***Доклады Международной конференции "Математическая биология и биоинформатика"**. Под ред. В.Д. Лахно. Том 8. Пущино: ИМПБ РАН, 2020. Статья № e32. doi: 10.17537/icmbb20.25*

Многомерный статистический анализ пространственных структур ДНК в интерфейсах комплексов семейства гомеодомен - ДНК: спиральные параметры

Полозов Р.В.[1], Грохлина Т.И.[2], Панченко Л.А.[3], Иванов В.В.[4]

[1]*Институт теоретической и экспериментальной биофизики РАН, Пущино, Россия*
[2]*ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН, Пущино, Россия*
[3]*Биологический факультет МГУ им. М.В. Ломоносова, Москва, Россия*
[4]*Объединенный институт ядерных исследований, Дубна, Россия*

polrob@mail.ru

Изучение строения семейства комплексов гомеодомен – ДНК и классификация комплексов белок-ДНК и их интерфейсов по физико-химическим, геометрическим параметрам связана с решением ряда задач вычислительного и статистического характера. В данной работе представлены результаты анализа спиральных параметров ДНК в интерфейсах 75 комплексов семейства гомеодомен – ДНК. Цель работы – исследовать взаимосвязь комплексов друг с другом, иметь возможность проследить их эволюцию и дать основу для формулировок правил узнавания и построения моделей узнавания. Для анализа данных мы используем методы многомерного статистического анализа: для проверки гипотезы независимости признаков – критерий Спирмена; ANOVA, непараметрического дисперсионного анализа – тест Крускала-Уоллиса, медианный критерий Брауна – Муда, которые показали, что структура комплекса не влияет на геометрические параметры ДНК; методы кластерного и дискриминантного анализа. Кластерный анализ разбил исходную выборку на три однородные группы, класса. Дискриминантный анализ показал, что значения корректных вероятностей классификации больше 75 %. Полученные результаты, возможно, позволят определить в семействе комплексов гомеодомен – ДНК филогенетические связи комплексов (по родству) или по типологии (по сходству).

Ключевые слова: *гомеодомен – ДНК, спиральные параметры, интерфейс, деформация ДНК, классификация, многомерный статистический анализ, строение семейства гомеодомен – ДНК.*

Multi-Dimensional Statistical Analysis of Spatial Structures of DNA in Interfaces Complexes of Homeodomain-DNA Family: Helical Parameters

Polozov R.V.[1], Grokhlina T.I.[2], Panchenko L.A.[3], Ivanov V.V.[4]

[1]*Institute of Theoretical and Experimental Biophysics, RAS, Pushchino, Russia*
[2]*IMPB RAS – Branch of KIAM RAS, Pushchino, Russia*
[3]*Biological Faculty, Lomonosov Moscow State University, Moscow*
[4]*Joint Institute for Nuclear Research, Dubna, Russia*

The study of the structure of the family of homeodomain – DNA complexes and the classification of protein-DNA complexes and their interfaces according to physicochemical, geometric parameters are associated with solving a number of problems of a computational and statistical nature. This paper presents the results of the analysis of the helical parameters of DNA in the interfaces of 75 complexes of the homeodomain – DNA family. The aim of the work is to investigate the interconnection of complexes with each other, to be able to trace their evolution and provide a basis for formulating the rules of recognition and building recognition models. For data analysis, we use the methods of multivariate statistical analysis: to test the hypothesis of independence of features – Spearman's test; ANOVA, nonparametric analysis of variance – Kruskal-Wallis test, Brown-Mud median test, which showed that the complex structures does not affect the geometric parameters of DNA; methods of cluster and discriminant analysis. Cluster analysis divided the original sample into three homogeneous groups, classes. The discriminant analysis showed that the values of the correct classification probabilities are more than 75 %. The results obtained will probably make it possible to determine the phylogenetic relations of the complexes (by kinship) or by typology (by similarity) in the family of complexes homeodomain - DNA.

***Proceedings of the International Conference "Mathematical Biology and Bioinformatics".** Ed. V.D. Lakhno. Vol. 8. Pushchino: IMPB RAS, 2020. Paper No. e32. doi: 10.17537/icmbb20.25*

Key words: homeodomain-DNA, helical parameters, interface, DNA deformation, classification, multidimensional statistical analysis, the structure of the homeodomain family – DNA.

Введение

Регуляция процессов транскрипции белками (факторами транскрипции) осуществляется посредством их избирательного связывания со специфическими участками ДНК (биомолекулярное узнавание), другой способ регуляции транскрипции – посредством коротких одноцепочечных РНК. Структурные типы факторов транскрипции и их комплексов с ДНК весьма разнообразны и образуют несколько семейств. Одно из этих семейств – это семейство гомеодоменов и их комплексов с ДНК, которое регулируют процессы дифференцировки клеток. Специфические комплексы гомеодоменов с ДНК образуются при связывании белка с ДНК в B-форме по широкому желобу и формируют достаточно плотную упаковку с атомными группами α-спирали гомеодомена («узнающая спираль»). Эта упаковка дает основу для формирования надлежащей, вполне определенной ориентации атомов ДНК и белка, и образования специфических атомных контактов между белком и ДНК, в том числе, через молекулу воды.

В этой области тесного соприкосновения белка и ДНК («интерфейс») молекула ДНК деформируется: искажается как «внутренняя геометрия» нуклеотидов, так и форма двойной спирали ДНК (геликоид). Степень деформации зависит от нуклеотидной последовательности ДНК в области интерфейса и уникальна для каждого из комплексов гомеодомен-ДНК. Совокупность всех этих факторов обеспечивает надежность, гибкость и точность узнавания. Мы представили лишь схему образования интерфейсов комплексов гомеодомен-ДНК. Более конкретное и подробное изложение проблемы узнавания ДНК белками (не только факторами транскрипции) приведено в обзоре [1], в котором дан большой список цитированных работ и приведены соответствующие доводы и доказательства.

Транскрипция – это сложная (меру сложности мы пока не знаем) сеть биохимических и физико-химических взаимодействий. Принципы организации и работы этой сети неизвестны. Эта область исследований в настоящее время бурно развивается. Цель нашей работы более проста. Мы исследуем самые элементарные свойства регуляторов этой сети, а именно, строение семейства комплексов гомеодомен – ДНК. При взаимодействии с белком молекула ДНК в интерфейсе деформируется, пространственная структура гомеодомена искажается незначительно. Это дает основание для того, чтобы свести задачу изучения строения семейства комплексов к задаче строения интерфейсов комплексов этого семейства. Поэтому можно полагать, что характеристическим свойством комплекса является уникальная (специфическая) деформация структуры и формы ДНК в интерфейсе. Наша работа основана на этом предположении.

2. Материалы и методы

Для изучения строения семейства комплексов мы используем методы многомерного статистического анализа, что позволит получить сведения о взаимосвязи комплексов друг с другом, оценить меру этой связи и найти основы для формулирования правил узнавания и построения моделей узнавания белком ДНК.

Трудности, которые возникают при постановке и решении такого рода задач, связаны с огромной многомерностью комплексов, числом комплексов в семействе, выбором дескрипторов, необходимостью обработки большого объема данных, выбором адекватных подходов и методов. Мы используем в качестве дескрипторов координаты ДНК в комплексах гомердомен-ДНК из Protein Data Bank (PDB) – банка данных трехмерных структур белков и нуклеиновых кислот, и геометрические параметры ДНК, вычисленные с использованием пакета программ 3DNA, предназначенного для анализа, реконструкции и визуализации трехмерных структур нуклеиновых кислот [2, 3].

В работе мы используем следующие группы дескрипторов - геометрических параметров, характеризующих структуру ДНК в интерфейсах комплексов:

1. Параметры, описывающие пространственное расположение соседних пар оснований, т. е. положение и ориентацию одной пары относительно другой: сдвиги *Shift*, *Slide* и *Rise* по осям *Ox*, *Oy*, *Oz* соответственно и три поворота *Tilt*, *Roll* и *Twist* (параметры твердого тела).
2. Параметры *Shift*, *Slide* и *Ris*e по осям *Ox*, *Oy*, *Oz*, соответственно, и три поворота *Tilt*, *Roll* и *Twist*, аналогичные первой группе, они характеризуют положение соседних азотистых оснований в цепочке ДНК.
3. Параметры, описывающие регулярность спирали (спиральные параметры): смещения по осям *x, y* и *z* – *dx, dy*, *helical rise* (*h*) соответственно, и повороты вокруг осей – *inclination* (η), *tip* (θ) и *helical twist* (Ω*h*).
4. Параметры, описывающие отклонения положения оснований в паре от идеального их положения – *Shear, Stretch, Stagger, Buckle, Propeller, Opening* («диккерсоновы параметры»), где *Shear* и *Stretch* определяют смещения оснований в плоскости пары, *Opening* – угол между основаниями в плоскости пары. Три оставшихся параметра характеризуют смещение оснований из плоскости пары. «Идеальным» здесь является положение оснований в абсолютно плоской паре, где все шесть параметров равны.

Ранее нами были проанализированы параметры, описывающие пространственное расположение соседних пар оснований («диккерсоновы параметры»): сдвиги *Shift*, *Slide* и *Rise* по осям *Ox*, *Oy*, *Oz* соответственно и три поворота *Tilt*, *Roll* и *Twist* [4].

Данными для этой работы являлись спиральные параметры пространственных структур ДНК в интерфейсах 75 комплексов гомеодомен-ДНК (PDB). Для сравнения и классификации спиральных параметров мы применяем методы многомерного статистического анализа: методы ANOVA, непараметрического дисперсионного анализа (тест Крускала – Уоллиса), методы кластерного и дискриминантного анализов.

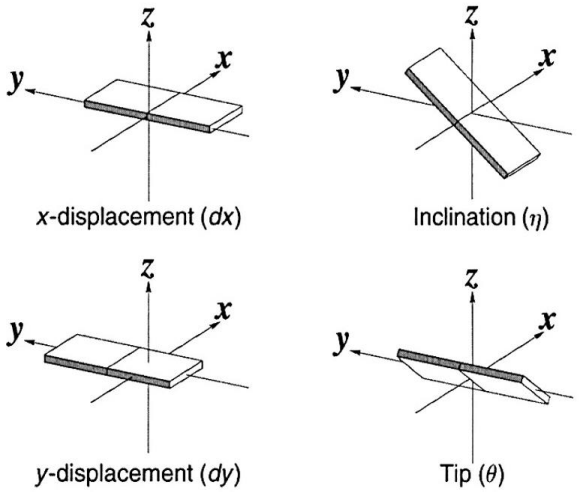

Рис. 1. Спиральные параметры пространственных структур ДНК в интерфейсах комплексов гомеодомен-ДНК.

3. Результаты и выводы

Для проверки гипотезы независимости параметров, H_0, против альтернативы, что параметры зависимы, использовали критерий Спирмена. Результаты анализа (см. рис. 2) показали, что гипотеза о независимости от нуклеотидной последовательности фрагмента ДНК в интерфейсе гомеодомен-ДНК, *Gr*, только для переменной *Incl* отвергается. Для выявления влияния качественного фактора, *Gr*, на параметры, описывающие регулярность спирали, спиральные параметры, смещения по осям *x*, *y* и *z*, *x*-displacement, *y*-displacement, helical rise соответственно, и повороты вокруг осей спирали – inclination (η), tip (θ) и helical twist (Ω*h*), мы использовали однофакторный дисперсионный анализ, ANOVA, непараметрические критерии, тест Крускала – Уоллиса и медианный критерий Брауна – Муда [5]. Результаты анализа позволяют сделать вывод, что влияние нуклеотидной последовательности на каждый из этих параметров статистически незначимо на уровне $\alpha = 0.05$. Отметим, что результаты, полученные с помощью однофакторного дисперсионного анализа, подтверждаются и непараметрическими методами (рис. 3).

	Spearman Rank Order Correlations (Data_II_N_1.sta) Marked correlations are significant at p <,05000						
Variable	X-disp	Y-disp	h-Rise	Incl	Tip	h-Twist	Gr
X-disp	1,00	-0,01	0,32	-0,53	0,03	0,52	-0,02
Y-disp	-0,01	1,00	0,01	-0,05	-0,14	-0,05	-0,01
h-Rise	0,32	0,01	1,00	-0,22	0,02	0,40	-0,02
Incl	-0,53	-0,05	-0,22	1,00	0,04	-0,21	-0,06
Tip	0,03	-0,14	0,02	0,04	1,00	0,02	-0,04
h-Twist	0,52	-0,05	0,40	-0,21	0,02	1,00	-0,02
Gr	-0,02	-0,01	-0,02	-0,06	-0,04	-0,02	1,00

Рис. 2. Корреляционная матрица Спирмена.

На следующем этапе мы использовали методы классификации многомерных данных: кластерный и дискриминантный анализы. Для классификации результатов наблюдений был использован агломеративно-иерархический кластерный анализ с евклидовым расстоянием, а в качестве меры различия – метод Варда [5]. Кластерный анализ, разбил исходную выборку на три однородные группы (рис. 4). Таким образом, мы построили группирующую переменную, Gr_3 (рис. 4). Дискриминантный анализ позволяет определить величины корректных вероятностей классификации. На рисунке 5 приведены результаты дискриминантного анализа, где показано, что вероятность корректной классификации равна ~ 80 %. В дискриминантном анализе используется понятие канонических переменных, которые несут максимальную информацию о взаимном расположении групп. Поэтому они полезны для геометрического представления результатов дискриминантного анализа. На рисунке 6 приведено графическое представление, т. е. проекция результатов наблюдений на плоскость первых двух канонических переменных. Для оценки влияния переменной Gr_3 на исходные параметры были использованы непараметрические критерии Крускала – Уоллиса и Брауна – Муда. Необходимо отметить, что влияние этой группирующей переменной на каждый из параметров, кроме tip (θ), статистически значимо на уровне $\alpha = 0.05$.

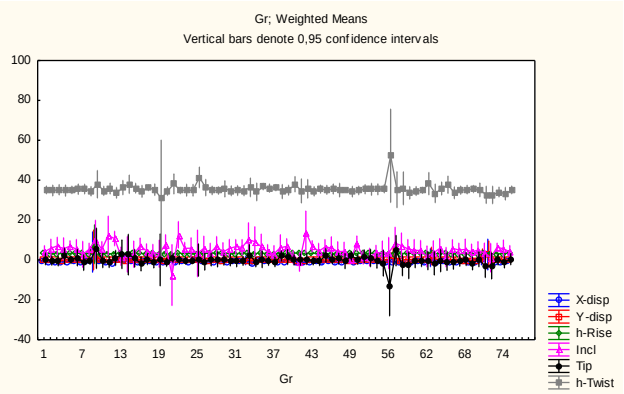

Рис. 3. Взвешенные средние значения параметров.

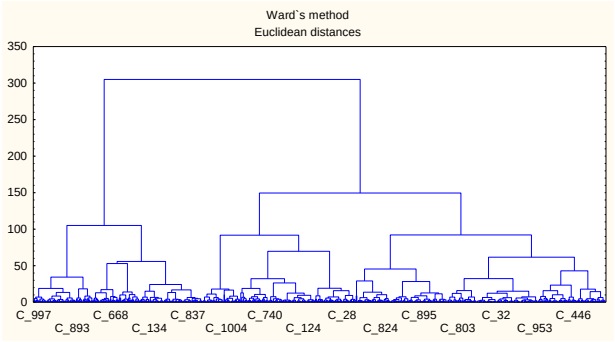

Рис. 4. Диаграмма – результат кластерного анализа.

Classification Matrix
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	G_1:1 p=,29484	G_2:2 p=,26859	G_3:3 p=,43657
G_1:1	77,74	262	2	73
G_2:2	75,57	3	232	72
G_3:3	87,17	30	34	435
Total	81,28	295	268	580

Рис. 5. Классификационная матрица.

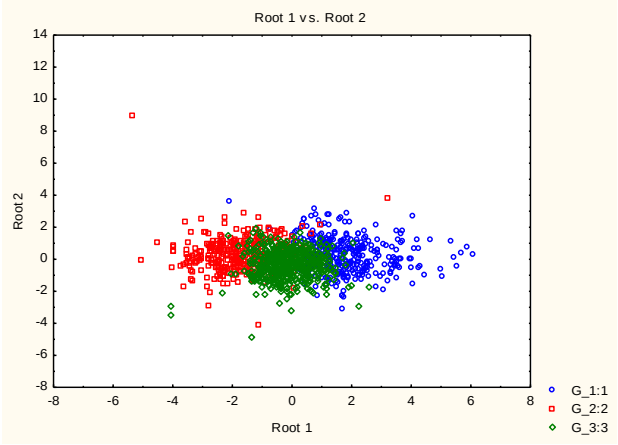

Рис. 6. Проекция результатов наблюдений на плоскость первых двух канонических переменных.

Заключение

В работе [4] и данной работе представлены результаты многомерного статистического анализа только для двух типов параметров ДНК – «диккерсоновых параметров» [4] и спиральных параметров ДНК. Необходимо решить аналогичные задачи для других групп геометрических характеристик ДНК. Затем, по совокупности решений всех задач, можно будет предложить адекватную интерпретацию полученных результатов. Это, возможно, позволит определить филогенетические связи комплексов (по родству) или по типологии (по сходству).

Список литературы

1. Siggers T., Gordan R. Protein–DNA binding: complexities and multi-protein codes. *Nucleic Acids Research*. 2014. V. 42. № 4. P. 2099–2111. doi: 10.1093/nar/gkt1112.
2. Lu X.-J., Olson W.K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* 2003. V. 31. P. 5108–5121. doi: 10.1093/nar/gkg680.
3. Lu X.-J., Olson W.K. Characterization of base pair geometry. *Computational Crystallography Newsletter*. 2016. V. 7. P. 6–9.
4. Грохлина Т.И., Панченко Л.А., Полозов Р.В., Сивожелезов В.С., Иванов В.В. Классификация комплексов семейств белков: гомеодомены – ДНК, цинковые пальцы – ДНК. Статистический анализ структур ДНК в интерфейсах комплексов гомеодомен–ДНК. В: *Доклады Международной конференции «Математическая биология и биоинформатика»*. Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e65. doi: 10.17537/icmbb18.98.
5. Zar J.H. *Biostatistical Analysis*. N.J.: Prentice-Hall, 1999. 663 p.