		**N**	**X**	**Wb**		**Nb**	**Xb**	**W**		**N**	**Xb**	**W**
1	x1	1		1	x1	3		1	x1	3		1	x1	1		1	x1	1
2	x2	1		2	x2	**0**		3	x3	2		1	x1	1		2	x1	1
3	x3	1		3	x3	2		4	x4	1		1	x1	1		3	x1	1
4	x4	1		4	x4	1		5	x5	1		3	x3	1		4	x3	1
5	x5	1		5	x5	1		6	x6	2		3	x3	1		5	x3	1
6	x6	1		6	x6	2		8	x8	1		4	x4	1		6	x4	1
7	x7	1		7	x7	**0**		9	x9	1		5	x5	1		7	x5	1
8	x8	1		8	x8	1		11	x11	1		6	x6	1		8	x6	1
9	x9	1		9	x9	1		14	x14	2		6	x6	1		9	x6	1
10	x10	1		10	x10	**0**		15	x15	1		8	x8	1		10	x8	1
11	x11	1		11	x11	1						9	x9	1		11	x9	1
12	x12	1		12	x12	**0**						11	x11	1		12	x11	1
13	x13	1		13	x13	**0**						14	x14	1		13	x14	1
14	x14	1		14	x14	2						14	x14	1		14	x14	1
15	x15	1		15	x15	1						15	x15	1		15	x15	1

Интерактивные статистические пакеты, управляемые пользователем через меню, плохо приспособлены к этой ситуации. Популярные пакеты-интерпретаторы с входными языками типа R или MATLAB тоже не всегда удобны, хотя бутстреп входит в эти пакеты как самостоятельная процедура. Поэтому мы разработали собственное модульное программное обеспечение – пакет Jacobi 4 [11], позволяющее легко включать бутстреп (и классический, и якорный) в разнообразные алгоритмы статистической обработки, при необходимости расширяя список модулей. Правильность расчетов по этим модулям проверяется с помощью пакетов PAST [12] и Statistica [13].

ВФ для ГК кодирующей последовательности гена *SLC9A1*

Кодирующая последовательность (CDS) гена *SLC9A1* (2445 нуклеотидов) и соответствующая ей аминокислотная последовательность (815 аминокислот) взяты из GenBank (код доступа NM_003047.5) [14]. В случае последовательностей любой природы, включая нечисловые, обычным приемом исследования является применение скользящего окна ширины L с некоторым шагом K.

В нашем случае $K = 3$, а ширина окна и, соответственно, длина получаемых фрагментов L для кодирующей последовательности выбрана равной 18 кодонам. Каждому фрагменту кодирующей последовательности однозначно соответствует фрагмент аминокислотной последовательности (18 остатков, примерно пять витков α-спирали в аминокислотной последовательности). В соответствии с методом PCA-Seq [1] между всеми 798 фрагментами кодирующей последовательности вычислена матрица квадратов евклидовых расстояний (p-дистанция) и по ней – матрица ГК кодирующей последовательности методом главных координат (PCo) Гауэра [15].

Доля несовпадающих нуклеотидов (или аминокислот) при попарном сравнении их позиций во фрагменте, или *p*-дистанция, представляет собой простейшее из эволюционных расстояний [16]. Заметим, что в нашем случае фрагменты являются наборами символов конечного алфавита, а дистанции – неотрицательными действительными числами. Квадратный корень из *p*-дистанции, в отличие от других эволюционных расстояний, всегда является евклидовой метрикой [17]. Матрица евклидовых расстояний между фрагментами однозначно задает их взаимное расположение как точек в евклидовом пространстве. Направление, в проекции на которое дисперсия множества фрагментов максимальна, является первой ГК и т.д. Для расчета всех ГК по матрице расстояний использован предложенный Дж. Гауэром еще в 1966 г. метод главных координат (PCo) [15]. Так как все фрагменты упорядочены в соответствии с их положением в последовательности, то методом PCA-Seq получаются числовые ГК символьной последовательности, в данном случае, ГК кодирующей последовательности. Для первых четырех ГК вычислены коэффициенты корреляции между ними и ВФ (табл. 2, 3).

Для чего нужны корреляции ГК с ВФ? Дело в том, что в самом методе PCA-Seq не используется никакой информации о природе элементов последовательности. Если заменить алфавит последовательности на любой другой с таким же числом букв, то ничего не изменится. Например, останется той же самой доля несовпадений для каждой пары фрагментов (*p*-дистанция). Так как матрица квадратов расстояний между фрагментами тоже останется той же самой, то не изменятся и ГК. Именно поэтому метод PCA-Seq является универсальным и применим к последовательностям элементов любого типа (позволяя при этом получать числовые ГК).

Но как интерпретировать найденные ГК с биологической точки зрения? Очевидно, в каждом отдельном случае надо использовать уже известную информацию об элементах последовательности, входящих в каждый фрагмент, особенно их числовые характеристики. Из этих характеристик можно составить числовые последовательности с уже понятным нам биологическим смыслом – ВФ, которые соответствуют рассматриваемой нами символьной последовательности. Поскольку они никак не были задействованы при расчете ГК методом PCA-Seq, то и по отношению к ГК они являются ВФ. Поэтому между ГК и ВФ можно просто вычислить корреляции.

На самом деле, одной кодирующей последовательности, представленной фрагментами, можно сопоставить очень много ВФ. Каждый фрагмент кодирующей последовательности состоит из 18 кодонов, т.е. из 54 нуклеотидов. Каждому кодону соответствует одна аминокислота, а каждому фрагменту – 18 аминокислот. Для данной аминокислотной последовательности рассчитана ВС белка (см. ниже), поэтому для каждой аминокислоты во фрагменте известно, входит ли она в α-спираль и/или в мембрану. Кроме того, у каждой аминокислоты имеются физико-химические свойства.

Какие ВФ стоит рассматривать в первую очередь? Поскольку фрагменту кодирующей последовательности однозначно соответствует аминокислотный фрагмент, естественно, надо брать содержание и нуклеотидов, и аминокислот в каждом фрагменте (табл. 2). Кроме того, желательно учесть свойства аминокислот. Для 20 протеиногенных аминокислот на сегодняшний день известно более 500 физико-химических и биохимических характеристик (AA-индексов), собранных в базе данных AAindex [18, 19], которая время от времени пополняется и обновляется. Многие из них коррелируют между собой. Этот массив периодически обрабатывается МГК, чтобы получить несколько интегрированных факторов, компактно представляющих все множество физико-химических свойств [20–24]. Мы пересчитали ГК по состоянию на момент написания статьи [18]. Очевидно, что ГК не может получиться больше 19, так как аминокислот всего 20.

Чтобы исключить путаницу с названиями, мы использовали для обозначения ГК кодирующей последовательности префикс Seq, а для обозначения ГК АА-индексов – префикс АА (т.е. ГК кодирующей последовательности именуются Seq1, Seq2 и т.д., а ГК физико-химических свойств – АА1, АА2 … АА19).

Для расчета АА-компонент использовано 553 полных АА-индекса, исключены 13 индексов с пропусками данных. Все 19 АА-компонент включены в анализ как ВФ (табл. 2). Конфигурация аминокислот на плоскости первых двух АА-компонент (50.2 % общей дисперсии) приведена на рисунке 1.

Общепринятая трактовка первой ГК (АА1) как оси “гидрофобность–гидрофильность” в целом подтверждается, хотя положение неполярного аланина несколько отличается от положения остальных неполярных аминокислот. Возможно, это не относится к пролину, который отличается от остальных аминокислот по многим свойствам, строго говоря, даже не является аминокислотой, и поэтому может оказаться на графике, где угодно.

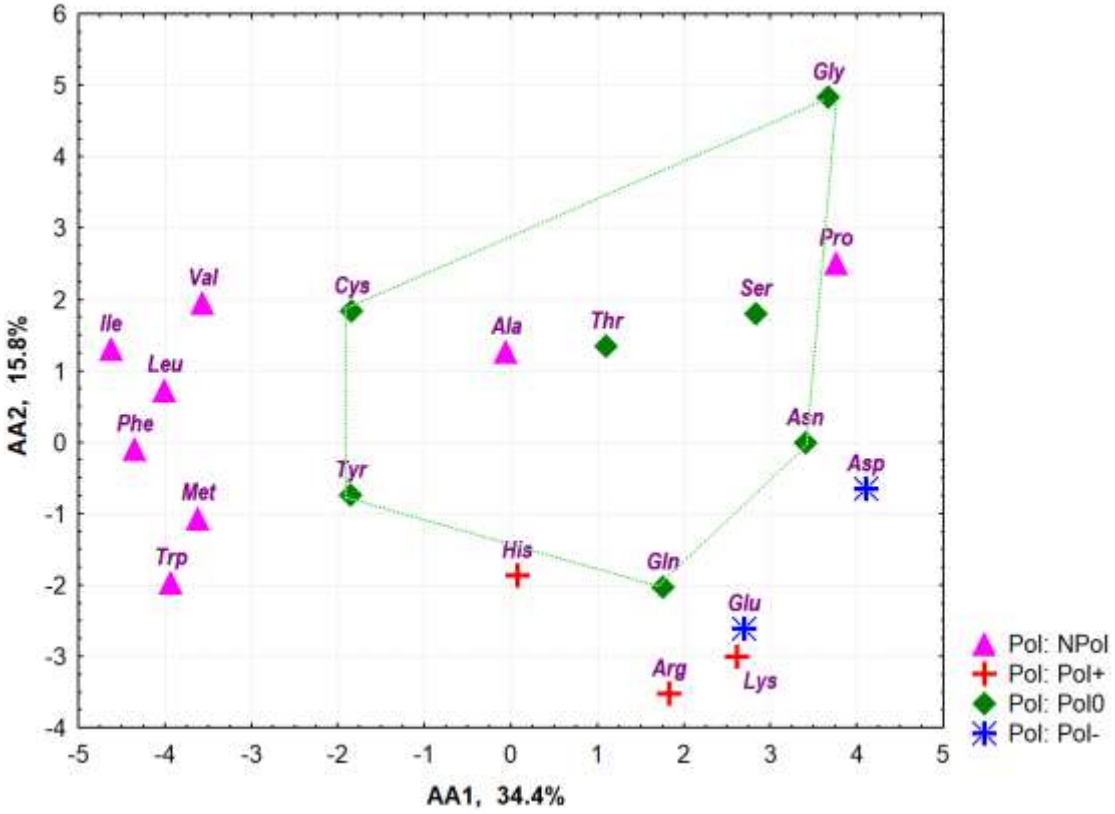

Рис. 1. Конфигурация аминокислот на плоскости первых двух ГК физико-химических свойств; NPol – неполярные (гидрофобные), Pol0 – полярные незаряженные, Pol+ – заряженные положительно, Pol- – заряженные отрицательно.

В предыдущих работах, относящихся к аминокислотным последовательностям гена *CYTB* человека и дрозофилы [1, 11], было найдено, что их первые ГК значимо коррелируют с ВС белка. Институт биологии развития общества Макса Планка (Тюбинген, Германия) предоставляет в свободное пользование бесплатный универсальный веб-сервис MPI Bioinformatics Toolkit [25] для биоинформатического анализа белков [26]. В частности, набор Quick2D этого сервиса содержит восемь различных алгоритмов распознавания ВС белка по ее аминокислотной последовательности: SS_PSIPRED, SS_SPIDER3, SS_PSSPRED4, SS_DEEГKNF, SS_NETSURFP2, TM_TMHMM, TM_PHOBIUS, TM_POLYPHOBIUS) [27–34]. Префикс SS означает, что алгоритм предсказывает расположение аминокислоты в α-спирали, TM

– в мембране. Воспользовавшись этим веб-сервисом, мы получили для исследуемой аминокислотной восемь двоичных последовательностей, характеризующих степень присутствия аминокислот в ВС (1, если аминокислота находится в α-спирали/мембране; 0 – в противном случае).

Обычно у мембранных белков α-спирали находятся в мембране. Однако у последовательности *SLC9A1* после 510-й позиции имеется С-терминальный цитозольный домен, содержащий α-спирали, но находящийся вне мембраны. Поэтому ГК для SS и TM последовательностей рассчитаны раздельно и взяты с этими же префиксами в качестве ВФ (табл. 2).

Таблица 2. Список ВФ для кодирующей последовательности: обозначения и описания

Обозначение	*N*	Описание	Список
Nucl	4	Содержание нуклеотидов во фрагменте кодирующей последовательности	A C G T
АК	20	Содержание аминокислот в аминокислотном фрагменте	Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val
AA	19	ГК физико-химических свойств аминокислот	AA1 AA2 … AA19
SS	5	ГК степени присутствия аминокислотного фрагмента в α-спирали	SS1 SS2 ... SS5
TM	3	ГК степени присутствия аминокислотного фрагмента в мембране	TM1 TM2 TM3

Для первых четырех ГК кодирующей последовательности и всех ВФ (табл. 2) вычислена матрица коэффициентов корреляции Пирсона (табл. 3).

Бутстреп-распределения корреляций между первой ГК кодирующей последовательности гена *SLC9A1* и ВФ

Для Seq1 – первой ГК кодирующей последовательности – классическим бутстрепом получено 1000 бутстреп-копий и вычислены их корреляции с ВФ. Тем самым найдены выборочные распределения коэффициентов корреляции в случае справедливости нулевой гипотезы (все генеральные корреляции равны нулю). Для них приведены нижние и верхние выборочные 1% и 5% квантили ±Q1% и ±Q5% (табл. 4, Null).

После этого расчет повторен еще 1000 раз с помощью якорного бустрепа (табл. 1). Тем самым получены выборочные распределения всех коэффициентов корреляции Seq1 с ВФ в случае справедливости альтернативной гипотезы (все генеральные корреляции равны рассчитанным по исходной выборке). Для них тоже приведены нижние и верхние выборочные 1% и 5% квантили ±Q1% и ±Q5% (табл. 4, Alt).

Расчеты проведены с помощью пакетов Statistica8 [13], PAST4 [12] и Jacobi4 [11].

РЕЗУЛЬТАТЫ

Корреляции

В таблице 3 приведены коэффициенты корреляции (×1000) первых 4 ГК кодирующей последовательности гена *SLC9A1* с ВФ (выборочно). Для Seq1 коэффициенты корреляции с некоторыми ВФ достаточно велики и, по всей видимости, отражают реальные структурные закономерности. В первую очередь нас интересует связь Seq1 с ВС белка, которая характеризуется компонентами SS1 и TM1 (степень нахождения фрагмента в α-спирали и мембране, соответственно). Поэтому, основываясь на таблицах

3 и 4, выберем дополнительно другие ВФ: AA1, A, T, Leu, Phe, Pro, коррелирующие с Seq1 или SS1 или TM1 выше порога 0.56 (см. след. подраздел). Подматрица корреляций этих ВФ и Seq1 без учета знака обработана методом квазиметрического двумерного шкалирования Крускала [35] (рис. 2).

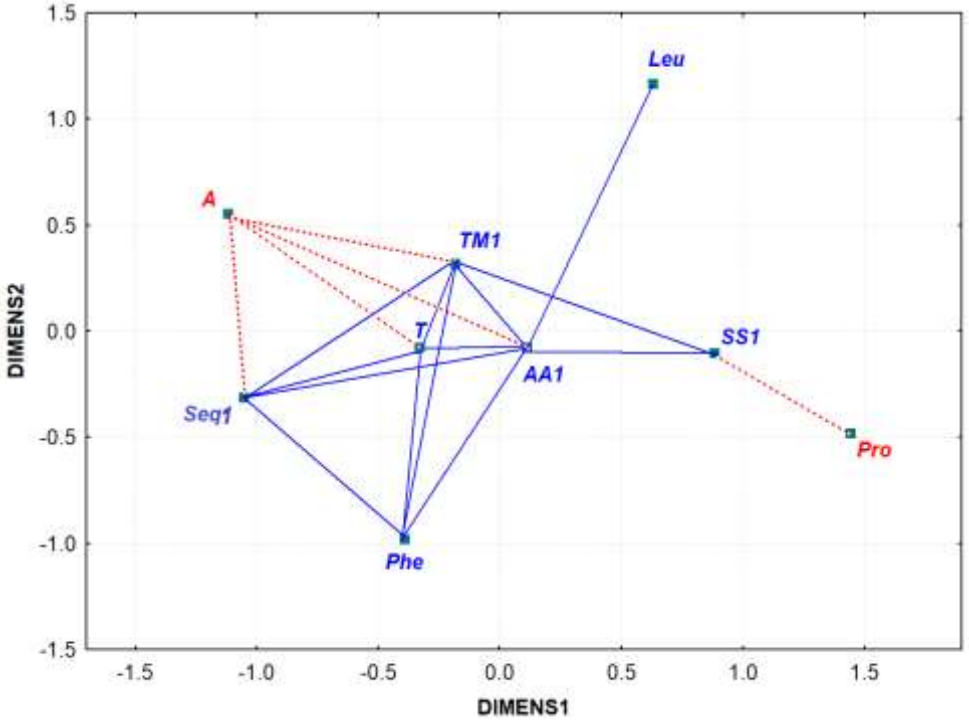

Рис. 2. Расположение Seq1 и коррелирующих с ней ВФ на плоскости квазиметрического двумерного шкалирования. Положительные связи выделены синим цветом, отрицательные – красным. TM1 (“трансмембранность”) находится ближе к Seq1, чем SS1 (“α-спиральность”).

Близость показателей на рисунке 2 означает их более высокую корреляцию по модулю. Положительные связи выделены синим цветом, отрицательные – красным. На рисунке видно, что Seq1 входит в практически полносвязную группу, состоящую из T, A, TM1 (маркера нахождения в мембране), AA1 (маркера гидрофобности) и Phe.

Сразу отметим, что SS1 – маркер присутствия фрагмента в α-спирали – хотя и находится вблизи TM1 и AA1, но, вместе с Leu и Pro, все же расположен на периферии группы. Это важно потому, что α-спирали обычно находятся в мембране, и тогда SS1 и TM1 практически совпадают. В нашем случае они расходятся после 510 позиции, отражая наличие α-спиралей, находящихся вне мембраны (рис. 3).

Фенилаланин на рис. 2 расположен несколько ближе к TM1, чем к SS1. Это отчетливо видно на его графике (рис. 4). На этом же рисунке видно, что график AA1 – маркера гидрофобности – занимает промежуточное положение между SS1 и TM1, в полном соответствии с рисунком 2. На рисунке 2 между Seq1 и TM1 расположились аденин (A) и тимин (T) с противоположными по знаку корреляциями. На рисунке 5 заметно большее расхождение их графиков сразу после 510 позиции. Это означает, что целесообразнее рассматривать разность T-A, чтобы они не погашали, а усиливали друг друга. И, действительно, корреляции разности T-A с Seq1 и ВФ выше, чем только одного тимина: Seq1 – 0.715; TM1 – 0.774; AA1 – 0.755. При сравнении графиков Seq1 и T-A (рис. 6) с рисунком 3 видно, что они больше соответствуют нахождению в мембране, чем в α-спирали.

Таблица 3. Коэффициенты корреляции (×1000) между первыми ГК кодирующей последовательности гена *SLC9A1* и ВФ (выборочно)

SLC9A1	Seq1	Seq2	Seq3	Seq4	SS1	TM1	AA1	AA2	_A	_C	_G	_T
SS1	347	-175	193	-2	1000	633	789	-40	-204	-224	-11	473
SS2	-168	-107	101	-13	-34	-34	-51	-34	27	-282	228	25
SS3	-35	-185	18	13	-77	-26	-40	139	-141	-108	110	153
SS4	-9	-78	165	-12	85	57	134	38	-61	-213	-15	310
SS5	-74	73	-72	-24	-172	-158	-145	-87	150	88	-59	-194
TM1	675	-299	369	12	633	1000	826	574	-697	6	42	713
TM2	-88	-186	31	17	-93	22	-26	102	-132	-112	132	123
TM3	-67	-148	57	19	-75	-78	-12	127	-80	-154	186	51
AA1	663	-215	357	0	789	826	1000	334	-577	-25	-138	808
AA2	602	-189	399	1	-40	574	334	1000	-802	247	152	454
AA3	131	356	-435	2	-444	-344	-365	-29	135	592	-518	-223
AA4	26	222	100	-46	328	83	129	-123	203	-305	143	-52
AA5	-220	-38	-237	28	-498	-392	-422	-102	135	193	-124	-220
AA6	-57	98	-349	37	-297	-316	-264	-157	186	310	-321	-189
AA7	202	33	101	-12	9	161	129	110	-59	32	-168	211
AA8	389	335	-273	4	113	73	217	-42	65	477	-774	251
AA9	188	254	-386	13	-185	-73	-165	-59	84	495	-391	-200
A	-683	330	-394	-7	-204	-697	-577	-802	1000	-282	-131	-656
C	559	316	-349	15	-224	6	-25	247	-282	1000	-665	-42
G	-431	-395	427	-9	-11	42	-138	152	-131	-665	1000	-221
T	616	-274	346	2	473	713	808	454	-656	-42	-221	1000
Ala	30	-167	160	20	105	209	121	85	-231	20	186	33
Arg	-373	50	-305	10	-199	-438	-405	-508	343	16	62	-459
Asn	-288	167	-77	-19	173	-159	-77	-377	403	-207	-36	-184
Asp	-286	41	-65	1	-500	-352	-495	-100	189	-50	125	-288
Cys	156	-65	-65	-9	141	84	166	118	-97	1	-51	159
Gln	-347	71	-256	19	62	-341	-251	-422	335	16	57	-445
Glu	-553	-55	-28	-15	-206	-419	-396	-405	438	-382	247	-338
Gly	109	-141	467	-31	25	298	89	616	-410	-247	584	87
His	-10	223	-192	-9	28	-174	14	-233	223	132	-399	43
Ile	373	70	28	-2	324	280	408	142	-30	104	-304	248
Leu	209	-332	248	4	501	510	578	187	-414	-121	106	469
Lys	-556	125	-182	-6	-20	-439	-312	-637	673	-429	55	-341
Met	-34	40	-3	19	-6	-78	-1	-111	93	-89	63	-75
Phe	663	-20	164	-15	459	592	652	256	-460	176	-311	651
Pro	-118	-14	-276	39	-722	-399	-544	181	-53	467	-171	-256
Ser	200	171	-298	-2	-430	-167	-359	210	-17	427	-240	-179
Thr	212	356	-175	5	-159	-37	-54	44	110	383	-391	-108
Trp	200	97	-78	-13	258	188	252	-106	-32	76	-71	30
Tyr	7	113	2	-20	266	69	150	-227	80	24	-133	30
Val	263	-324	504	11	184	471	427	546	-558	-236	325	514

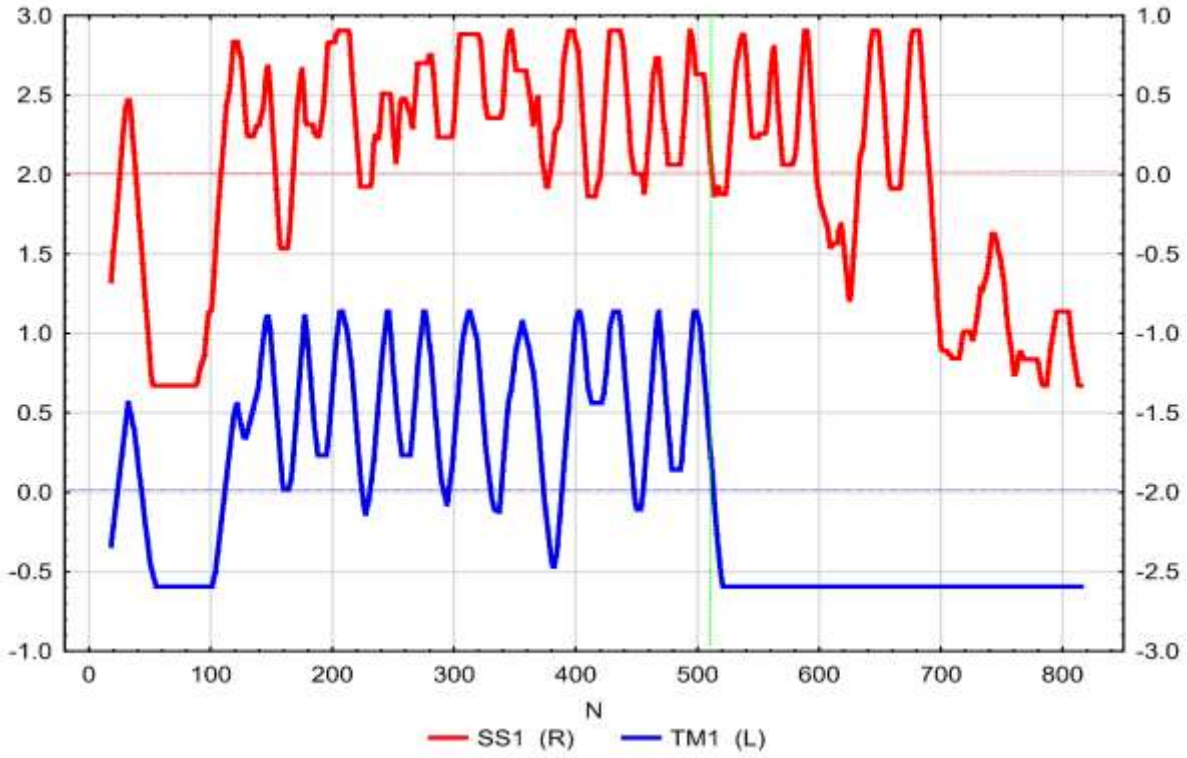

Рис. 3. Ген SLC9A1. SS1 – первая ГК степени присутствия АК-фрагмента в α-спирали; TM1 – первая ГК степени присутствия АК-фрагмента в мембране. $R = 0.633$. Обе кривые идут синхронно до 510-й позиции (α-спирали находятся в мембране). После 510-й позиции мембрана заканчивается и графики расходятся. SS1 до 700-й позиции отражает α-спирали, находящиеся в цитоплазме вне мембраны.

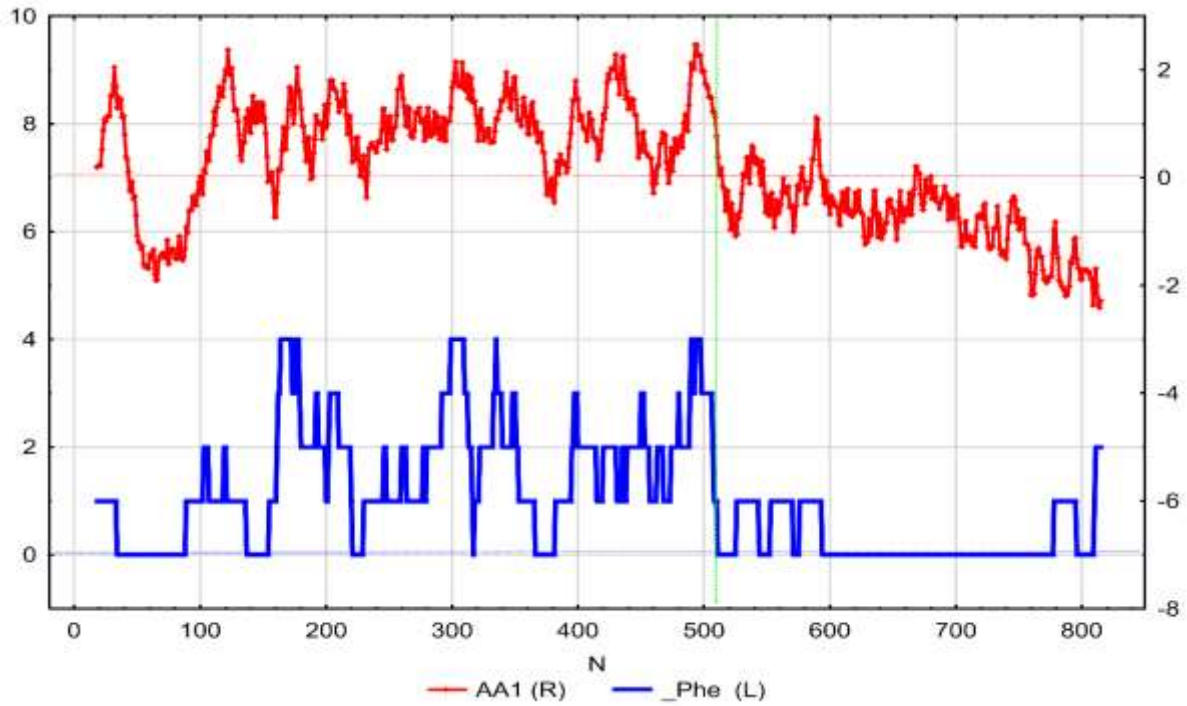

Рис. 4. Ген *SLC9A1*. Phe – содержание фенилаланина в АК-фрагменте; AA1 – первая ГК физико-химических свойств аминокислот (гидрофобность). $R = 0.652$. Динамика AA1 промежуточна между TM1 и SS1, а фенилаланина – больше похожа на TM1, чем SS1 (рис. 3).

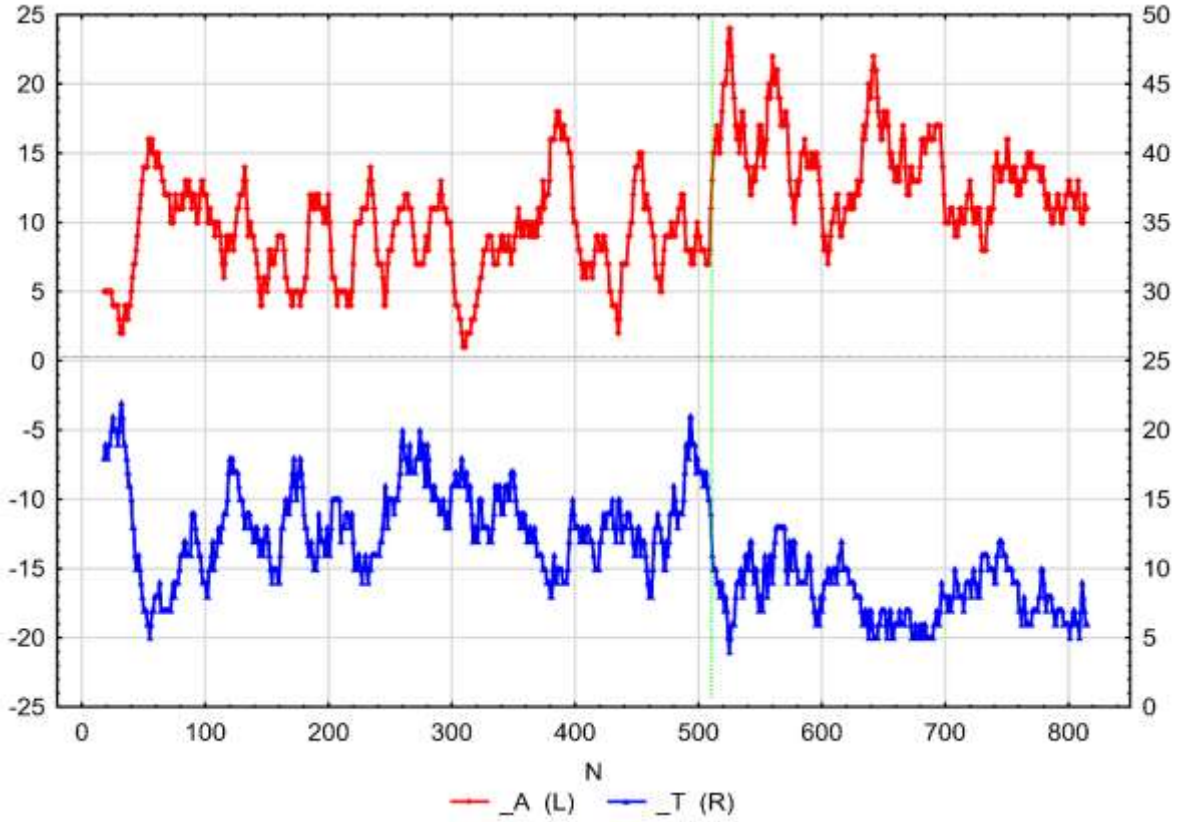

Рис. 5. Ген *SLC9A1*. A – содержание аденина в нуклеотидном фрагменте; T – содержание тимина в нуклеотидном фрагменте. $R = -0.656$. Графики тимина (T) и аденина (A) находятся в противофазе друг к другу. Заметно их расхождение сразу после 510 позиции.

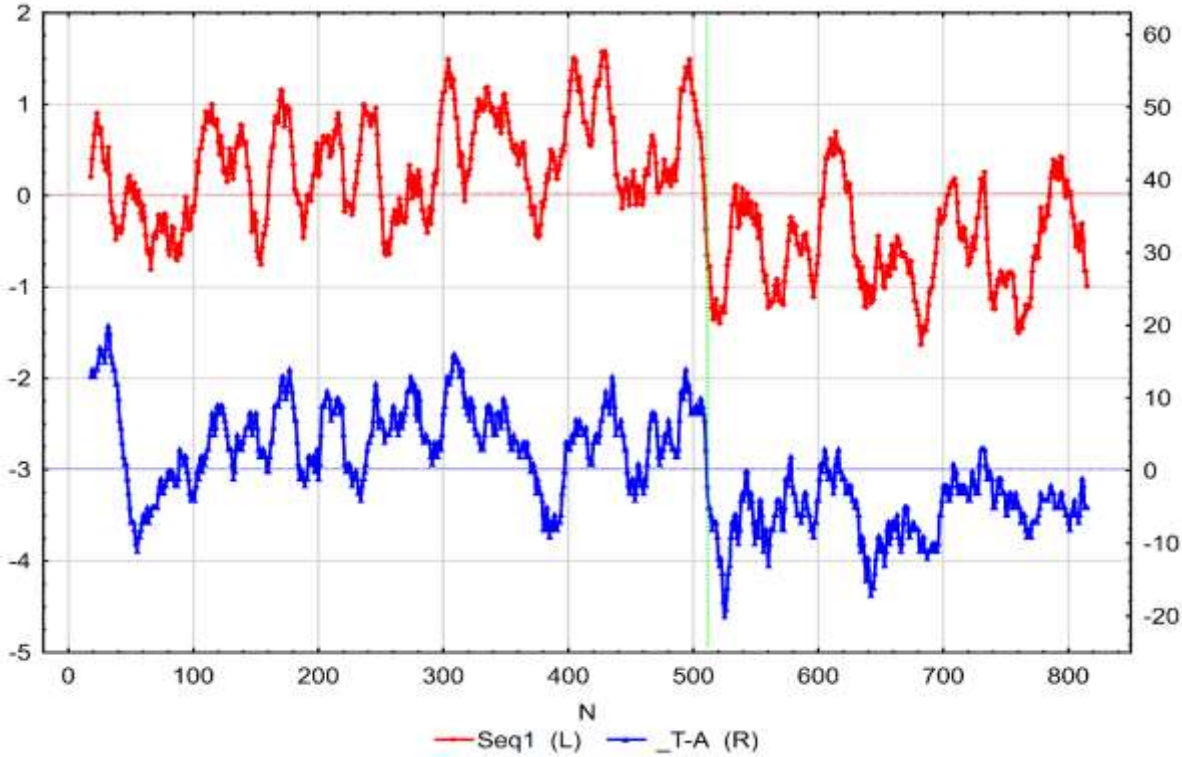

Рис. 6. Ген *SLC9A1*. Seq1 – первая ГК кодирующей последовательности; T-A – разность содержания тимина и аденина в нуклеотидном фрагменте. $R = 0.715$. При сравнении с рисунком 3 видно, что они больше соответствуют нахождению в мембране, чем в α-спирали.

Таблица 4. Коэффициенты корреляции (×1000) Seq1 с ВФ и квантили бутстреп-распределений (*Nboot* = 1000) при нулевой и альтернативной гипотезе.

Null	Seq1	-Q1%	-Q5%	+Q5%	+Q1%	Alt	Seq1	-Q1%	-Q5%	+Q5%	+Q1%
SS1	347	-91	-69	68	93	SS1	347	153	184	418	454
SS2	-168	-86	-64	73	95	SS2	-168	-253	-212	104	154
SS3	-35	-104	-69	73	101	SS3	-35	-168	-145	87	123
SS4	-9	-91	-73	64	88	SS4	-9	-128	-81	225	287
SS5	-74	-89	-72	66	92	SS5	-74	-195	-143	108	152
TM1	675	-88	-69	68	90	TM1	675	360	396	719	754
TM2	-88	-87	-66	76	108	TM2	-88	-225	-206	16	59
TM3	-67	-89	-64	71	91	TM3	-67	-154	-124	68	92
AA1	663	-93	-72	68	87	AA1	663	333	379	662	686
AA2	602	-90	-72	75	99	AA2	602	172	243	602	638
AA3	131	-91	-69	67	95	AA3	131	-197	-164	190	243
AA4	26	-94	-74	76	98	AA4	26	-118	-91	197	242
AA5	-220	-103	-73	69	95	AA5	-220	-409	-366	-157	-126
AA6	-57	-95	-73	68	88	AA6	-57	-230	-203	54	101
AA7	202	-92	-70	68	84	AA7	202	-83	-26	221	252
AA8	389	-96	-68	63	84	AA8	389	-16	44	400	444
AA9	188	-89	-71	65	80	AA9	188	-115	-54	212	250
A	-683	-99	-68	63	83	A	-683	-707	-668	-290	-185
C	559	-94	-72	66	90	C	559	-4	68	495	550
G	-431	-73	-60	73	109	G	-431	-476	-426	-48	7
T	616	-89	-72	64	81	T	616	257	297	613	633
Ala	30	-93	-67	70	99	Ala	30	-118	-82	158	222
Arg	-373	-92	-71	69	91	Arg	-373	-436	-415	-145	-81
Asn	-288	-95	-77	67	86	Asn	-288	-319	-294	-47	36
Asp	-286	-88	-67	66	94	Asp	-286	-410	-380	-202	-161
Cys	156	-91	-69	69	94	Cys	156	-119	-57	212	252
Gln	-347	-97	-68	73	98	Gln	-347	-404	-381	-153	-105
Glu	-553	-94	-74	71	95	Glu	-553	-578	-549	-240	-193
Gly	109	-80	-62	75	97	Gly	109	-54	-13	277	310
His	-10	-88	-74	66	97	His	-10	-190	-155	122	181
Ile	373	-98	-73	64	88	Ile	373	68	131	395	433
Leu	209	-87	-69	65	85	Leu	209	-31	25	337	378
Lys	-556	-89	-69	69	91	Lys	-556	-553	-525	-125	-76
Met	-34	-94	-74	68	85	Met	-34	-262	-203	87	146
Phe	663	-87	-68	63	83	Phe	663	240	300	618	657
Pro	-118	-91	-78	68	97	Pro	-118	-296	-259	-57	-22
Ser	200	-85	-68	67	84	Ser	200	-120	-87	213	269
Thr	212	-92	-66	68	85	Thr	212	-92	-51	280	319
Trp	200	-102	-76	69	86	Trp	200	3	39	280	305
Tyr	7	-91	-70	69	94	Tyr	7	-94	-56	191	249
Val	263	-93	-67	69	86	Val	263	74	116	405	453

Математическая биология и биоинформатика. 2021. Т. 16. № 2. doi: 10.17537/2021.16.299

Достоверность и якорный бутстреп

В таблице 4 через Q1% и Q5% обозначены найденные с помощью бутстрепа нижний и верхний квантили выборочных распределений коэффициентов корреляции Seq1 с ВФ с уровнем значимости 1% и 5%, соответственно. Это означает, что доверительными интервалами являются (–Q1%; +Q1%) и (–Q5%; +Q5%).

При нулевой гипотезе и *p-value* = 0.05 коэффициент корреляции не выходит за пределы ±0.078.

Теория, в случае справедливости гипотезы независимой выборки из генерального двумерного нормального распределения с нулевым коэффициентом корреляции, дает ±0.070 [29].

При *p-value* = 0.01 пределы равны ±0.109. Теория, при тех же предположениях, дает ±0.091. Разница, по-видимому, возникает из-за различий в конкретной конфигурации объектов на графике рассеяния (скаттер-диаграмме), которую игнорирует теория, но учитывает бутстреп.

При альтернативной гипотезе доверительный интервал зависит еще и от самого коэффициента корреляции. Как следует из расчетов якорного бутстрепа, для достоверности при *p-value* = 0.05 модуль коэффициента корреляции должен превышать 0.22, при *p-value* = 0.01 для некоторых конфигураций (связанных с C и G) нижняя граница корреляции поднимается до 0.56. Поэтому она выбрана как порог отсечения в предыдущем подразделе.

ОБСУЖДЕНИЕ

Для чего нужны корреляции с ВФ? Они полезны при интерпретации ГК. Сам по себе метод ГК (МГК) интерпретации не дает. Собственно говоря, никакой статистический метод интерпретации не дает. Обычно статистический метод дает рекомендацию принять или отвергнуть некоторую гипотезу, нулевую или альтернативную, на некотором уровне значимости. Считается, что гипотезу должен выдвигать специалист в предметной области (но при этом формулировать ее на языке математической статистики).

Отличие МГК (так же, как кластерного анализа и многомерного шкалирования) от других статистических методов заключается в том, что МГК не проверяет никаких гипотез. Он только помогает их выдвинуть. Это поисковый метод. Если мы изучаем изменчивость некоторой совокупности объектов, то неизбежно используем меры сходства/различия. Часто удобной математической моделью совокупности объектов является множество точек в многомерном геометрическом пространстве. Сходство/различие объектов отображается расстоянием между точками. Чем ближе точки друг к другу, тем более похожи друг на друга объекты, которых они представляют. МГК заключается в том, что в геометрическом пространстве ищутся направления, в проекции на которые дисперсия точек максимальна. Это и есть ГК. Они являются новыми признаками и часто отражают какие-то реальные закономерности. О природе найденных закономерностей МГК, естественно, сказать ничего не может. Зато могут сказать корреляции ГК с уже известными ВФ, если они имеются. Если корреляции достаточно велики и достоверны, это может служить подсказкой для интерпретации или даже самой интерпретацией, это уже компетенция специалиста в предметной области.

Что касается интерпретации Seq1 – найденной нами методом PCA-Seq первой ГК кодирующей последовательности гена *SLC9A1* – то очевидно, что Seq1 отражает главную изменчивость в изучаемой последовательности: нахождение в будущем кодируемого ею белка в мембране и вне ее (рис. 2, 6). Получается, что эта информация даже не закодирована, а записана открытым текстом в самой последовательности через повышенное содержание тимина и пониженное – аденина, хотя будет реализовываться

довольно сложными молекулярно-генетическими процессами. В эту закономерность хорошо укладывается и фенилаланин (Phe). Он кодируется всего двумя кодонами с максимально возможным содержанием тимина и полным отсутствием аденина – TTT, TTC. На рис. 1 он находится вблизи Seq1.

С физической точки зрения α-спираль выполняет роль проводника для передачи через мембрану: электрона – в одну сторону; протона, в следующем звене цепи – в обратную. Для этого нужна изоляция от влаги. Поэтому α-спираль обычно имеет гидрофобную оболочку, и она имеет смысл именно для прохождения через мембрану.

Из наших результатов получается, что разница в содержании тимина и аденина, по-видимому, кодирует не столько саму гидрофобность (“α-спиральность”), сколько именно будущее положение соответствующего фрагмента белка в мембране (“трансмембранность”). Для проверки этой гипотезы необходимы исследования других белков с α-спиралями, находящимися вне мембраны, поскольку для остальных трансмембранных белков гидрофобность, “α-спиральность” и “трансмембранность”, очевидно, практически неразличимы.

Одним из возможных биологических объяснений обсуждаемой связи является “конвергенция, то есть возникновение сходства между определенными участками генома в результате их независимой эволюции в одинаковом направлении под действием сходных ограничений” “… накладываемых ВС белков на порядок и состав аминокислот, а, следовательно, и на порядок и состав кодонов в генах (что и делает соответствующие участки генов более сходными, чем случайные последовательности)” [36, 37]. Иными словами, нуклеотидные кодоны отбираются эволюцией под ВС белка, а не наоборот. Первична именно функция, а молекулярно-генетическое наполнение может подбираться и отбираться в допускаемых функцией пределах.

В литературе, посвященной гену *SLC9A1*, несмотря на огромное количество публикаций, ничего похожего найти не удалось. Однако обнаружилось несколько ранних работ, оставшихся практически неизвестными из-за крайне низкой цитируемости, в которых на других генах найдена статистическая связь между повышенным содержанием тимина в нуклеотидных последовательностях, пониженным, соответственно, аденина, и гидрофобностью / “трансмембранностью” кодируемых ими белков [38–41]. Весьма вероятно, что закономерность имеет более общий характер и заслуживает большего внимания.

ЗАКЛЮЧЕНИЕ

Предложенный нами ранее метод PCA-Seq является универсальным и позволяет для любой последовательности рассчитывать числовые ГК, не используя никакой информации о природе элементов этой последовательности. Для содержательной интерпретации разумно использовать корреляции найденных ГК с рядами числовых характеристик элементов или фрагментов исходной последовательности. Для оценки достоверности получаемых корреляций в настоящей работе используется новый способ – якорный бутстреп для последовательностей. Отличие от классического заключается в том, что в якорном бутстрепе не делается традиционного предположения, что элементы последовательности являются выборкой из независимых одинаково распределенных случайных величин. Таким образом, полная технологическая цепочка для исследуемой последовательности включает расчет ГК, биологическую интерпретацию через расчет корреляций ГК с ВФ и бутстреп-оценку достоверности этих корреляций.

Для примера взята кодирующая последовательность гена *SLC9A1* (*NHE1*). Судя по литературе, кодируемый белок состоит из двух блоков – трансмембранного домена, состоящего из 12 гидрофобных α-спиралей, и С-терминальной хвостовой части, расположенной за пределами мембраны, в которой тоже находятся 5 гидрофобных α-

спиралей. При расчете ГК эта информация не использовалась. Тем не менее, по первой ГК четко проявились различия между этими блоками. Доверительные интервалы для корреляций рассчитывались с помощью классического и якорного бутстрепов (*Nboot* = 1000). Достоверные корреляции первой ГК, превышающие значение 0.6, выявились с гидрофобностью / “трансмембранностью” соответствующих фрагментов аминокислотной последовательности, содержанием в них фенилаланина, а также разностью содержания тимина и аденина в нуклеотидных фрагментах. Похожая закономерность была найдена ранее другими авторами на других генах и весьма вероятно, что она имеет более общий характер.

Работа выполнена при поддержке гранта РФФИ № 19-07-00658-а и Бюджетного проекта ИЦиГ СО РАН №0259-2021-0009.

Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Конфликт интересов отсутствует.

СПИСОК ЛИТЕРАТУРЫ

1. Ефимов В.М., Ефимов К.В., Ковалева В.Ю. Метод главных компонент и его обобщения для последовательности любого типа (PCA-Seq). *Вавиловский журнал генетики и селекции.* 2019. Т. 23. № 8. С. 1032–1036. doi: 10.18699/VJ19.584
2. Duras T. The fixed effects PCA model in a common principal component environment. *Communications in Statistics-Theory and Methods.* 2020. P. 1–21. doi: 10.1080/03610926.2020.1765255
3. Efron B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*. 1979. V. 7. P. 1–26. doi: 10.1214/aos/1176344552
4. Timmerman M.E., Kiers H.A., Smilde A.K. Estimating confidence intervals for principal component loadings: a comparison between the bootstrap and asymptotic results. *British Journal of Mathematical and Statistical Psychology*. 2007. V. 60. № 2. P. 295–314. doi: 10.1348/000711006X109636
5. Linting M., Meulman J.J., Groenen P.J., Van der Kooij A.J. Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap. *Psychological methods*. 2007. V. 12. № 3. P. 359. doi: 10.1037/1082-989X.12.3.359
6. Efimov V., Efimov K., Kovaleva V. Anchored Bootstrap. In: *2020 Cognitive Sciences, Genomics and Bioinformatics (CSGB). IEEE.* 2020. P. 32–35. doi: 10.1109/CSGB51356.2020.9214598
7. Hendus-Altenburger R., Vogensen J., Pedersen E.S., Luchini A., Araya-Secchi R., Bendsoe A.H., Nanditha Shyam Prasad, Andreas Prestel, Marité Cardenas, ... Kragelund B.B. The intracellular lipid-binding domain of human Na+/H+ exchanger 1 forms a lipid-protein co-structure essential for activity. *Communications Biology*. 2020. V. 3. № 1. P. 1–18.doi: 10.1038/s42003-020-01455-6
8. Koch A., Schwab A. Cutaneous pH landscape as a facilitator of melanoma initiation and progression. *Acta Physiologica.* 2019. V. 225. № 1. P. e13105. doi: 10.1111/apha.13105
9. Böhme I., Schönherr R., Eberle J., Bosserhoff A.K. Membrane Transporters and Channels in Melanoma. In: *Reviews of Physiology, Biochemistry and Pharmacology*. 2020. P. 1–106. doi: 10.1007/112_2020_17
10. Pethő Z., Najder K., Carvalho T., McMorrow R., Todesca L.M., Rugi M., Bulk E., Chan A., Löwik C.W.G.M., Reshkin S.J., Schwab A. pH-channeling in cancer: How pH-dependence of cation channels shapes cancer pathophysiology. *Cancers*. 2020. V. 12. № 9. P. 2484. doi: 10.3390/cancers12092484
11. Polunin D., Shtaiger I., Efimov V. JACOBI4 software for multivariate analysis of biological data. *bioRxiv*. 2019. P. 803684. doi: 10.1101/803684

12. Hammer Ø., Harper D.A., Ryan P.D. PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica.* 2001. V. 4. № 1. URL: http://palaeo-electronica.org/2001_1/past/issue1_01.htm (дата обращения: 05.09.2021).
13. Hill T., Lewicki P. *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining.* Tulsa, Okla., UK: StatSoft Ltd. 2006. 719 p. ISBN: 9781884233593
14. *NCBI*. URL: https://www.ncbi.nlm.nih.gov (дата обращения: 05.09.2021).
15. Gower J.C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*. 1966. V. 53. № 3–4. P. 325–338. doi: 10.1093/biomet/53.3-4.325
16. Ней М., Кумар С. *Молекулярная эволюция и филогенетика.* Киев: КВІЦ, 2004. ISBN: 966-7192-53-9
17. Ефимов В.М., Мельчакова М.А., Ковалева В.Ю. Геометрические свойства эволюционных дистанций. *Вавиловский журнал генетики и селекции*. 2015. Т. 17. № 4/1. С. 714–723.
18. *AAindex* (v.9.2 от 13.02.2017). URL: https://www.genome.jp/aaindex (дата обращения: 05.09.2021).
19. Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*. 2008. V. 36. № 1. P. D202–D205. doi: 10.1093/nar/gkm998
20. Sneath P.H.A. Relations between chemical structure and biological activity in peptides. *Journal of Theoretical Biology*. 1966. V. 12. № 2. P. 157–195. doi: 10.1016/0022-5193(66)90112-3
21. Hellberg S., Sjoestroem M., Skagerberg B., Wold S. Peptide quantitative structure-activity relationships, a multivariate approach. *Journal of Medicinal Chemistry*. 1987. V. 30. № 7. P. 1126–1135. doi: 10.1021/jm00390a003
22. Sandberg M., Eriksson L., Jonsson J., Sjöström M., Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *Journal of Medicinal Chemistry.* 1988. V. 41. № 14. P. 2481–2491. doi: 10.1021/jm9700575
23. Kosky A.A., Dharmavaram V., Ratnaswamy G., Manning M.C. Multivariate analysis of the sequence dependence of asparagine deamidation rates in peptides. *Pharmaceutical Research*. 2009. V. 26. № 11. P. 2417–2428. doi: 10.1007/s11095-009-9953-8
24. Zbacnik N.J., Henry C.S., Manning M. C. A Chemometric Approach Toward Predicting the Relative Aggregation Propensity: Aβ (1–42). *Journal of Pharmaceutical Sciences*. 2020. V. 109. № 1. P. 624–632. doi: 10.1016/j.xphs.2019.10.014
25. *MPI Bioinformatics Toolkit.* URL: https://toolkit.tuebingen.mpg.de (дата обращения: 05.09.2021).
26. Zimmermann L., Stephens A., Nam S.Z., Rau D., Kübler J., Lozajic M., Gabler F., Söding J., Lupas A.N., Alva V. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *Journal of Molecular Biology*. 2018. V. 430. № 15. P. 2237–2243. doi: 10.1016/j.jmb.2017.12.007
27. Jones D.T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*. 1999. V. 292. № 2. P. 195–202. doi: 10.1006/jmbi.1999.3091
28. Heffernan R., Yang Y., Paliwal K., Zhou Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics.* 2017. V. 33. № 18. P. 2842–2849. doi: 10.1093/bioinformatics/bty1006

29. Yan R., Xu D., Yang J., Walker S., Zhang Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Scientific Reports*. 2013. V. 3. P. 2619. doi: 10.1038/srep02619
30. Wang S., Peng J., Ma J., Xu J. Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports.* 2016. V. 6. P. 18962. doi: 10.1038/srep18962
31. Klausen M.S., Jespersen M.C., Nielsen H., Jensen K.K., Jurtz V.I., Soenderby C.K., Sommer M.O.A., Winther O., Nielsen M., Petersen B., Marcatili P. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics.* 2019. V. 87. № 6. P. 520–527. doi: 10.1002/prot.25674
32. Krogh A., Larsson B., Von Heijne G., Sonnhammer E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*. 2001. V. 305. № 3. P. 567–580. doi: 10.1006/jmbi.2000.4315
33. Käll L., Krogh A., Sonnhammer E.L. A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*. 2004. V. 338. № 5. P. 1027–1036. doi: 10.1016/j.jmb.2004.03.016
34. Käll L., Krogh A., Sonnhammer E.L. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*. 2005. V. 21. № 1. P. i251–i257. doi: 10.1093/bioinformatics/bti1014
35. Kruskal J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika.* 1964. V. 29. № 2. P. 1–27. doi: 10.1007/BF02289565
36. Кель Э.А., Колчанов Н.А., Соловьев В.В. Конвергентное происхождение повторов в генах, кодирующих глобулярные белки. Анализ факторов, обуславливающих наличие прямых повторов. *Журн. общ. биол.* 1988. Т. 49. № 3. С. 343–354.
37. Колчанов Н.А., Кель Э.А., Соловьев В.В. Конвергентное происхождение повторов в генах, кодирующих глобулярные белки. Моделирование конвергентного возникновения прямых повторов. *Журн. общ. биол*. 1988. Т. 49. № 6. С. 723–728.
38. Chen C. P., Kernytsky A., Rost B. Transmembrane helix predictions revisited. *Protein Science.* 2002. V. 11. № 12. P. 2774–2791. doi: 10.1110/ps.0214502
39. Lesnik T., Reiss C. Detection of transmembrane helical segments at the nucleotide level in eukaryotic membrane protein genes. *Biochem. Mol. Biol. Int.* 1998. V. 44. № 3. P. 471–479. doi: 10.1080/15216549800201492
40. Nakashima H., Yoshihara A., Kitamura K.I. Favorable and unfavorable amino acid residues in water-soluble and transmembrane proteins. *J. Biomedical Science and Engineering*. 2013. V. 6. № 1. P. 36–44. doi: 10.4236/jbise.2013.61006
41. Vakirlis N., Acar O., Hsu B., Coelho N.C., Van Oss S.B., Wacholder A., Medetgul-Ernar K., Bowman II R.W., Hines C.P., Iannotta J. et all. *De novo* emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature communications*. 2020. V. 11. № 1. P. 1–18. doi: 10.1038/s41467-020-14500-z

Рукопись поступила в редакцию 10.05.2021, переработанный вариант поступил 30.07.2021.
Дата опубликования 10.09.2021.

Principal Components of Genetic Sequences: Correlations and Significance

Efimov V.M.[1,2,3,4], Efimov K.V.[5], Kovaleva V.Yu.[2], Matushkin Yu.G.[1]

[1]*Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[2]*Institute of Systematics and Ecology of Animals SB RAS, Novosibirsk, Russia*
[3]*Novosibirsk State University, Novosibirsk, Russia*
[4]*Tomsk State University, Tomsk, Russia*
[5]*HSE School of Economics, Moscow, Russia*

Abstract. Any numerical series can be decomposed into principal components using singular spectral analysis. We have recently proposed a new analysis method – PCA-Seq, which allows calculating numerical principal components for a sequence of elements of any type. In particular, the sequence may be composed of nucleotide base pairs or amino acid residues. Two questions inevitably arise about interpretation of the obtained principal components and about the assessment of their reliability. For interpretation of the symbolic sequence principal components, it is reasonable to evaluate their correlations with numerical characteristics of the sequence elements. To assess the significance of correlations between sequences, one should bear in mind that standard significance criteria are based on the assumption of independence of observations, which, as a rule, is not fulfilled for real sequences. The article discusses the use of an anchor bootstrap technique for these purposes also previously developed by the authors of the article. In this approach it is assumed, that points of a metric space can represent the objects. When taken together they make up some fixed structure in it, in particular, a sequence. The objects are assigned the same random integer weights as in the classical bootstrap. This is sufficient to obtain the bootstrap distribution of the correlation coefficients and assess their significance. The coding sequence of the *SLC9A1* gene (synonyms *APNH*, *NHE1*, *PPP1R143*) were taken as an example of use the anchor bootstrap technique in the genetic sequence analysis. Significant correlations of the first principal component were revealed with the hydrophobicity / "transmembraneity" of the corresponding fragments of the amino acid sequence, the phenylalanine content in them, as well as the difference in the T- and A-content in the corresponding nucleotide fragments. Earlier a similar pattern was found by other authors for other genes. Very likely, that it is of a more general nature.

Key words: *SSA, PCA-Seq, SLC9A1 (NHE1) gene, CDS, protein secondary structure, external factors, anchor bootstrap.*