УДК 519.6

И. В. Басов[1], студент, e-mail: generalrot@rambler.ru,
С. В. Грошев[2], ст. преподаватель, e-mail: groshev_sergey@mail.ru,
А. П. Карпенко[3], д-р физ.-мат. наук, проф., зав. кафедрой, e-mail: apkarpenko@mail.ru,
К. В. Шайтан[3], д-р физ.-мат. наук, проф., e-mail: shaytan49@yandex.ru,
Д. Л. Шуров[4], аспирант, e-mail: d.l.shurov@gmail.com,
[1] Московский институт государственного и муниципального управления,
[2]Московский государственный технический университет им. Н. Э. Баумана,
[3]МГУ имени М. В. Ломоносова, Биологический факультет, Кафедра биоинженерии,
[4]Институт химической физики им. Н. Н. Семенова Российской академии наук

Метод построения и кластерного анализа карт вероятности заселенности конформаций дипептидов

Представлены динамические аналоги известных карт Рамачандрана, именуемые картами вероятности заселенности конформаций. Предложенные карты показывают вероятности реализации значений торсионных углов основной цепи полипептидов. Карты заселенности получены на основе предварительно определенных траекторий молекулярной динамики, а затем кластеризованы с использованием самоорганизующейся карты Кохонена. Представлен метод построения и кластерного анализа карт заселенности конформаций для всех возможных 400 дипептидов.

Ключевые слова: *дипептиды, конформационная динамика полипептидов, карта Рамачандрана, карта вероятности заселенности конформаций, кластерный анализ, самоорганизующаяся карта Кохонена*

Введение

Изучение пространственной структуры и структурной динамики полипептидов имеет большое значение для понимания их функционирования и разработки методов направленной модификации биологических функций этих макромолекул. Полипептиды являются гетерополимерами, состоящими из природных аминокислотных остатков, соединенных пептидными связями. Высокомолекулярные функционально активные полипептиды (белки) формируют уникальные пространственные структуры с образованием ионных, водородных, дисульфидных связей и с участием гидрофобных взаимодействий [1—3]. Пространственная структура полипептидов испытывает тепловые флуктуации с характерными временами от долей наносекунд и выше и амплитудами от 0,05 нм и более. Эти флуктуации пространственной структуры называются конформационной динамикой полипептидов и играют важную роль в функционировании биополимеров.

Основной вклад в конформационную динамику и пространственную структуру полипептидов вносят изменения двугранных (торсионных) углов при поворотах аминокислотных остатков вокруг одинарных химических связей в полипептидной цепи. Выделяют типы торсионных углов φ, ψ, ω [3, 4]. С практической точки зрения угол ω пептидной связи —CO—NH— не представляет интереса, поскольку поворот по этому углу связан с преодолением относительно высокого энергетического барьера, вследствие чего значение угла ω обычно близко к 180° (реже к 0°) [1].

Для изучения конформационной динамики биополимеров в настоящее время широко используют метод молекулярной динамики, основанный на численном решении систем уравнений механики макромолекул. Существует открытая база данных белков PDB (Protein Data Bank) [5], в которой содержится информация о примерно 120 000 белковых структур. Каждый белок в базе данных PDB имеет уникальный четырехзначный код.

Работа посвящена исследованию способов представления информации для кластерного анализа энергетических поверхностей *дипептидов*, которые являются первыми представителями в линейке всех возможных полипептидов. Предлагаем изучать и сравнивать динамические аналоги известных *карт Рамачандрана — карты вероятности заселенности*

конформаций, определяющие вероятности реализации значений пар торсионных углов φ, ψ основной цепи полипептидов. Карты заселенности получаем на основе предварительно полученных траекторий молекулярной динамики (см., например, [6]), а затем кластеризуем с использованием самоорганизующейся карты Кохонена.

Известно 20 природных аминокислот, которые встречаются в различных сочетаниях в полипептидах. Эти аминокислоты принято обозначать трехбуквенным кодом (GLY, LYS, TYR и т. д.). Ставим задачу выполнить молекулярное моделирование и разработать методы автоматизации кластерного анализа карт заселенности конформаций для всех возможных 400 дипептидов.

1. Карты Рамачандрана и карты заселенности конформаций

Карта Рамачандрана (Сасисекхарана-Рамакришнана-Рамачандрана) представляет собой двумерную диаграмму возможных значений торсионных углов φ, ψ аминокислотных остатков, входящих в структуру пептида. Карта предложена в 1963 г. Г. Н. Рамачандраном, С. Рамакришнаном, В. Сасисекхараном [7]. Каждую точку на карте определяют координаты, которые соответствуют углам φ, ψ в отдельном аминокислотном остатке. Совокупность всех таких точек на карте позволяет качественно и количественно оценить наличие тех или иных элементов вторичных структур в белке.

Картами Рамачандрана удобно характеризовать распределение конформаций аминокислотных остатков в молекулах белков. На рис. 1 в качестве примера показана карта Рамачандрана для кристаллической структуры одного из человеческих белков, имеющего код 1HMP в базе данных PDB. Точки на карте определенным образом сгруппированы, что позволяет судить о наличии в белке устойчивых вторичных структур. Точки, расположенные в пределах области, выделенной серым цветом, соответствует энергетически выгодным конформациям — различным видам спиралей и листов [8]. Точки за пределами этой области соответствуют относительно небольшому числу энергетически напряженных конформаций.

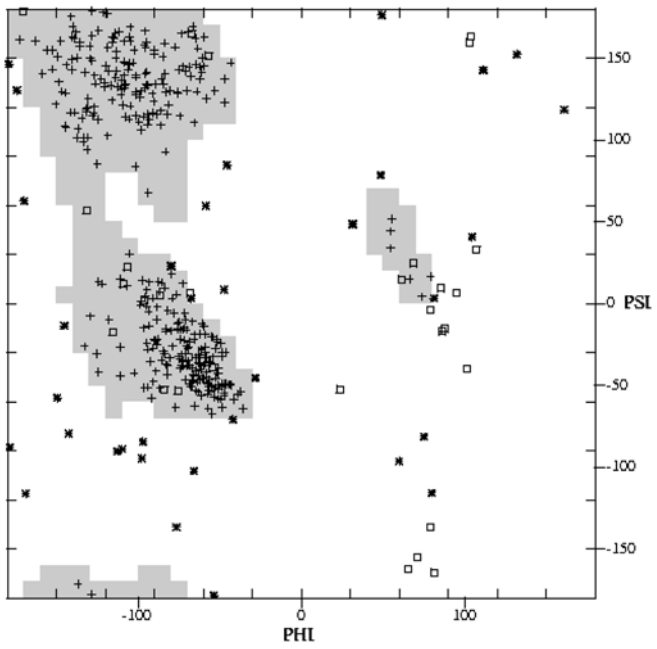

Рис. 1. Карта Рамачандрана для белка 1HMP [9]

Энергетическая (потенциальная) поверхность пептидной цепи представляет собой скалярную функцию от пространственных переменных, определяющих положение атомов этой цепи:

$$U(\mathbf{r}) = \Sigma U_{\text{связей}}(\mathbf{r}) + \Sigma U_{\text{другие}}(\mathbf{r}).$$

Здесь **r** — вектор расстояний между атомами цепи; $\Sigma U_{\text{связей}}(\mathbf{r})$ — потенциальная энергия химических связей; $\Sigma U_{\text{другие}}(\mathbf{r})$ — потенциальная энергия невалентных взаимодействий [10].

Компонентами слагаемого $\Sigma U_{\text{связей}}(\mathbf{r})$ являются следующие функции.

- Энергия химической связи, зависящая от расстояния между атомами и равная

$$U_b = k_b(r_{ij} - r_0)^2,$$

где k_b — коэффициент упругости (связь рассматривается как линейная пружина); r_{ij} — текущее расстояние между атомами; r_0 — равновесная валентная длина.

- Энергия связи, зависящая от валентного угла и определяемая выражением

$$U_a = k_a(\theta - \theta_0)^2;$$

$$U_a = \frac{k_a}{2}(\cos(\theta) - \cos(\theta_0))^2,$$

где k_a — коэффициент упругости (связь рассматривается как пружина кручения); θ — текущий валентный угол; θ_0 — равновесный валентный угол.

- Энергия связи, зависящая от торсионного угла ψ

$$U_d = \begin{cases} k_d(1 + \cos((\nu\psi + \phi)), \nu > 0, \\ k_d(\psi - \phi)^2, \nu = 0. \end{cases}$$

Здесь k_d — коэффициент упругости (связь рассматривается как пружина кручения); ν — кратность торсионного барьера; ψ — торсионный угол; ϕ — угол сдвига фаз.

Компоненты невалентных взаимодействий $\Sigma U_{\text{другие}}(\mathbf{r})$ обычно определяют потенциалы Ленард—Джонса (реже Борна—Хаггинса—Мейера, Букингема, Морса), кулоновские потенциалы [9].

Далее потенциальную энергию полипептида рассматриваем как функцию его торсионных углов в виде

$$U = U(\varphi_1, ..., \varphi_n, \psi_1, ..., \psi_n),$$

где n — число аминокислотных остатков в пептидной цепи.

На рис. 2 (см. вторую сторону обложки) изображена двумерная проекция энергетической поверхности одного из белков. Как правило, эта поверхность имеет один глобальный минимум, отвечающий наиболее энергетически выгодной конформации, а также некоторое число локальных минимумов. Прежде чем перейти в наиболее оптимальное по энергии состояние, белок в процессе молекулярной динамики проходит через некоторое число локально оптимальных состояний. Данный процесс называется сворачиванием (или фолдингом) белка.

Самый крупный из известных белков — титин состоит из 38 138 аминокислотных остатков. Отсюда следует, что энергетическая поверхность белка имеет очень высокую размерность, затрудняющую ее исследование.

2. Предлагаемая процедура получения карт заселенности конформаций для дипептидов

Суть процедуры получения карт заселенности конформаций заключается в том, что в качестве набора точек (см. рис. 1) мы используем набор пар углов φ, ψ для звеньев полипептидной цепи, определяемый по ее молекулярно-динамической траектории с некоторым шагом по времени. Каждую из точек карты окрашиваем в соответствии с вероятностью нахождения молекулы (дипептида) в этом состоянии (рис. 3, см. вторую сторону обложки).

Оранжевый цвет на рис. 3 показывает, в каком состоянии дипептид может находиться с наибольшей вероятностью, синий — с наименьшей вероятностью. Белые участки соответствуют конформациям, реализация которых практически невозможна. Указанные вероятности пропорциональны числу точек молекулярно-динамической траектории дипептида, попавших в соответствующие подобласти пространства конформаций. Таким образом, карта заселенности конформаций дает наглядное представление о конформационной структуре дипептида в пространствах углов φ, ψ одного или нескольких аминокислотных остатков. Благодаря тому, что вероятность нахождения белка в том или ином состоянии связана с потенциальной энергией этого состояния, такая диаграмма также дает наглядное представление о том, какие состояния белка являются наиболее или наименее энергетически выгодными.

Автоматизированная процедура построения карт заселенности конформаций дипептида включает следующие шаги:

- формирование пространственной молекулярной структуры дипептида по заданным кодам его аминокислот;
- запуск программы молекулярной динамики для сформированной пространственной структуры дипептида и сбор данных о значениях углов φ, ψ на каждом временном шаге для каждой аминокислоты, входящей в дипептид;
- оценка двумерных плотностей вероятности $f_1(\varphi, \psi)$, $f_2(\varphi, \psi)$ нахождения дипептида в заданной точке конформационного пространства для первой и второй аминокислот соответственно;
- визуализация карт заселенности конформаций для каждой аминокислоты, входящей в дипептид, на основе полученных на предыдущем шаге двумерных плотностей вероятности.

Для последующего решения задачи кластеризации карт заселенности необходимо учесть, что ландшафт энергетической поверхности дипептида характеризуется одновременно двумя его аминокислотными остатками и, значит, двумя картами заселенности. Из этого следует, что кластеризацию необходимо осуществлять по объединенному набору дискретных функций плотности вероятности $\tilde{f}_1(\varphi, \psi)$, $\tilde{f}_2(\varphi, \psi)$, который представляем в виде вектора $X(j)$, где j — номер соответствующей ячейки сетки, покрывающей область допустимых значений углов φ, ψ.

3. Постановка задачи кластеризации и используемые алгоритмы кластеризации

Рассматриваем задачу кластеризации в следующей постановке. Полагаем, что объект $X \in R^m$ набора $\mathbf{X}$ имеет набор характеристик $(x_1, ..., x_m)$. Определена функция расстояния между объектами (метрика)

$$\rho(X_i, X_j) \to [0; \infty),\ X_i, X_j \in \mathbf{X}.$$

Задано число кластеров k.

Требуется найти разбиение множества объектов $\mathbf{X}$ на кластеры

$$C = \{C_1, ..., C_k\},$$

такое, что каждый объект $X_i \in \mathbf{X}$ принадлежит одному и только одному кластеру.

Задачу кластеризации ставим как задачу минимизации средней квадратичной ошибки разбиения

$$e^2(\mathbf{X}, C) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \| X_{j_i} - \mathbf{X}_j^C \|^2, \qquad (1)$$

где $\mathbf{X}_j^C$ — вектор координат центра масс кластера C_j; n_j — число объектов в этом кластере; $\sum_{j=1}^{k} n_j = n$ — число объектов в наборе $\mathbf{X}$.

Вообще говоря, в качестве метрики ρ могут быть использованы степенное расстояние, евклидово и квадрат евклидова расстояний, манхэттенское расстояние, расстояние Чебышева [12]. Мы используем евклидово расстояние.

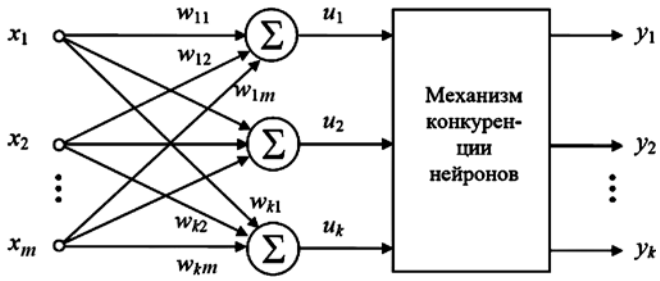

Рис. 4. Схема соединения нейронов типа *WTA* [13]

Структура нейронной сети *WTA* [13, 14] представлена на рис. 4. Все нейроны сети получают на вход один и тот же входной сигнал $x_1, ..., x_m$. На основе анализа их выходных сигналов u_i, $i = 1, 2, ..., k$, проводится выбор нейрона-победителя как нейрона, имеющего наибольшее значение u_i. Победителю ставят в соответствие сигнал $y_i = 1$, а всем другим проигравшим нейронам — нулевой сигнал $y_i = 0$.

С точки зрения реализации, нейрон Σ_i представляет собой сумматор, на который подается линейная комбинация компонентов вектора входных сигналов $\mathbf{X}$ и вектора весов его синаптических связей $\mathbf{W}_i$:

$$u_i = (\mathbf{W}_i, \mathbf{X}) = w_{i1}x_1 + ... + w_{im}x_m,\ i = 1, 2, ..., k.$$

Начальные значения синаптических весовых коэффициентов генерируем случайными, равномерно распределенными в интервале [0; 1]. Векторы входных данных нормализуем с целью избежать лишних шагов обучения и решить, хотя бы частично, проблему "мертвых" нейронов [12].

Наиболее простым методом обучения нейронной сети *WTA* является правило Гроссберга [15]. В каждом цикле обучения t в этом методе побеждает тот нейрон, чей текущий вектор входных весов $\mathbf{W}_i^t$ наиболее близок текущему входному вектору $\mathbf{X}^t$. При этом вектор $\mathbf{W}_i^t$ корректируют в сторону вектора $\mathbf{X}^t$. Поэтому в результате обучения каждая группа близких друг другу входных векторов (кластер) начинает обрабатываться одним нейроном.

В настоящей работе использован алгоритм обучения Кохонена [12—14]. В отличие от метода обучения Гроссберга этот алгоритм позволяет корректировать веса не только нейронов-победителей, но и их ближайших соседей (в смысле используемой топологии нейронной сети). Говоря более точно, веса корректируются у всех нейронов, но с ростом расстояния данного нейрона до нейрона-победителя сила коррекции слабеет.

В качестве функции соседства используем функцию Гаусса

$$G^t(i, X^t) = \exp\left(-\frac{d^2(W_i^t, X^t)}{2(\sigma^t)^2}\right),$$

где $d^2(W_i^t, X^t)$ — расстояние от i-го нейрона до нейрона-победителя в t-м цикле обучения; σ^t — ширина функции Гаусса на этом цикле обучения. Полагаем, что $d^2(W_i^t, X^t) = 0$ для нейрона-победителя; $d^2(W_i^t, X^t) = 1$ — для всех его ближайших соседей; $d^2(W_i^t, X^t) = 2$ — для соседей второго уровня и т. д.

Ширину функции Гаусса σ уменьшаем в ходе обучения по правилу

$$\sigma^t = \sigma^{\max}\left(\frac{\sigma^{\min}}{\sigma^{\max}}\right)^{\frac{t}{t^{\max}}},$$

где $\sigma^{\max}$, $\sigma^{\min}$ — соответственно максимальное и минимальное значение величины σ; $t^{\max}$ — число итераций обучения.

Коррекцию весов i-го нейрона сети *WTA* осуществляем по формуле

$$W_i^{t+1} = W_i^t + \eta_i^t G^t(W_i^t, X^t)(X^t - W_i^t),\ i \in [1{:}k],$$

где η_i^t — коэффициент обучения i-го нейрона на t-м цикле обучения. Значения этих коэффициентов в процессе обучения уменьшаем по правилу

$$\eta_i^t = \eta^{\max}\left(\frac{\eta^{\min}}{\eta^{\max}}\right)^{\frac{t}{t^{\max}}},$$

где $\eta^{\max}$, $\eta^{\min}$ — максимальное и минимальное значения этого коэффициента соответственно.

4. Программная реализация

Разработка программного комплекса для кластеризации конформационных карт дипептидов, названного DPPMClusterizer (Dipeptide Probability Population Map Clusterizer), выполнена на языке *C*, в силу его универсальности, кроссплатформенности и высокой производительности. Программирование утилит и вспомогательных инструментов, обеспечивающих конвертацию данных и сопряжение интерфейсов основных программ, осуществлено на языке программирования *Python* и языке скриптов *GNU Bourne-Again SHell (bash)*. Для компиляции использовался набор компиляторов *GCC 4.9.0 (GNU Compiler Collection)* и интерпретаторов *GNU Python 2.7.7* и *GNU bash 4.3*.

Моделирование динамики дипептидов реализовано с помощью программного пакета молекулярной динамики *NAMD* [16].

Использованы следующие сторонние библиотеки:

- *GNU SciPy 0.14.0* — универсальная научная библиотека языка *Python* для сравнительного анализа алгоритмов кластеризации;
- *GNU Matplotlib 1.3.1* — библиотека визуализации научных данных для языка *Python* и библиотеки *SciPy*.

Для графического вывода результатов работы программы использована утилита *Gnuplot 4.6*. Разработка велась в среде операционной системы *Linux* дистрибутива *Novell OpenSuse 13.1*.

Все перечисленное выше программное обеспечение является свободным и распространяется под лицензией *GPL (GNU Public License)*.

Ниже представлены основные блоки программной модели.

Сборка дипептидов. Данный блок осуществляет последовательную сборку всех вариантов дипептидов из набора 20 аминокислот. Разработаны сценарии, осуществляющие сборку исследуемого дипептида в три этапа:

1) создание конфигурационного файла для сборки;

2) сборка атомистической структуры дипептида по аминокислотной последовательности;

3) генерация структурного файла дипептида на основе файла атомистической структуры.

Запуск пакета молекулярной динамики *NAMD*. Для запуска этого пакета требуется наличие следующих данных:

- файл атомистической структуры дипептида в формате *PDB* (*Protein Data Bank*), содержащий координаты атомов [17]. Файлы PDB могут быть созданы вручную, либо загружены из открытого банка данных белковых структур;
- файл белковой структуры в формате *PSF* (*Protein Structure File*), который содержит информацию о взаимодействии ковалентно-связанных атомов;
- файл параметров силового поля, содержащий данные для вычисления потенциалов сил, действующих на атомы системы [18];
- конфигурационный файл, в котором пользователь задает параметры для запуска процесса молекулярной динамики.

Построение 2D-гистограммы вероятностей заселенности конформаций. Блок реализует следующие функции:

- анализ результатов моделирования молекулярной динамики и формирование текстового файла, содержащего соответствующие 2D-гистограммы;
- отображение гистограммы в графическом виде как карты заселенности конформаций.

Визуализация карт заселенности конформаций — создания набора изображений для обработки алгоритмами кластеризации.

Кластеризация методом *WTA*. Блок выполняет кластеризацию полученных карт заселенности по методу *WTA* с применением алгоритма обучения Кохонена.

5. Вычислительный эксперимент

Цель вычислительного эксперимента состояла в построении для всех возможных дипептидов карт конформационной заселенности в пространстве углов φ, ψ, отображающих вероятностное распределение состояний дипептида. Время моделирования методом молекулярной динамики принято равным 10 нс, шаг моделирования — 1 фс. Динамику молекул моделируем в вакууме без водного окружения при температуре, равной 300 K. Используем модель силового поля CHARMM версии 2.7 [19].

С помощью разработанного программного обеспечения выполнен анализ 400 пар φ, ψ-карт вероятности заселенности конформаций дипептидов.

Поскольку число кластеров на изображениях априори неизвестно, одной из задач исследования являлся поиск минимально достаточного числа кластеров. Если расстояние между центром кластера и ближайшим к нему дипептидом превышало заданное значение, то программный комплекс DPPMClusterizer повторял кластеризацию заданное число раз. В случае отсутствия успеха число кластеров увеличивалось. Кроме того, комплекс DPPMClusterizer для каждого кластера вычислял максимальное расстояние между его представителями (т. е. радиус кластера). Если, хотя бы для одного из текущих кластеров, это значение превышало заданное значение, то кластеризация повторялась с числом кластеров, увеличенным на единицу. Дипептиды, расстояние от которых до типичного представителя данного кластера превышало некоторый установленный пользователем порог, помечались как требующие визуальной проверки (желтый цвет). Аналогично дипептиды за пределами второго порога помечались, как, вероятнее всего, не относящиеся к данному кластеру (красный цвет). Таким образом, существенно упрощался экспертный анализ данных, полученных в результате вычислительного эксперимента.

Примеры полученных кластеров показаны на рис. 5 (см. третью сторону обложки) и рис. 6 (см. четвертую сторону обложки). Полученные кластеры также позволяют построить таблицы возможных аминокислотных замен. Эти таблицы могут быть визуализированы в виде графов связности, где ребро между пептидами соответствует возможной замене, а цвет — близости дипептидов.

Заключение

В работе предложен способ построения карт заселенности конформаций дипептидов с помощью разработанного программного комплекса DPPMClusterizer, а также метод кластерного анализа полученных изображений. По результатам ра-

боты создан атлас дипептидов, содержащий карты заселенности для всех возможных пар дипептидов с указанием их схожести. Полученный результат представляет интерес с точки зрения биофизики в контексте исследования пространственной структуры белков.

Развитие работы предполагает анализ более сложных пептидов, состоящих из трех и более аминокислот. При этом увеличение числа звеньев приведет к увеличению числа степеней свободы, а значит и размерности карт заселенности конформаций, а также к увеличению числа получаемых карт. Данное обстоятельство делает целесообразным разработку алгоритмов "свертки" многомерных изображений, а также численных мер их близости. В конечном счете при значительном увеличении размерности задачи может потребоваться принципиально иной подход к ее решению, основанный на новых методах анализа и визуализации многомерных данных.

Работа выполнена при частичной поддержке гранта Российского научного фонда № 14-50-00029 и гранта РФФИ № 16-07-00287.

Список литературы

1. **Северин Е. С.** Биохимия: учебник для вузов. М.: ГЭОТАР-Медиа, 2003. 779 с.

2. **Первичная** структура белков — аминокислоты: Лекция 1: URL: http://hpc.mipt.ru/wp-content/uploads/2012/05/Lecture01.pdf (дата обращения: 11.05.2017).

3. **Финкельштейн А. В.** Введение в физику белка. Курс лекций: URL: http://phys.protres.ru/lectures/protein_physics/ (дата обращения: 11.05.2017).

4. **Пименова И. Н., Пименов А. В.** Лекции по общей биологии. М.: Лицей, 2003. 215 с.

5. **RCSB** Protein Data Bank. URL: http://www.rcsb.org/pdb/ (дата обращения: 10.05.2017).

6. **Шайтан К. В., Ермолаева М. Д., Сарайкин С. С.** Молекулярная динамика олигопептидов 3 // Биофизика, 1999. Т. 44. С. 18—21.

7. **Ramachandran G. N., Ramakrishnan C., Sasisekharan V.** Stereochemistry of polypeptide chain configurations // *Journal of Molecular Biology*. 1963. N. 7. P. 95—103.

8. **Вторичные** структуры белков [Электронный ресурс]. URL: http://www.chem.msu.su/rus/teaching/kolman/74.htm (дата обращения: 10.05.2017).

9. **Generating** Ramachandran (phi/psi) plots for Proteins [Электронный ресурс]. URL: http://www2.warwick.ac.uk/fac/sci/moac/people/students/peter_cock/python/ramachandran/other/ (дата обращения: 11.05.2017).

10. **Шайтан К. В., Балабаев Н. К.** Алгоритмы и методы молекулярной динамики. М., 2010. 166 с.

11. **Protein** Folding and Denaturation. URL: http://www1.lsbu.ac.uk/water/protein_denatured.html (дата обращения: 11.05.2017).

12. **Котов А., Красильников Н.** Семинары: Кластеризация данных. URL: http://logic.pdmi.ras.ru/ (дата обращения: 11.05.2017).

13. **Федорук В. Г.** Искусственные нейронные сети. URL: http://bigor.bmstu.ru/?cnt/?doc = NN/base.cou (дата обращения: 11.05.2017).

14. **Грошев С. В., Смольникова Е. А.** Обзор подходов к визуализации многомерных данных // Технологии инженерных и информационных систем. 2015. № 3. С. 3—13.

15. **Инстар** Гроссберга. URL: http://bigor.bmstu.ru/?cnt/?doc= NN/013-neurons.mod/?cou = NN/base.cou (дата обращения: 11.05.2017).

16. **NAMD** — Scalable Molecural Dynamics. URL: https://www.ks.uiuc.edu/Research/namd (дата обращения: 11.05.2017).

17. **PDB** File Format. URL: https://www.rcsb.org/pdb/static/pdb. (дата обращения: 11.05.2017).

18. **NAMD** Input and Output types. URL: http://www.ks.uiuc.edu/Research/namd/2.9/ug/node10.html (дата обращения: 11.05.2017).

19. **The CHARMM** Force Field. URL: https://www.ks.uiuc.edu/science/node5 (дата обращения: 11.05.2017).

I. V. Basov[1], Student, e-mail: generalrot@rambler.ru,
S. V. Groshev[2], Senior Lecturer, e-mail: groshev_sergey@mail.ru,
A. P. Karpenko[2], Dr. Sci. (Phys.-Math.), Prof., Head of the Department, e-mail: apkarpenko@mail.ru,
K. V. Shaitan[3], Dr. Sci. (Phys.-Math.), Prof., e-mail: shaytan49@yandex.ru,
D. L. Shurov[4], Graduate student, e-mail:d.l.shurov@gmail.com,
Moscow, [1]Moscow Institute of State and Municipal Administration,
[2]Bauman Moscow State Technical University,
[3]Lomonosov Moscow State University, Faculty of Biology, Department of Bioengineering,
[4]Semenov Institute of Chemical Physics of Russian Academy of Sciences

Method for Construction and Cluster Analysis of Conformational Occupancy Density Maps of Dipeptides

We propose dynamic analogues of known Ramachandran plots, called conformational occupancy density maps. The proposed maps show the probability of values of the torsion angles of the main polypeptide chain. Maps are obtained based on molecular dynamics trajectories, and then are clustered using self-organizing maps. We present a method for constructing and cluster analysis of conformational occupancy density maps for all possible 400 dipeptides.

Keywords: *dipeptides, conformational dynamics of dipeptides, Ramachandran plot, conformational occupancy density map, cluster analysis, self-organizing map*

References

1. **Severin E. S.** *Biokhimiya: uchebnik dlya vuzov*, Moscow, GEOTAR-Media, 2003, 779 p. (in Russian).
2. **Pervichnaya** *struktura belkov — aminokisloty*, Lektsiya 1, URL: http://hpc.mipt.ru/wp-content/uploads/2012/05/Lecture01.pdf (data of access: 11.05.2017) (in Russian).
3. **Finkel'shtein A. V.** *Kurs lektsii: Vvedenie v fiziku belka*, URL: http://phys.protres.ru/lectures/protein_physics/ (data of access: 11.05.2017) (in Russian).
4. **Pimenova I. N., Pimenov A. V.** *Lektsii po obshchei biologii*, Moscow, Litsei, 2003, 215 p. (in Russian).
5. **RCSB** Protein Data Bank. URL: http://www.rcsb.org/pdb/ (data of access: 10.05.2017).
6. **Shaitan K. V., Ermolaeva M. D., Saraikin S. S.** Molekulyarnaya dinamika oligopeptidov 3, *Biofizika*, 1999, vol. 44, pp. 18—21. (in Russian).
7. **Ramachandran G. N., Ramakrishnan C., Sasisekharan V.** Stereochemistry of polypeptide chain configurations, *Journal of Molecular Biology*, 1963, no. 7: 95—103.
8. **Vtorichnye** *struktury belkov*, URL: http://www.chem.msu.su/rus/teaching/kolman/74.htm (data of access: 10.05.2017) (in Russian).
9. **Generating** Ramachandran (phi/psi) plots for Proteins, URL: http://www2.warwick.ac.uk/fac/sci/moac/people/students/peter_cock/python/ramachandran/other/ (data of access: 11.05.2017).
10. **Shaitan K. V., Balabaev N. K.** *Algoritmy i metody molekulyarnoi dinamiki*, Moscow, 2010, 166 p. (in Russian).
11. **Protein** Folding and Denaturation, URL: http://www1.lsbu.ac.uk/water/protein_denatured.html (data of access: 11.05.2017).
12. **Kotov A., Krasil'nikov N.** Seminary: *Klasterizatsiya dannykh*, URL: http://logic.pdmi.ras.ru/ (data of access: 11.05.2017) (in Russian).
13. **Fedoruk V. G.** *Iskusstvennye neironnye seti*, URL: http://bigor.bmstu.ru/?cnt/?doc=NN/base.cou (data of access: 11.05.2017) (in Russian).
14. **Groshev S. V., Smol'nikova E. A.** Obzor podkhodov k vizualizatsii mnogomernykh dannykh, *Tekhnologii inzhenernykh i informatsionnykh sistem*, 2015, no. 3, pp. 3—13 (in Russian).
15. **Instar** Grossberga. URL: http://bigor.bmstu.ru/?cnt/?doc=NN/013-neurons.mod/?cou=NN/base.cou (data of access: 11.05.2017) (in Russian).
16. **NAMD** — Scalable Molecural Dynamics, URL: https://www.ks.uiuc.edu/Research/namd (data of access: 11.05.2017).
17. **PDB** File Format, URL: https://www.rcsb.org/pdb/static/pdb (data of access: 11.05.2017).
18. **NAMD** Input and Output types, URL: http://www.ks.uiuc.edu/Research/namd/2.9/ug/node10.html (data of access: 11.05.2017).
19. **The CHARMM** Force Field, URL: https://www.ks.uiuc.edu/science/node5 (data of access: 11.05.2017).

Адрес редакции:

107076, Москва, Стромынский пер., 4

Телефон редакции журнала **(499) 269-5510**

E-mail: it@novtex.ru

Технический редактор *Е. В. Конова*.

Корректор *Е. В. Комиссарова*.

Сдано в набор 10.07.2017. Подписано в печать 23.08.2017. Формат 60×88 1/8. Бумага офсетная.

Усл. печ. л. 8,86. Заказ IT917. Цена договорная.

Журнал зарегистрирован в Министерстве Российской Федерации по делам печати, телерадиовещания и средств массовых коммуникаций.

Свидетельство о регистрации ПИ № 77-15565 от 02 июня 2003 г.

Оригинал-макет ООО "Адвансед солюшнз". Отпечатано в ООО "Адвансед солюшнз".

119071, г. Москва, Ленинский пр-т, д. 19, стр. 1.