УДК 519.222
DOI: 10.25559/SITITO.14.201803.672-678

# СРАВНИТЕЛЬНЫЙ АНАЛИЗ СВЯЗАННЫХ РЯДОВ И ИХ ПРИРАЩЕНИЙ НА ОСНОВЕ ДИСКРИМИНАНТНОГО АНАЛИЗА

С.Н. Истомина
Московский авиационный институт (национальный исследовательский университет), г. Москва, Россия

# COMPARATIVE ANALYSIS OF RELATED SEQUENCES AND THEIR INCREMENTS ON THE BASIS OF DISCRIMINANT ANALYSIS

Svetlana N. Istomina
Moscow Aviation Institute (National Research University), Moscow, Russia



**Ключевые слова**

Ортогональное многомерное преобразование; главная компонента; кластерный анализ; корреляция; доверительный интервал; регрессионный анализ; критерий Фишера; ортологичные белки; дискриминантный анализ; обучающая выборка.

**Аннотация**

Работа посвящена изучению связи длин ортологичных белков четырех организмов, один из которых принят за базисный (в сумме более 1200 белков). Использованы методы многомерного статистического анализа, который применяется к парам, тройкам и четвёркам (строкам), составленным из длин ортологичных белков. Таких строк от 200 до 400. Анализ парных корреляций, ортогональное преобразование и кластерный анализ позволили выделить два однородных кластера четвёрок длин. Параллельно изучали приращения длины ортологичного белка относительно базисного организма. Показали, что строки образуют неоднородную выборку, а приращения образуют однородную выборку. Далее задача состояла в расширении кластеров строками с неполными данными. Показали, что для этого кластерный анализ не применим, поэтому использовали дискриминантный анализ с обучающей выборкой – кластеризацией с полными данными. Получено 100-процентное разделение всех неполных строк по кластерам; с последующим описанием по кластерам зависимости длин от базисных. Проверена адекватность полученных уравнений регрессии. В результате статистического анализа сделаны следующие выводы. Для множества длин ортологичных рядов получен обобщающий фактор, назовем его размером ортологичного объекта из 4-х длин ортологичных беков. Для данной задачи получены такие размеры объектов, причем их средние групповые отличаются они образуют два отдельных интервала значений, по одному для каждой группы из полученных другими методами. Для рядов приращений длин ортологичных белков из объектов по четыре анализ показал всеми методами однородность множества. Показано, что длины ортологичных белков имеют значимую автокорреляцию, как и бывает у рядов, связанных с одним и тем же базисным рядом.

**Keywords**

Orthogonal multivariate transformation, principal component, cluster analysis, correlation, confidence interval, regression analysis, F-test, orthologous proteins, discriminant analysis, training set.

**Abstract**

The article is devoted to the study of the relationship between the lengths of orthologous proteins of four organisms, one of which is taken as the basic one ( more than 1200 proteins in total). The methods of multivariate statistical analysis are used, it is applied to pairs, triples and fours (strings) composed of lengths of orthologous proteins. The number of such lines is from 200 to 400.
The analysis of pair correlations, orthogonal transformation and cluster analysis allowed us to distinguish two homogeneous clusters of four-lengths. At the same time, we studied the increments of the length of the orthologous protein relative to the basic organism. We showed that the lines form a non-uniform sample, and the increments form a homogeneous sample. Then the task was to expand the clusters with rows with incomplete data. It was shown that cluster analysis is not applicable for this task, so we used discriminant analysis with a training sample — clustering with complete data. A 100 percent separation of all incomplete rows by clusters was obtained; with the following description of the length dependences of clusters on the base. The adequacy of the resulting regression equations was tested. As a result of statistical analysis, the following conclusions were made. For a set of lengths of orthologous series, a generalizing factor was ob-

Об авторе:

**Истомина Светлана Николаевна**, кандидат химических наук, доцент, Московский авиационный институт (национальный исследовательский университет) (125993, Россия, г. Москва, Волоколамское ш., д. 4), ORCID: http://orcid.org/0000-0002-9206-2641, stomina_sn45@mail.ru

tained, let's call it the size of an orthologic object from 4 lengths of orthologous proteins. For the given task such sizes of objects were obtained, and their average group values differ, they form two separate ranges of values, one for each group of the values obtained by other methods. For series of increments of the lengths of orthologous proteins from objects of four, an analysis performed by all methods showed homogeneity of the set. It was shown that the lengths of orthologous proteins have significant autocorrelation, as is the case with rows associated with the same basic series.

## Введение

В данной работе исследовали связи длин ортологичных белков с учетом свойства ортологичности белков, которое связывает ряды длин белков по четыре в строке. Для этого использовали дискриминантный метод многомерного статистического анализа. Содержание рассматриваемой задачи и направление, к которому она относится, описаны в [2–5].

## Теоретическая часть: результаты

Методы статистического анализа многомерных данных применены для выявления взаимосвязей рядов длин ортологичных белков по четырем организмам. Обозначим S – столбец длин белков базисного организма, T, N и P – столбцы длин белков трех организмов ортологичных базисному. В приложении эти буквы обозначают соответственно цианобактерии, токсоплазмы, неоспоры и плазмодии. В дополнительных материалах к [1] длины белков приведены парами столбцов (ST), (SN) и (SP), где число пар длин в столбцах разное: 318, 279 и 245 соответственно. Там же приведена большая часть этих данных, организованная в матрицу по 4 ортологичных белка (STNP), таких строк по 4 оказалось 198. В биоинформатике принято рассматривать также связи приращений длин ортологичных белков относительно длин белков базисного организма. Работа [6] посвящена исследованию взаимосвязей длин ортологичных белков, а также их приращений относительно длин базисного ряда попарно для базисного ряда и ему ортологичного ряда длин и ряда приращений отдельно.

В работе выделены два этапа. На первом исследовали в качестве исходных данных матрицу длин ортологичных белков (STNP), где строки упорядочены по возрастанию длин белков организма S, а также аналогичную матрицу приращений этих длин относительно длин белков базисного организма (это матрицы 4 на 198). На втором этапе исследовали те значения длин ортологичных белков, для которые не вошли в матрицу 4 на 198.

В таблице 1 приведены числовые характеристики рядов длин и приращений длин ортологичных белков, составляющих описанные матрицы по 4 значения в строке.

**Таблица 1.** Основные числовые характеристики для рядов длин и приращений длин ортологичных белков

Table 1. Basic numerical characteristics for series of lengths and length increments of orthologous proteins

| Ряд | Среднее значение | Среднеквадратическое отклонение | Значение дисперсии |
|---|---|---|---|
| S | 393,33 | 228,417 | 52174,343 |
| T | 575,82 | 372,87 | 139034,312 |
| N | 521,67 | 358,66 | 128639,381 |
| P | 504,28 | 348,85 | 121698,334 |
| T-S | 182,49 | 229,27 | 52567,317 |
| N-S | 128,34 | 200,04 | 40016,895 |
| P-S | 110,95 | 175,91 | 30945,642 |

Выполнена проверка нормальности распределений этих вариационных рядов. Имеется достаточное приближение к нормальному распределению длин каждого из этих рядов, что позволяет использовать кластерный, факторный и регрессионный анализ.

Для рядов длин ортологичных белков по 4 в строке провели корреляционный анализ для выяснения наличия и силы связи между рядами длин ортологичных белков. Коэффициенты парной корреляции для всех длин ортологичных белков приведены в таблице 2 с доверительными интервалами.

**Таблица 2.** Парные коэффициенты корреляции с доверительными интервалами для матрицы длин

Table 2. Paired correlation coefficients with confidence intervals for the length matrix

| Корреляция | S | T | N | P |
|---|---|---|---|---|
| S | 1 | 0,814±0,06 | 0,859±0,072 | 0,897±0,03 |
| T | 0,814±0,06 | 1 | 0,941±0,024 | 0,89±0,045 |
| N | 0,859±0,072 | 0,941±0,024 | 1 | 0,899±0,039 |
| P | 0,897±0,03 | 0,89±0,045 | 0,899±0,039 | 1 |

Как видим, ряды длин S, T, N и P связаны практически линейно. Высокое значение парных корреляций длин белков может быть следствием связанности рядов ортологичных белков с базисным рядом.

Коэффициенты парных корреляций приращений длин белков относительно базисных, приведены в таблице 3 также вместе с доверительными интервалами.

**Таблица 3.** Парные коэффициенты корреляции с доверительными интервалами для матрицы приращений длин

Table 3. Paired correlation coefficients with confidence intervals for the length increment matrix

| Корреляция | S | T-S | N-S | P-S |
|---|---|---|---|---|
| S | 1 | 0,327±0,18 | 0,399±0,18 | 0,48±0,15 |
| T-S | 0,327±0,18 | 1 | 0,834±0,06 | 0,673±0,09 |
| N-S | 0,399±0,18 | 0,834±0,06 | 1 | 0,647±0,12 |
| P-S | 0,48±0,15 | 0,673±0,09 | 0,647±0,12 | 1 |

Здесь корреляции слабее, хотя и значимы. Отметим, что все корреляции положительны по обеим таблицам. Но часть значений коэффициентов корреляции, например, для пары рядов (TN) парная корреляция 0,94. Поэтому определяли значения автокорреляция, которая оказалась значимой, но не достаточно большой, чтобы исказить парные уравнения регрессии.

Судя по результатам корреляционного анализа ряды длин белков T, S и N попарно связаны линейной регрессией, но нельзя получить надежную многомерную регрессию. Поэтому нами был проведен факторный анализ методом главных компонент с ортогональным преобразованием исходных данных.

В результате такого преобразования для 198 строк по 4 значениям длин ортологичных белков, как объектов, выделена одна главная компонента для которой начальное собственное значение равно 3,65, а для других эти значения 0,2 и меньше.

Причем полная объясненная дисперсия корреляционной матрицы для рядов длин белков составляет 91,26 % для первой компоненты. Значит, можно пренебречь другими тремя компонентами, объясняющими 5,20 %; 2,23 % и 1,31 % соответственно. Эта главная компонента включает в линейную комбинацию, то есть в латентную (скрытую) переменную Y, значения длин всех четырех рядов и имеет вид:

$Y$=0,933*S+0,954*T+0,968*N+0,965*P.

Здесь коэффициенты Y являются значениями факторных нагрузок признаков. Они показывают вес каждого признака – длин белков четырех разных организмов, ортологичные белки которых исследуются. Интересно, что факторные нагрузки приблизительно равны и строке значений четырех длин соответствует одно значение Y, которое характеризует общий или суммарный (с некоторой поправкой) размер объекта из четырех длин ортологичных белков.

Для приращений длин белков выделена также одна главная компонента, у которой начальное собственное значение больше нуля, но она объясняет только 67,97 % от полной дисперсии, а включение еще двух мало значимых компонент не оправдано.

При анализе парных регрессий рядов по всем точкам несмотря на высокие значения коэффициентов корреляции заметны сгущения точек в одной области, отличающие ее от других областей на всех парах рядов. Поэтому было решено провести кластерный анализ исходных данных (198 объектов по четыре), чтобы с помощью группировки объектов и вычислений расстояний между разными группами выделить компактно расположенные и удаленные друг от друга группы объектов. За меру расстояния между объектами выбрали евклидово расстояние.

Для выделения кластеров использовали наиболее распространенный иерархический агломеративный метод, алгоритм которого представим в виде последовательности операций. Принадлежность объекта к конкретному кластеру выясняли методом k-средних кластерного анализа, основанного на процедуре оптимального разбиения начальных данных на определенное число кластеров. В итоге получили три кластера объектов для длин белков и 2 кластера для приращений длин, причем последние кластеры состоят из трех одних и тех же строк.

Результаты кластеризации для длин белков приведены в таблице 4.

**Таблица 4.** Распределение по кластерам объектов из длин рядов S, T, N, P

**Table 4.** The distribution over clusters of objects from the lengths of the rows S, T, N, P

| № кластера | Количество наблюдений | Наблюдения |
|---|---|---|
| 1 | 158 | не вошедшие в кластеры №2 и №3 |
| 2 | 37 | 3-6, 8,9, 11, 14, 16, 19, 21, 23, 24, 27,29, 37, 46, 49, 52, 54, 62, 66, 84, 86, 100, 102-104, 107,108, 117, 118, 132, 150, 156, 191, 195 |
| 3 | 3 | 1,2,10 |

Для объектов с приращениями результаты приведены в таблице 5.

**Таблица 5.** Распределение по кластерам объектов приращений длин белков

**Table 5.** Distribution of protein length increments over object clusters

| № кластера | Количество наблюдений | Наблюдения |
|---|---|---|
| 1 | 195 | все, кроме объектов 1,2,10 |
| 2 | 3 | 1,2,10 |

Для значений главной компоненты (Y), вычисленных по приведенной выше линейной комбинации в соответствии с описанной кластеризацией, в таблице 6 приведены средние значения и концы интервалов.

**Таблица 6**. Диапазон значений латентной переменной Y с учетом кластеризации

**Table 6.** The range of values of the latent variable Y with regard to clustering

| Кластер | Количество наблюдений | Минимальное значение | Максимальное значение | Среднее значение |
|---|---|---|---|---|
| №1 | 158 | 281,16 | 3106,48 | 1441,8803 |
| №2 | 37 | 1958,11 | 5172,01 | 3459,8920 |
| №3 | 3 | 6511,72 | 7467,00 | 7105,965 |

Далее получили уравнения линейной регрессии, попарно связывающие ортологичные ряды. Это позволило не только представить характер связи каждого ортологичного белка с базисным, но и сравнить эти связи между собой.

Зависимость ряда от ряда длин белка S рядов длин белков T, P и N соответственно, причем по двум кластерам отдельно:

$T_{1r}$ = 1.007 * S + 130.808,$F_{факт}$ = 146, n = 158;

$T_{2r}$ = 0,598 * S + 88,142,$F_{факт}$ = 10,548, n = 37.

$P_{1r}$ = 1.02 * S + 44,346,$F_{факт}$ = 423,702, n=158;

$P_{2r}$ = 0,807 * S + 391.875,$F_{факт}$ = 32,078, n=37.

$N_{1r}$ = 0,947 * S + 92,517,$F_{факт}$ = 243,261, n=158;

$N_{2r}$ = 0,786 * S + 422,522,$F_{факт}$= 21,069, n=37.

Графики этих уравнений регрессии для наглядного сравнения результатов будут приведены ниже.

Кроме описанных регрессий были рассчитаны также уравнения для пар (TP), (TN) и (NP) также по кластерам. Приведем только одно из них.

$T_{1r}$ = 0,98 * N + 64,449,$F_{факт}$ =54,27, n=158;

$T_{2r}$ = 0,765 * N + 262,39,$F_{факт}$ = 10,158, n=37.

Графики этих уравнений приведены на рисунке 1 вместе с эмпирическими точками.

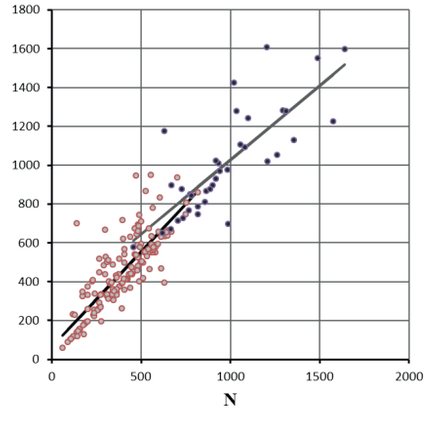


**Рис. 1.** Графики уравнений парной линейной регрессии белков N и T и точек с координатами, равными значениям длин белков

**Fig. 1.** Graphs of equations of pairwise linear regression of N and T proteins and points with coordinates equal to protein lengths

Изучение рядов длин ортологичных белков четверками, как связанных именно базисным рядом, убедительно показало, что имеется возможность по этому признаку разделить объекты на два кластера. Поэтому встал вопрос нельзя ли определить, какому из выделенных кластеров принадлежат те белки, что не вошли в матрицу из 198 объектов. Напомним, что для организмов T, P и N таких белков осталось соответственно: для T – 110, для P - 47 и для N – 81. Их длины дополняют соответствующие ряды длин T, P и N, а также для базисного ряда длин S, составляющего ортологичные пары к ним.

В таблице 7 приведены число значений дополненных рядов и другие характеристики.

Таблица 7. Числовые характеристики дополненных пар ортологичных рядов (ST), (SN) и (SP)

Table 7. Numerical characteristics of augmented pairs of orthologous series (ST), (SN) and (SP)

| | $S_t$ | T | $S_n$ | N | $S_p$ | P |
|---|---|---|---|---|---|---|
| Число значений ряда | 318 | 318 | 279 | 279 | 245 | 245 |
| | | | | | | |
| Стд. отклонение | 233.636 | 361.230 | 230.817 | 337.23 | 224.204 | 331.198 |
| Размах | 1162.00 | 2539.00 | 1135.00 | 2430.00 | 1162.00 | 2339.00 |
| Минимум | 38.00 | 37.00 | 65.00 | 59.00 | 38.00 | 36.00 |
| Максимум | 1200.00 | 2576.00 | 1200.00 | 2489.00 | 1200.00 | 2375.00 |

Чтобы распределить эти дополнительные значения по кластерам сначала использовали кластерный анализ для пар, состоящих из длин базисных белков и ортологичных ему дополненных рядов, а результаты сравнили с предыдущим исследованием объектов по четыре в строке. Результаты оказались не убедительными на границах кластеров. Поэтому применили дискриминантный анализ для каждой пары рядов длин (ST), (SN), (SP), образованных значениями, не входящими в первую четырехмерную матрицу. Реализован дискриминантный анализ с двумя обучающими выборками для классификации именно по этим двум кластерам. с обучающей выборкой, в качестве которой использовали кластеры, полученные по объектам из длин по 4 в строке. Ниже приведены результаты этого исследования.

При реализации дискриминантного анализа определяется дискриминантная переменная, которая используется для предсказания принадлежности новых значений пар рядов длин (ST), (SN), (SP) соответственно к первой или второй группе, основами для которых послужат кластеры, полученные при исследовании четверками.

Для каждой пары длин рядов первая обучающая выборка состояла из 158 значений, а вторая из 37.

Присутствие разделяющих особенностей у каждой группы и значимость различий между переменными после дискриминации видно по следующим показателям в таблице 8.

**Таблица 8.** Критерий равенства групповых средних для пар рядов (S,P), (S,N) и (S,T).

**Table 8.** The criterion for the equality of group averedges for pairs of rows (S, P), (S, N) and (S, T)

| | Лямбда Уилкса | F-критерий | Степеней свободы 1 | Степеней свободы 2 | Уровень значимости |
|---|---|---|---|---|---|
| S | 0,497 | 195,079 | 1 | 193 | 0,0001 |
| P | 0,389 | 302,590 | 1 | 193 | 0,0001 |
| S | 0,497 | 195,079 | 1 | 193 | 0,0001 |
| N | 0,472 | 215,712 | 1 | 193 | 0,0001 |
| S | 0,497 | 195,079 | 1 | 193 | 0,000 |
| T | 0,472 | 215,712 | 1 | 193 | 0,000 |

Получены весьма значимые различия между группами. Таким образом, ни одна переменная не может быть исключена из списка без ухудшения качества классификации значений. О хорошем качестве дискриминации объектов по два значения в строке говорят данные таблицы 9.

**Таблица 9.** Анализ коэффициентов дискриминантной функции

**Table 9.** Analysis of the coefficients of the discriminant function

| Функция и ряды | Собственное значение | Объясненная дисперсия, % | Кумулятивный процент, % | Каноническая корреляция |
|---|---|---|---|---|
| 1; S,N | 1,315[a] | 100,0 | 100,0 | 0,754 |
| 1; S,P | 1,568[a] | 100,0 | 100,0 | 0,781 |
| 1; S,T | 1,315[a] | 100,0 | 100,0 | 0,754 |

Собственные значения матрицы, выражаемые как отношение межгрупповой суммы квадратов к внутригрупповой сумме квадратов, в таблице 24 позволяют судить о том, что обе группы, полученные в результате классификации объектов каждой пары рядов, дискриминируются одной, достаточно точно подобранной, функцией (так как собственные значения превышают единицу). Каждая из трех функций учитывает 100% дисперсии переменных. Кроме того, корреляция между исходными и расширенными данными - сильная – порядка 0,75 по каждой группе. Значимость дискриминантной функции по каждому критерию позволяет использовать ее для классификации значений.

Вклад каждой независимой переменной в различие двух групп можно оценить по нормированным коэффициентам канонической дискриминантной функции, полученные в результате вычета средних значений и деления на дисперсию соответствующих коэффициентов дискриминантных функций. Для пар рядов (S,P); (S,N); (S,T) эти коэффициенты соответственно равны

(0,02;0,984); (0,495;0,614) и (0,495;0,614). Из этого видно, что наибольший вклад в классификацию вносят не базисные, а ряды, длины которых связаны с базисным по признаку ортологичности.

В результате применения дискриминантного анализа было получено деление по обучающим выборкам, значений длин рядов ортологичных белков, не кластеризованных по четырем рядам:

Из 46 несгруппированных значений пары рядов (SP) 46 были отнесены к первой группе, образованной первым кластером, а 1 значение – ко второй, соответственно образованной вторым кластером.

Из 135 несгруппированных значений пары рядов (ST) 95 были отнесены к первой группе, а 40 значений – ко второй.

Из 81 несгруппированного значения пары рядов (SN) 69 были отнесены к первой группе, а 12 значений – ко второй.

После дискриминации исходные ряды длин S, N, P и T, приведенные на сайте (ссылка на 3 стр. этой работы), парами, базисный и ему ортологичный ряд, оказались также парами разделены на две группы. Исключены лишь 3 четверки длин белков, которые были выделены всеми методами, и классифицированы здесь как выбросы. Более подробное теоретическое исследование этого вопроса выполнено в [7–17].

После дискриминации для полных рядов по расширенным кластерам получены линейные уравнения регрессии, с зависимостью от ряда длин белка S рядов длин белков T, P и N соответственно. Приведены вместе со значениями критерия Фишера и числом значений ряда.

$T_{1г} = 0{,}89 * S + 158{,}084, F_{факт} = 190$, n = 254;

$T_{2г} = 0{,}467 * S + 708{,}555, F_{факт} = 8{,}5$, n = 61.

$P_{1г} = 1.011 * S + 59{,}723, F_{факт} = 534$, n=204;

$P_{2г} = 0{,}814 * S + 383{,}800, F_{факт} = 33$, n=38.

$N_{1г} = 0{,}954 * S + 130{,}419, F_{факт} = 209$, n=227;

$N_{2г} = 0{,}509 * S + 624{,}252, F_{факт} = 8{,}6$, n=49.

Как видим, регрессии по расширенным группам после дискриминации, адекватны описываемым данным и мало отличаются от регрессий для групп четырехмерной кластеризации.

На рисунках 4, 5 и 6 приведены точки, координаты которых – длины соответствующих белков и для сравнения графики итоговых уравнений регрессии после дискриминации вместе с графиками регрессий, полученных по данным кластеризации 198 объектов.

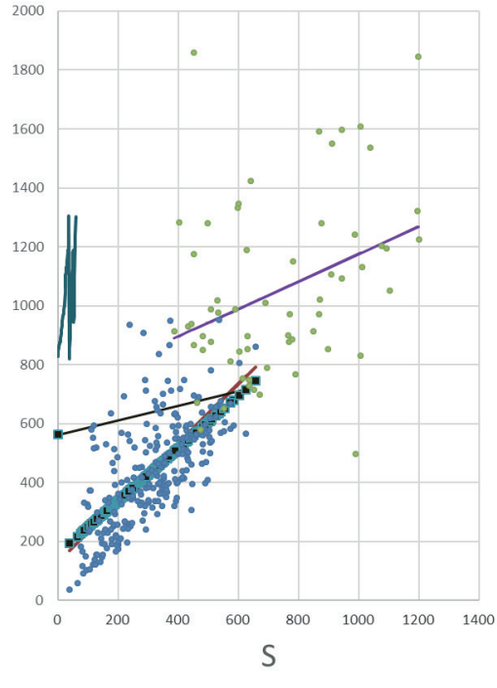


**Рис. 4**. Уравнения регрессий для рядов S,T для групп четырехмерной кластеризации и после дискриминации

**Fig. 4.** Regression equations for the S, T series for four-dimensional clustering groups and after discrimination

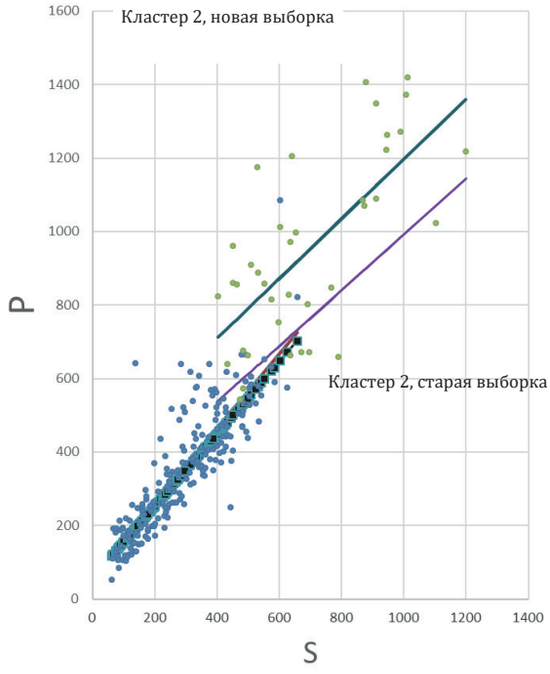


**Рис. 5**. Уравнения регрессий для рядов S,P для групп четырехмерной кластеризации и после дискриминации

**Fig. 4.** Regression equations for the S, T series for four-dimensional clustering groups and after discrimination

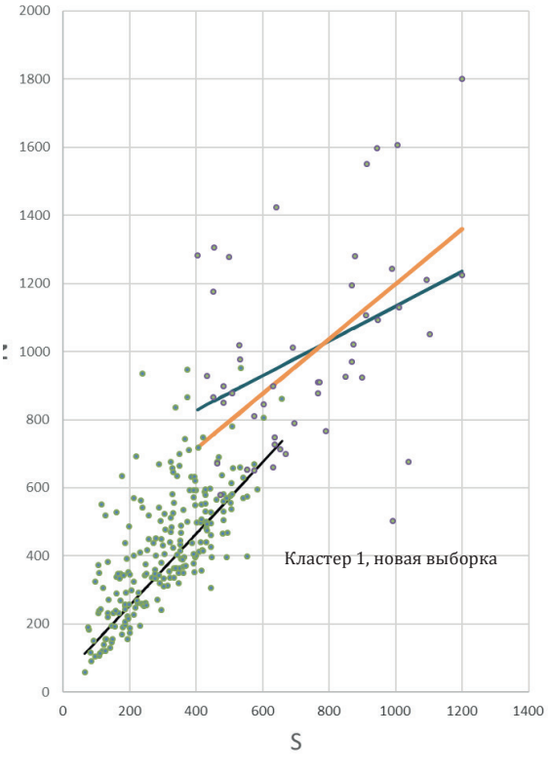


**Рис. 6**. Уравнения регрессий для рядов S,N для групп четырехмерной кластеризации и после дискриминации

**Fig. 6.** Regression equations for the S, N series for four-dimensional clustering groups and after discrimination

На приведенных выше рисунках видим не принципиальное различие регрессий на фоне разброса данных. Это подтверждает обоснованность результатов дискриминантного анализа. Но что очень важно, подтверждает также наличие двух кластеров рядов длин при учете ортологичности белков по разным организмам.

## Заключение

В работе многомерными методами изучены как ряды длин, так и ряды приращений длин относительно базисного ряда с учетом того, что свойство ортологичности проявляется в связанности рядов длин. При этом показано различие свойств множества длин и множества приращений белков: последние менее коррелированы, но однородны, а длины связаны практически линейно и разделяются на две группы, большая часть которых это малые и средние длины, а меньшую группу составляют белки с большой длиной.

Для множества длин ортологичных рядов получен обобщающий фактор, назовем его размером ортологичного объекта из 4-х длин ортологичных беков. Для данной задачи получены такие размеры объектов, причем их средние групповые отличаются они образуют два отдельных интервала значений, по одному для каждой группы из полученных другими методами.

Для рядов приращений длин ортологичных белков из объектов по четыре анализ показал всеми методами однородность множества.

Показано, что длины ортологичных белков имеют значимую автокорреляцию, как и бывает у рядов, связанных с одним и тем же базисным рядом.

## Список использованных источников


[1] *Seliverstov A.V., Zverkov O.A., Istomina S.N., Pirogov S.A., Kitsis P.S.* Comparative Analysis of Apicoplast-Targeted Protein Extension Lengths in Apicomplexan Parasites // BioMed Research International. 2015. Vol. 2015, Article ID 452958. DOI: 10.1155/2015/452958

[2] *Seliverstov A.V., Rubanov L.I., Shilovsky G.A., Zverkov O.A., Lyubetsky V.A.* Longevity in euarchontoglires: lost genes as a determinant // FEBS Open Bio. 2018. Vol. 8, Suppl. 1. Pp. 456-457. DOI: 10.1002/2211-5463.12453

[3] *Lyubetsky V.A., Gershgorin R.A., Gorbunov K.Yu.* Chromosome structures: reduction of certain problems with unequal gene content and gene paralogs to integer linear programming // BMC Bioinformatics. 2017. Vol. 18, Article 537, 18 pp. DOI: 10.1186/s12859-017-1944-x

[4] *Lyubetsky V.A., Korotkova D.D., Ivanova A.S., Rubanov L.I., Seliverstov A.V., Zverkov O.A., Nesterenko A.M., Tereshina M.B., Zaraisky A.G.* Novel transmembrane protein c-Answer revealed by bioinformatic screening of genes present only in well regenerating animals // FEBS Journal. 2017. Vol. 284, issue S1. Pp. 155. DOI: doi.org/10.1111/febs.14174

[5] *Korotkova D.D., Ivanova A.S., Lyubetsky V.A., Seliverstov A.V., Martynova N.Yu., Nesterenko A.M., Tereshina M.B., Zaraisky A.G.* Novel FGF-signaling modulator c-Answer revealed by bioinformatics screening for genes present only in well-regenerative animals // Mechanisms of Development. 2017. Vol. 145. Pp. S49. DOI: doi.org/10.1016/j.mod.2017.04.089

[6] *Истомина С.Н.* Сравнительный анализ связанных рядов: длин ортологичных белков и их приращений относительно длин базисных белков // Современные информационные технологии ИТ образования. 2015. Том 11, № 2. С. 594-599. URL: https://elibrary.ru/item.asp?id=26167553 (дата обращения: 05.08.2018).

[7] *Gorbunov K.Yu., Lyubetsky V.A.* The minimum-cost transformation of graphs // Doklady Mathematics. 2017. Vol. 96, no. 2. Pp. 503–505. DOI: 10.1134/S1064562417050313

[8] *Горбунов К.Ю., Любецкий В.А.* Линейный алгоритм минимальной перестройки структур // Проблемы передачи информации. 2017. Том 53, вып. 1. С. 60-78. URL: https://elibrary.ru/item.asp?id=28876248 (дата обращения: 05.08.2018).

[9] *Lyubetsky V.A.* Linear algorithm for minimal rearrangement of structures // Problems of Information Transmission. 2017. Vol. 53, issue 1. Pp. 55–72. DOI: 10.1134/S0032946017010057

[10] *Gorbunov K.Yu., Lyubetsky V.A.* A linear algorithm for the shortest transformation of graphs with different operation costs // Journal of Communications Technology and Electronics. 2017. Vol. 62, no. 6. Pp. 653–662. DOI: 10.1134/S1064226917060092

[11] *Korolev S.A., Zverkov O.A., Seliverstov A.V., Lyubetsky V.A.* Ribosome reinitiation at leader peptides increases translation of bacterial proteins // Biology Direct. 2016. Vol. 11, no. 1, Article 20. 6 pp. DOI: 10.1186/s13062-016-0123-8

[12] *Lyubetsky V.A., Gershgorin R.A., Seliverstov A.V., Gorbunov K.Yu.* Algorithms for reconstruction of chromosomal structures // BMC Bioinformatics. 2016. Vol. 17, no. 1, Article 40. 23 pp. DOI: 10.1186/s12859-016-0878-z

[13] *Rubanov L.I., Seliverstov A.V., Zverkov O.A., Lyubetsky V.A.* A method for identification of highly conserved elements and evolutionary analysis of superphylum Alveolata // BMC Bioinformatics. 2016. Vol. 17, no. 1, Article 385. 16 pp. DOI: 10.1186/s12859-016-1257-5

[14] *Gorbunov K.Yu., Gershgorin R.A., Lyubetsky V.A.* Rearrangement and Inference of Chromosome Structures // Molecular Biology. 2015. Vol. 49, no. 3. Pp. 327–338. DOI: 10.1134/S0026893315030073

[15] *Zverkov O.A., Seliverstov A.V., Lyubetsky V.A.* A Database of Plastid Protein Families from Red Algae and Apicomplexa and Expression Regulation of the moeB Gene // BioMed Research International. 2015. Vol. 2015, Article ID 510598. 5 pp. DOI: 10.1155/2015/510598

[16] *Rusin L.Yu., Lyubetskaya E.V., Gorbunov K.Yu., Lyubetsky V.A.* Reconciliation of Gene and Species Trees // BioMed Research International. 2014. Vol. 2014, Article ID 642089. 22 pp. DOI: 10.1155/2014/642089

[17] *Lyubetsky V.A., Korolev S.A., Seliverstov A.V., Zverkov O.A., Rubanov L.I.* Gene expression regulation of the PF00480 or PF14340 domain proteins suggests their involvement in sulfur metabolism // Computational Biology and Chemistry. 2014. Vol. 49. Pp. 7-13. DOI: 10.1016/j.compbiolchem.2014.01.001


Поступила 05.08.2018; принята в печать 02.09.2018;
опубликована онлайн 30.09.2018.

## References


[1] Seliverstov A.V., Zverkov O.A., Istomina S.N., Pirogov S.A., Kitsis P.S. Comparative Analysis of Apicoplast-Targeted Protein Extension Lengths in Apicomplexan Parasites. *BioMed Research International*. 2015; 2015:452958. DOI: 10.1155/2015/452958

[2] Seliverstov A.V., Rubanov L.I., Shilovsky G.A., Zverkov O.A., Lyubetsky V.A. Longevity in euarchontoglires: lost genes as a determinant. *FEBS Open Bio*. 2018; 8(Suppl. 1):456-457. DOI: 10.1002/2211-5463.12453

[3] Lyubetsky V.A., Gershgorin R.A., Gorbunov K.Yu. Chromosome structures: reduction of certain problems with unequal gene content and gene paralogs to integer linear programming. *BMC Bioinformatics*. 2017; 18:537. 18 pp. DOI: 10.1186/s12859-017-1944-x

[4] Lyubetsky V.A., Korotkova D.D., Ivanova A.S., Rubanov L.I., Seliverstov A.V., Zverkov O.A., Nesterenko A.M., Tereshina M.B., Zaraisky A.G. Novel transmembrane protein c-Answer revealed by bioinformatic screening of genes present only in well regenerating animals. *FEBS Journal*. 2017; 284(S1):155. DOI: doi.org/10.1111/febs.14174

[5] Korotkova D.D., Ivanova A.S., Lyubetsky V.A., Seliverstov A.V., Martynova N.Yu., Nesterenko A.M., Tereshina M.B., Zaraisky A.G. Novel FGF-signaling modulator c-Answer revealed by bioinformatics screening for genes present only in well-regenerative animals. *Mechanisms of Development*. 2017; 145:S49. DOI: doi.org/10.1016/j.mod.2017.04.089

[6] Istomina S.N. Comparative analysis of related series: lengths of orthologous proteins and their increments relative to the lengths of basic proteins. *Modern Information Technology and IT-education*. 2015; 11(2):594-599. Available at: https://elibrary.ru/item.asp?id=26167553 18 (accessed 05.08.2018). (In Russian)

[7] Gorbunov K.Yu., Lyubetsky V.A. The minimum-cost transformation of graphs. *Doklady Mathematics*. 2017; 96(2):503–505. DOI: 10.1134/S1064562417050313

[8] Gorbunov K.Yu., Lyubetsky V.A. Linear algorithm for minimal rearrangement of structures. *Problems of Information Transmission*. 2017; 53(1):60-78. Available at: https://elibrary.ru/item.asp?id=28876248 (accessed 05.08.2018). (In Russian)

[9] Lyubetsky V.A. Linear algorithm for minimal rearrangement of structures. *Problems of Information Transmission*. 2017; 53(1):55–72. DOI: 10.1134/S0032946017010057

[10] Gorbunov K.Yu., Lyubetsky V.A. A linear algorithm for the shortest transformation of graphs with different operation costs. *Journal of Communications Technology and Electronics*. 2017; 62(6):653–662. DOI: 10.1134/S1064226917060092

[11] Korolev S.A., Zverkov O.A., Seliverstov A.V., Lyubetsky V.A. Ribosome reinitiation at leader peptides increases translation of bacterial proteins. *Biology Direct*. 2016; 11(1):20. 6 pp. DOI: 10.1186/s13062-016-0123-8

[12] Lyubetsky V.A., Gershgorin R.A., Seliverstov A.V., Gorbunov K.Yu. Algorithms for reconstruction of chromosomal structures. *BMC Bioinformatics*. 2016; 17(1):40. 23 pp. DOI: 10.1186/s12859-016-0878-z

[13] Rubanov L.I., Seliverstov A.V., Zverkov O.A., Lyubetsky V.A. A method for identification of highly conserved elements and evolutionary analysis of superphylum Alveolata. *BMC Bioinformatics*. 2016; 17(1):385. 16 pp. DOI: 10.1186/s12859-016-1257-5

[14] Gorbunov K.Yu., Gershgorin R.A., Lyubetsky V.A. Rearrangement and Inference of Chromosome Structures. *Molecular Biology*. 2015; 49(3):327–338. DOI: 10.1134/S0026893315030073

[15] Zverkov O.A., Seliverstov A.V., Lyubetsky V.A. A Database of Plastid Protein Families from Red Algae and Apicomplexa and Expression Regulation of the moeB Gene. *BioMed Research International*. 2015; 2015:510598. 5 pp. DOI: 10.1155/2015/510598

[16] Rusin L.Yu., Lyubetskaya E.V., Gorbunov K.Yu., Lyubetsky V.A. Reconciliation of Gene and Species Trees. *BioMed Research International*. 2014; 2014:642089. 22 pp. DOI: 10.1155/2014/642089

[17] Lyubetsky V.A., Korolev S.A., Seliverstov A.V., Zverkov O.A., Rubanov L.I. Gene expression regulation of the PF00480 or PF14340 domain proteins suggests their involvement in sulfur metabolism. *Computational Biology and Chemistry*. 2014; 49:7-13. DOI: 10.1016/j.compbiolchem.2014.01.001


Submitted 05.08.2018; revised 02.09.2018;
published online 30.09.2018.

About the author:

**Svetlana N. Istomina**, Candidate of Sciences in Chemistry, Associate Professor, Moscow Aviation Institute (National Research University) (4 Volokolamsk highway, Moscow 125993, Russia), ORCID: http://orcid.org/0000-0002-9206-2641, stomina_sn45@mail.ru