БИОФИЗИКА, 2018, том 63, вып. 2, с. 225–231

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.3

НЕКАНОНИЧЕСКИЕ И СТРОГО ЗАПРЕЩЕННЫЕ КОНФОРМАЦИИ ОСТОВА ПОЛИПЕПТИДНЫХ ЦЕПЕЙ ГЛОБУЛЯРНЫХ БЕЛКОВ

© 2018 г. И.Ю. Торшин, А.В. Батяновский*, Л.А. Урошлев, Н.Г. Есипова**, В.Г. Туманян****

Химический факультет Московского государственного университета имени М. В. Ломоносова, 119991, Москва, Ленинские горы, 1/3

**Институт биофизики и клеточной инженерии НАН Беларуси 220072, Минск, ул. Академическая, 27, Беларусь*

***Институт молекулярной биологии им. В.А. Энгельгардта РАН, 119991, Москва, ул. Вавилова, 32*

E-mail: tuman@eimb.ru

Поступила в редакцию 02.12.17 г.

С использованием современных методов распознавания и классификации проведено построение, предложено новое представление и выполнен исчерпывающий анализ аналога карты Рамачандрана. Сверхбольшие карты, включающие более 50 млн точек, опираются на последние релизы банков данных структур глобулярных белков. Установлены области карты A, B, B′, C и D, соответствующие строго запрещенным конформациям, их суммарная площадь составляет 25% площади всей карты. Область неканонических конформаций определяется путем вычитания строго запрещенных и разрешенных областей изо всей площади карты. Получены аргументы в пользу новой классификации конформаций остова полипептидной цепи белка.

Ключевые слова: глобулярные белки, карта Рамачандрана, конформационный анализ, распознавание и классификация, статистический и кластерный анализ, строго запрещенные, разрешенные, неканонические конформации.

Предложенный Рамачандраном и соавторами [1,2] способ описания и анализа конформаций основной цепи полипептида в виде двумерных карт основан на том обстоятельстве, что число главных конформационных параметров на остаток, определяющих структуру основной цепи, равно всего двум (в полинуклеотидной цепи число таких параметров на единицу равно пяти). Это двугранный угол φ между плоскостями, проходящими через атомы C′, N, C^{α} и атомы N, C^{α}, C′, и двугранный угол ψ между плоскостями, проходящими через атомы N, C^{α}, C′ и атомы C^{α}, C′, N. Угол ψ между плоскостями, определяемыми атомами C^{α}, C′, N и C′, N, C^{α}, в первом приближении можно считать фиксированным, что связано с фундаментальным свойством пептидной связи – sp^2 гибридизацией атомов азота аминной группы и углерода карбоксильной группы, приводящей к практической планарности пептидной связи. Выигрыш в числе независимых переменных при описании конформации в пространстве естественных параметров по сравнению с декартовыми координатами атомов в случае полипептидной цепи особенно велик, ведь так называемая дипептидная единица, включающая не менее тринадцати атомов, определяется всего двумя параметрами. Поэтому, поскольку имеет место взаимно-однозначное соответствие конформации главной цепи белка и пар значений двугранных углов φ и ψ, представление конформационной картины как функции двух независимых переменных стало золотым стандартом в конформационном анализе белков.

Конформационный анализ имеет двоякую цель: с одной стороны – найти разрешенные конформации молекулы или ее сегмента, с другой – запрещенные конформации. Считается, что первым соответствует более благоприятная энергия, тогда как вторым – менее благоприятная энергия. На практике, кроме всего, остаются конформации промежуточного типа, принадлежащие к так называемой серой зоне, вопрос трактовки которой может решаться поразному.

Уже первые варианты конформационных карт, содержащие контуры разрешенных областей и точки экспериментально определенных пар значений (φ, ψ), позволили Рамачандрану и соавторам [1,2] охарактеризовать как разре-

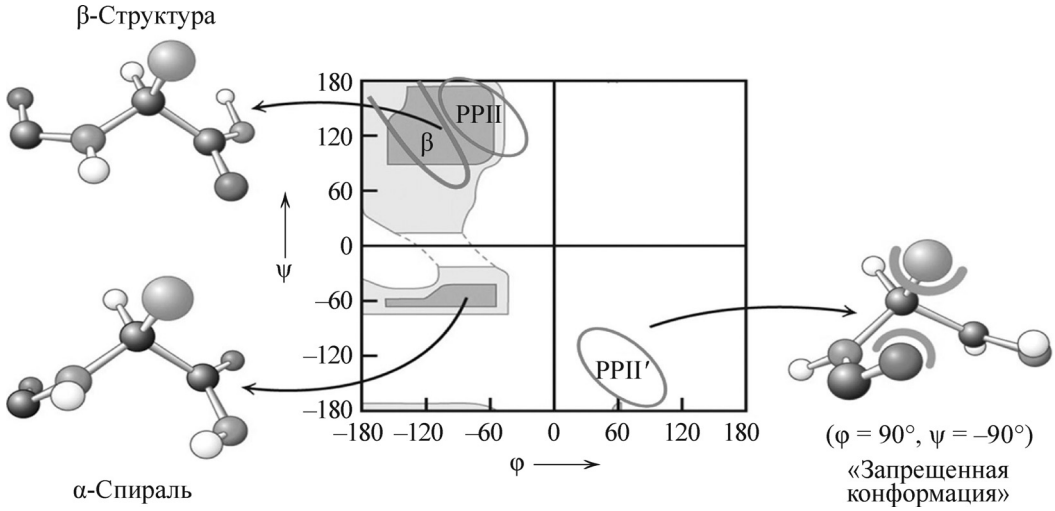

Рис. 1. Конформации остова полипептидной цепи. Схематическое представление карты Рамачандрана. Приведены области на карте и конформации, соответствующие α-спиралям, β-структурам и PPII (левоспиральная конформация типа полипролин II [3,4]). Показан также пример «запрещенной» или неразрешенной конформациии (комментарии в тексте).

шенные, так и запрещенные области. Анализ полученных в этих работах карт позволяет прийти к принципиальному выводу о том, что площадь разрешенных областей на карте для дипептидной единицы примерно в четыре раза меньше, чем площадь запрещенных областей. В ходе развития расчетных и экспериментальных исследований соотношение разрешенных и запрещенных областей и контуры этих областей, естественным образом, продолжают уточняться.

Что касается разрешенных областей, то путем анализа экспериментально определенных структур белков в согласии с расчетами были обнаружены два ярко выраженных разрешенных кластера, соответствующих α-спиралям и β-структурам, а также небольшой кластер, отвечающий α-спиралям левого знака (рис. 1). В ходе дальнейших исследований было обнаружено, что в левом верхнем квадранте находится не одно, а два сгущения точек: кроме области, соответствующей β-структурам, имеет место область PPII (левоспиральная конформация типа полипролин II) [3,4].

На рис. 1, кроме разрешенных областей, показан пример «запрещенной» или неразрешенной конформациии. Эта конформация соответствует β-изгибу типа II′, или, согласно номенклатуре, предложенной в [5,6], конформации PPII′, симметричной PPII.

Вопрос о запрещенных областях карты Рамачандрана требует более подробного изложения. Уже результаты первых рентгеноструктурных экспериментов, в ходе которых было расшифровано относительно немного (десятки или сотни) структур белков, позволили обнаружить, наряду с разрешенными, области карты Рамачандрана, которые остаются практически незаполненными в согласии со стерическими запретами для полипептидной цепи в соответствующей конформации. Появление отдельных экспериментальных точек в этих стерически неразрешенных областях было естественно трактовать либо как ошибки эксперимента, либо как феномен, требующий специального изучения. По мере роста числа белковых структур, депонированных в банки данных, и повышения качества (разрешения) этих структур осуществлялась элиминация ошибок эксперимента и отбор достоверных, хотя и попадающих в неразрешенные области конформаций.

Такого рода конформации, приобретшие статус факта, получили наименование «конформационно-неразрешенных конформаций» («conformationally disallowed»), поскольку им соответствуют стерически менее выгодные конформации по сравнению, например, с областями, соответствующими α-спиралям и β-структурам (см. рис. 1). Было предложено несколько отличающихся друг от друга вариантов определения «запрещенных» областей [7,8]. Авторы работы [9] избегают термина «неразрешенная конформация» и говорят о «редких конформациях» (аглийский термин «sparsely» означает «редко, негусто, скудно»); в их число попадают конформации, которые в более ранних работах назывались «запрещенными». Таким образом,

наметилась тенденция перехода от термина, подразумевающего оценку энергии конформации, к чисто феноменологическому термину, отражающему частоту реализации соответствующих конформаций. Надо сказать, что использование терминологии, основанной на статистической оценке, представляется более корректным. Действительно, энергию локальной структуры трудно оценить, поскольку роль играет как контекст, так и трудно контролируемые факторы, например, квантово-химической природы.

Забегая вперед, скажем, что мы поставили себе целью получить аргументы в пользу разделения карты Рамачандрана на области с благоприятной энергией, строго запрещенные области и все остальные, которые разумно назвать неканоническими областями. Исследование базируется на анализе последних версий банка PDB и новых методов кластеризации точек, в данном случае в пространстве конформаций.

Один из поводов возвращаться к проблеме карты Рамачандрана – все ускоряющийся рост числа вновь депонированных структур в банк PDB и повышение качества этих структур. Последняя достаточно полная ревизия карты Рамачандрана была выполнена в работах [5,6], причем в вышедшей в 2010 г. работе [6] был проведен анализ выборки структур с высоким разрешением (≤ 1,2 Å), включившей данные по 72000 аминокислотным остаткам. В настоящей работе с использованием банка данных PDB (релиз 2016 г.) проводится анализ более 50 миллионов аминокислотных остатков. Кроме того, развиваются новые методы обработки больших данных. Итак, в настоящей работе мы на основе всего накопленного к настоящему времени экспериментального материала и с использованием новых методических приемов обработки данных проводим построение, анализ и интерпретацию карты Рамачандрана.

На какие же вопросы в принципе может быть получен ответ путем конформационного анализа с применением карты Рамачандрана? Как и в общем случае, конформационный анализ дает возможность помочь найти предпочтительные конформации и объяснить причины их стабилизации. С другой стороны, с помощью конформационного анализа можно очертить круг конформаций, запрещенных в той или иной степени, с выяснением природы такого рода запретов.

Для анализа достаточно большого массива данных по значениям углов (φ, ψ) представляет интерес использовать новейшие подходы к кластеризации, которые, во-первых, позволяют обрабатывать большие массивы данных (в настоящем исследовании каждой точке отвечает пара углов φ, ψ и таких точек десятки миллионов) за приемлемое время, и, во-вторых, характеризуются более высокой чувствительностью в смысле возможности детектирования «сгущений» (кластеров) точек. В настоящей работе для кластерного анализа использован подход, основанный на фундаментальной концепции *метрики* (в математике «метрика» – это положительно определенная симметричная функция для измерения расстояния между парами точек, которая удовлетворяет неравенству треугольника). На основе измерения попарных расстояний между этими точками становится возможным нахождение *метрических сгущений точек* (кластеров близко лежащих точек с высокой плотностью точек) [10]. Соответственно, могут быть найдены и «разрежения» точек (множества точек с низкими значениями плотности), соответствующие «запрещенным» областям.

МАТЕРИАЛЫ И МЕТОДЫ

Выборка данных. Был проведен анализ базы данных PDB (релиз 2016 г.). Из базы данных были удалены файлы с идентичными последовательностями и с разрешением хуже, чем 2,0 Å. В результате была сформирована выборка из структур 121450 белковых цепей, взятых из 62096 PDB-файлов. Значения (φ, ψ) были рассчитаны для 52563104 аминокислотных остатков с известными координатами каждого неводородного атома главной цепи.

Методы кластеризации для поиска «сгущений» и «разрежений» точек. Строгое математическое обоснование и описание алгоритмов поиска *метрических сгущений точек* (кластеров близко лежащих точек с высокой плотностью точек) на основе набора точек с заданной метрикой (так называемых *метрических конфигураций)* представлено в работе [10]. В соответствии с проведенными нами экспериментами на модельных кластерах с разной степенью «размытости», использованная процедура кластеризации позволяет идентифицировать кластеры даже при минимальных колебаниях в значениях плотности точек. Идентифицируемые процедурой кластеры не могут быть найдены с использованием таких стандартных подходов, как DBSCAN, OPTICS, DeLi-Clu, EM-clustering.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Отображение пар значений (φ, ψ) для 52563104 аминокислотных остатков на карту Рамачандрана позволило обнаружить сущест-

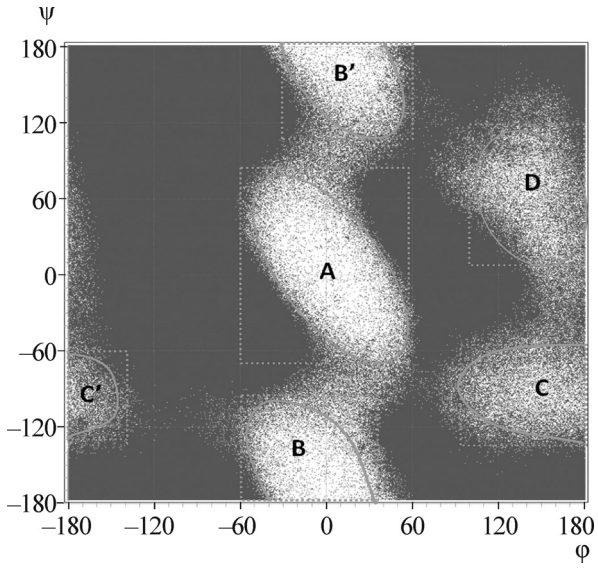

Рис. 2. «Строго запрещенные» области на карте Рамачандрана для 52563104 аминокислотных остатков, соответствующие областям с максимальной разреженностью точек. Очевидно существование четырех выраженных областей «разрежений» точек. В силу того, что границы «разрежений» размыты, овалы (сплошные линии) ограничивают области разрежений, включающие более 90% точек в каждом «разрежении». Прямые (пунктирные линии) аппроксимируют области «разрежений» прямоугольниками в терминах значений углов φ и ψ.

вование четырех выраженных областей «разрежения» точек, которые можно считать соответствующими конформациям, которые избегаются. Это область A, область (B, B′), область (C, C′) и область D (рис. 2). Мы предлагаем называть эти области «строго запрещенными» областями потому, что даже в выборке из более чем 50 миллионов точек существует крайне мало примеров аминокислотных остатков с конформациями, соответствующими данным областям. Интервалы значений углов φ и ψ, характеризующих эти области, очевидны из рис. 2. Существенно, что результат носит максимально общий характер: не проводилось никакой селекции в плане аминокислотного состава, типа локальных структур, стереохимии полипептидной цепи и тому подобное.

Проведем сравнение расположения полученных «строго запрещенных» областей на карте Рамачандрана (см. рис. 2) с известными литературными источниками. Мы будем использовать данные по атом-атомным контактам атомов главной цепи, проявляющимся на карте Рамачандрана, из работы [11], которые считаются общепринятыми, а также данные из более поздней работы [12].

Напомним обычно применяемую нумерацию атомов в так называемой дипептидной единице Рамачандрана. Атомы аминокислотного остатка в центре единицы обозначаются без индексов; атомы примыкающей слева карбонильной группы, принадлежащей предыдущему остатку, имеют индекс i–1; атомы аминной группы следующего остатка – индекс i+1.

Строго запрещенной области A соответствует контакт $O_{i-1}\ldots H_{i+1}$ согласно работе [11]; авторы работы [12] предлагают заменить этот контакт контактом $O_{i-1}\ldots N_{i+1}$ (области контактов ограничиваются эллипсом). Свой вклад в область A вносят контакты $C_{i-1}\ldots C$ (вертикальная полоса на карте) и $N\ldots H_{i+1}$ (горизонтальная полоса на карте).

Что касается областей B и B′, то здесь характерен контакт $O_{i-1}\ldots O$ (граница контакта эллипс) и тот же, что и в случае области A, контакт $C_{i-1}\ldots C$ (вертикальная полоса на карте).

Интерпретацию области C в терминах контактов можно связать с горизонтальной полосой, соответствующей контактам $C^{\beta}\ldots H_{i+1}$ (в работе [12] это контакт $C^{\beta}\ldots N_{i+1}$), и вертикальной полосой контактов $O_{i-1}\ldots C^{\beta}$.

Наконец, существование области D можно объяснить одновременным наложением контактов $O_{i-1}\ldots C^{\beta}$, $N\ldots H_{i+1}$ (а также $C^{\beta}\ldots O$ согласно работе [12]) и перекрыванием ван-дер-ваальсовых радиусов $H\ldots H_{i+1}$.

Можно видеть, что интерпретация областей D и C в терминах контактов не столь убедительна в смысле сходства формы областей, чем для областей A и B, но и запреты для первых двух областей выражены несколько слабее.

Надо подчеркнуть, что полученные нами результаты не могут быть воспроизведены, исходя из картины напряженных ван-дер-ваальсовых контактов: эти контакты имеют место и в строго запрещенных областях и в неканонических областях. В такую упрощенную концепцию не укладывается ни диагональная форма запрещенных областей, ни значительные промежутки между ними. Последние, когда речь идет о промежутках между областями A и B, хорошо согласуются с расположенными как раз между ними диагональными полосами, соответствующими реализации стабилизирующего $n \rightarrow \pi^*$ взаимодействия, чисто квантово-механического эффекта (см. рис. 3 из работы [13]).

После выявления областей достоверных разрежений мы можем перейти к областям очевидных сгущений. Наш анализ «сгущений» точек на карте Рамачандрана (рис. 3) с использованием высокочувствительного метода поиска метрических сгущений позволяет установить

три области сгущения точек: одну, соответствующую β-структурам и левым спиралям (см. область PPII на рис. 1), и вторую и третью, соответствующие правозакрученным и левозакрученным α-спиралям соответственно. Других сгущений не было установлено; все остальные области карты равномерно заполняются точками с невысокими значениями плотности (η_j < 0,03), без каких-либо выделенных пиков плотности.

Таким образом, если считать «строго запрещенными» области карты Рамачандрана с наименьшей, близкой к нулю плотностью точек (найденные выше «разрежения»), то в настоящем исследовании таковыми являются области A, (B, B′), (C, C′) и D на рис. 2. Суммарная площадь этих областей достаточно велика и составляет около 25% общей площади карты. Данное определение «строго запрещенных» областей получено на основании современных выборок структур белков с разрешением не хуже 2 Å. Если к этим областям добавить области разрешенных конформаций, показанные на рис. 3 (области очевидных сгущений), то оставшаяся после вычитания этих двух типов областей площадь карты Рамачандрана (см. рис. 4) должна соответствовать так называемым «неразрешенным» («disallowed») конформациям [7,8,15,16,17], или, пользуясь другой терминологией, «редким» («sparsely») конформациям [9]. В этой работе мы предпочитаем использовать термин «неканонические конформации».

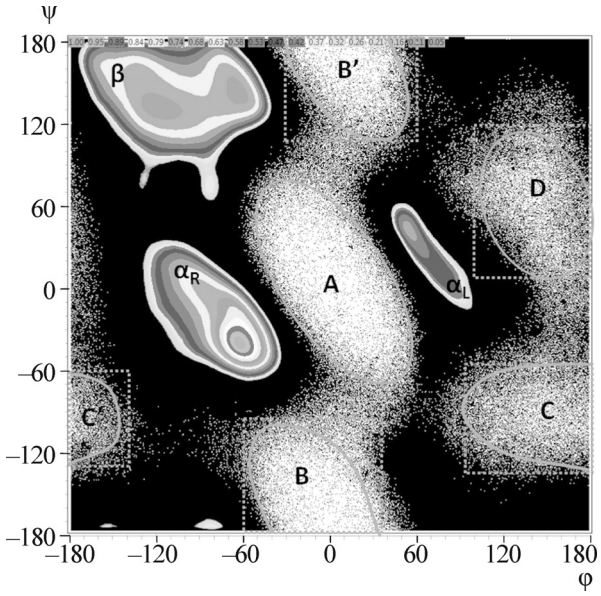

Рис. 3. «Сгущения» точек на карте Рамачандрана. Цветовая шкала (приведена в верхней части рисунка) отображает значения обобщенной плотности. Сгущение «β» соответствует β-структуре и области PPII, «α_R» – правозакрученным α-спиралям, «α_L» – левозакрученным альфа-спиралям. Пики плотности в области φ = –180°…–60°, ψ = –180°…–170° можно рассматривать как часть сгущения «β». На карте также показаны области разряжений точек; белый цвет соответствует нулевой плотности точек в соответствующей области диаграммы. Строго говоря, точные контуры областей сгущений зависят от типа аминокислотного остатка и от модуляции угла ω, задающего отклонение от плоскости пептидной группы, как это показано в работе [14].

На рис. 4 помещены области неразрешенных конформаций согласно работам [7,8]. Можно видеть, что они хорошо согласуются с участками конформационной карты, которые не являются по нашей терминологии ни строго запрещенными, ни разрешенными. И это утверждение остается верным при том, что результаты работ [7] и [8] не во всем согласуются друг с другом.

Итак, полученные данные по локализации «условно запрещенных» областей, предложенных в более ранних работах, и локализации «строго запрещенных» и разрешенных областей указывают на необходимость введения тройственной классификации областей карты Рамачандрана: «разрешенные области» (соответствующие пикам плотности на рис. 3), «строго запрещенные области» (области разрежений на рис. 2) и неканонические области. С последними проявляют определенное сходство «условно запрещенные» области из работ [7,8], т.е. не строго запрещенные и в то же время не разрешенные области. К ним примыкают так называемые «редкие» области из работы [9].

Условно запрещенными следует считать все области карты Рамачандрана, которые не являются «разрешенными» или «строго запрещенными». Дело в том, что использованная в настоящей работе процедура позволяет идентифицировать сгущения точек, если колебания плотности точек от сгущения к разрежению превышают 5–7%. В рамках такой точности не представляется возможным достоверно различать сгущения точек в областях «I», «II», «II′».

Мы рассмотрели результаты анализа конформаций главной цепи вне зависимости от типа аминокислотных остатков. Данные по картинам конформаций основной цепи для индивидуальных аминокислот приведены в работе [22].

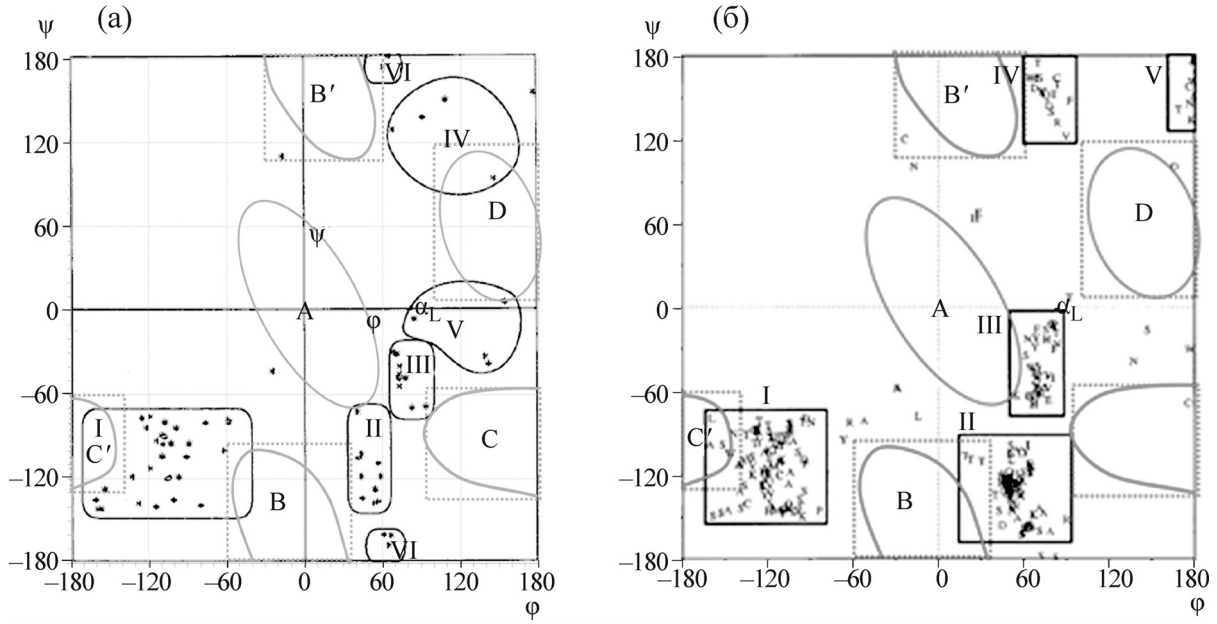

Рис. 4. Сравнение расположения «разрежений» с определениями «условно запрещенных» (disallowed) областей, полученными в работах [7] (а) и [8] (б).

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проекты №№ 17-04-02105, 16-54-00219-Бел) и Белорусского республиканского фонда фундаментальных исследований (проект № Б16Р-178).

СПИСОК ЛИТЕРАТУРЫ

1. G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, J. Mol. Biol. **7**, 95 (1963).
2. C. Ramakrishnan and G. N. Ramachandran, Biophys. J. **55**, 909 (1965).
3. A. A. Adzhubei, M. J. Sternberg, and A. A. Makarov, J. Mol. Biol. **425**, 2100 (2013).
4. N. G. Esipova and V. G. Tumanyan, Curr. Opin. Struct. Biol. **42**, 41 (2017).
5. S. A. Hollingsworth, D. S. Berkholz, and P. A. Karplus, Prot. Sci. **18**, 1321 (2009).
6. S. A. Hollingsworth and P. A. Karplus, Biomol. Concepts **1**, 271 (2010).
7. K. Gunasekaran, C. Ramakrishnan, and P. Balaram, J. Mol. Biol. **264**, 191 (1996).
8. D. Pal and P. Chakrabati, Biopolymers **63**, 195 (2002).
9. N. V. Kalmankar, C. Ramakrishnan, and P. Balaram, Proteins **82**, 1101 (2014).
10. I. Yu. Torshin and K. V. Rudakov, Pattern Recogn. Image Anal., No. **4**. 145 (2016).
11. N. Mandel, G. Mandel, B. L. Trus, et al., J. Biol. Chem. **252**, 4619 (1977).
12. B. K. Ho, A. Thomas, and R. Brasseur, Prot. Sci. **12**, 2508 (2003).
13. M. P. Hinderaker and R. T. Raines, Prot. Sci. **12**, 1188 (2003).
14. И. Ю. Торшин, А. В. Батяновский, Л. А. Урошлев и др., Биофизика **62** (3), 435 (2017).
15. M. C. Vega, J. C. Martinez, and L. Serrano, Prot. Sci. **9**, 2322 (2000).
16. Л. А. Урошлев, И. Ю. Торшин, А. В. Батяновский и др., Биофизика **60** (1), 5 (2015).
17. I. Yu. Torshin, N. G. Esipova, and V. G. Tumanyan, J. Biomol. Struct. Dynam. **32** (2), 198 (2014).

Non-canonical and Strongly Disallowed Conformations of the Backbone of Globular Proteins Polypeptide Chain

I.Yu. Torshin*, A.V. Batyanovskii, L.A. Uroshlev***, N.G. Esipova***, and V.G. Tumanyan*****

**Department of Chemistry, Lomonosov Moscow State University, Leninskie Gory 1/3, Moscow, 119991 Russia*

***Institute of Biophysics and Cell Engineering, National Academy of Sciences of Belarus, ul. Akademicheskaya 27, Minsk, 220072 Belarus*

****Engelhard Institute of Molecular Biology, Russian Academy of Sciences, ul. Vavilova 32, Moscow, 119991 Russia*

Using modern recognition and classification methods, the analog of the Ramachandran map was drawn, a fresh look at the map was taken and a thorough analysis was performed. Very large maps with the density of more than 50 million dots are created based on data sets derived from the latest releases of globular protein structure data banks. A, B, B', C, and D regions corresponding to strongly disallowed conformations are defined, they occupy 25% of the plot area. The region of non-canonical conformations can be determined by subtracting strongly disallowed and permitted regions from the overall plot area. The arguments in favor of new classification of conformations of the backbone of protein polypeptide chain are found.

Keywords: globular proteins, Ramachandran plot, conformational analysis, recognition and classification, statistical and cluster analysis; strongly disallowed, permitted, non-canonical conformations