

РАЗРАБОТКА АЛГОРИТМОВ ДЛЯ УПОРЯДОЧИВАНИЯ СТРУКТУР БЕЛКОВ

Соколов А.А.

студент НИУ ИТМО, ansokolmail@gmail.com

Буздалов М.В.

аспирант НИУ ИТМО, mbuzdalov@gmail.com

Аннотация: Данная работа посвящена моделированию движения белковых молекул. Исследование белковых взаимодействий является актуальной и малоизученной темой. Предлагается метод построения гладких конформационных движений белков.

Введение

Современная биоинформатика — молодая и бурно развивающаяся наука. Белки в клетках живых организмов выполняют множество разнообразных функций [1]. Основным предметом изучения протеомики являются белки и их взаимодействия. Одной из интереснейших задач протеомики является построение траекторий движения молекул.

Описание задачи

Для того чтобы получить представление о структуре белка, нужно использовать один из методов исследования химических объектов. Наиболее распространенные методы: спектроскопия ядерного магнитного резонанса и рентгеноструктурный анализ. В интернете доступны электронные банки белков, например, Protein Data Bank [2]. В указанном банке белков имеется 89 740 записей (по состоянию на 16 апреля 2013 г.), и их число постоянно растет.

Однако, эти отдельные снимки белка не позволяют получить полного представления о том, как молекула движется в пространстве. Возникает задача компьютерного построения траектории движения белка на основе некоторого числа конформаций, полученных из банков данных.

Метод построения графа гладких конформационных движений

Входными данными для описываемого метода является N конформаций исследуемого белка, полученных из банка белков. Назовем указанные конформации *базовыми*. Результатом работы метода является новый набор конформаций исследуемого белка и граф, описывающий, в каком порядке белок может принимать конформации из этого набора.

Метод состоит из следующих этапов:

- Вычисление конформационных движений между всеми парами конформаций и их стоимости.
- Удаление заведомо неверных конформационных движений.
- Для каждой базовой конформации: определение несовместимых конформационных движений из заданной конформации.
- Построение графа возможных конформационных движений.

Далее каждый этап будет изложен подробнее. В данном разделе считается, что базовые конформации исследуемого белка пронумерованы числами от одного до N .

Вычисление конформационных движений между всеми парами конформаций и их стоимости. Для всех пар конформаций (S, T) с номерами $1 \leq S < T \leq N$ и некоторого, заранее заданного числа K с использованием метода минимизации переноса масс [3, 4] вычисляется стоимость конформационного движения $L(S, T) = L(T, S)$ и цепочка $C(S, T)$ из $K - 2$ промежуточных конформаций. Результатом работы является неориентированный граф G_1 , вершинами которого служат базовые конформации, и между каждой парой различных вершин есть ребро, весом которого является стоимость конформационного движения между соответствующими конформациями.

В дополнение к этому, для каждой базовой конформации A выбирается один из ее твердотельных поворотов в пространстве $R(A)$, и для всех базовых конформаций $B \neq A$ строится цепочка промежуточных конформаций $F(A, B)$, эквивалентная цепочке $C(\min(A, B), \max(A, B))$ с точностью до порядка следования конформаций (от начала к концу или от конца к началу) и до твердотельного вращения конформаций. Направление этой цепочки выбирается таким, что первой конформацией является конформация A , а ее ориентация в пространстве совпадает с $R(A)$. Кроме того, все остальные конформации выравниваются относительно предыдущей путем минимизации RMSD.

Удаление заведомо неверных конформационных движений. По причине использования приближенных стохастических алгоритмов оптимизации конформационного движения может получиться так, что для некоторой тройки конформаций A, B, C выполняется неравенство $L(A, B) + L(B, C) \ll L(A, C)$. Это означает, что оптимизация конформационного движения от A до C проведена неудовлетворительным образом. Если $2 \times (L(A, B) + L(B, C)) < L(A, C)$, то ребро из A в C удаляется из графа.

Определение несовместимых конформационных движений из заданной конформации. Если белок при движении из конформации A в конформацию B проходит некоторую промежуточную конформацию C , то это движение из общих соображений должно происходить без резкого изменения направлений движения молекул, составляющих белок.

Из этого следует, что если белок пришел в конформацию C из некоторой конформации A , то он может продолжить движение не во все конфор-

мации — например, он не может вернуться обратно в конформацию A или двигаться в конформацию B , похожую на A , так как это повлечет изменение направления движений молекул на угол, близкий к 180° .

Определение несовместимых конформационных движений производится следующим образом. Пусть A — заданная базовая конформация. Для каждой базовой конформации $B \neq A$ строится список $D(A, B)$ разностей координат атомов между $R(A)$ и второй конформацией из цепочки $F(A, B)$. Тогда для атома с номером q величина $M_q(A, B) = D_q(A, B)/L(A, B)$ будет являться приближенным значением производной координат атома по времени при начале движения белка из конформации A в конформацию B . Далее, для каждой пары базовых конформаций (B, C) , $B \neq A$, $C \neq A$, $B \neq C$ вычисляется величина $Q(B, C) = \max_q (M_q(A, B) + M_q(A, C))$. Чем ближе к нулю эта величина, тем с большей долей уверенности можно утверждать, что движение из A в C может быть продолжением движения из B в A . Выберем пороговое значение ϵ и определим возможность движения из B в C через A следующим образом: $\theta(A, B, C) = 1$, если $Q(B, C) < \epsilon$, и 0 в противном случае.

Построение графа возможных конформационных движений. Строится ориентированный граф G_2 с $O(N^2)$ вершинами. В данном графе имеются вершины следующих типов:

- для каждой базовой конформации A — вершина S_A , соответствующая началу движения белка из конформации A ;
- для каждой базовой конформации A — вершина T_A , соответствующая окончанию движения белка в конформации A ;
- для каждой пары базовых конформаций A, B , $A \neq B$ — вершина S_{AB} , соответствующая началу движения белка из A в B ;
- для каждой пары базовых конформаций A, B , $A \neq B$ — вершина T_{AB} , соответствующая окончанию движения белка из A в B .

Приведем алгоритм построения множества ребер графа G_2 :

- Для всех упорядоченных пар базовых конформаций A, B , $A \neq B$:
 - добавляется ребро с весом 0 из S_A в S_{AB} ;
 - добавляется ребро с весом 0 из T_{BA} в T_A ;
 - если существует ребро E из A в B в графе G_1 , то добавляется ребро с весом, равным весу E , из S_{AB} в T_{AB} .
- Для всех упорядоченных троек базовых конформаций A, B, C , $B \neq A$, $C \neq A$, $B \neq C$:
 - $\theta(A, B, C) = 1$, то добавляется ребро с весом 0 из T_{BA} в S_{AC} .

Таким образом, в графе G_2 имеется $O(N^2)$ вершин и $O(N^3)$ ребер.

Свойством графа G_2 является то, что любой путь из S_A в T_B взаимно однозначно соответствует такому пути в графе G_1 , в котором любые два ребра подряд не являются несовместимыми, при этом веса этих путей равны. Используя граф G_2 , можно искать гладкие пути между конформациями.

Заключение

Построение возможных белковых траекторий позволяет получить представление о том, как движется белок в пространстве, и понять, какие конформации этого белка возможны. Исследование белков и их взаимодействий позволит ускорить разработку лекарственных средств.

Литература

1. *Финкельштейн А.В. Птицын О.Б.* Физика Белка. Курс лекций. 2005, 456 с.
 2. RCSB Protein Data Bank <http://www.rcsb.org>
 3. *Krass A., Nikolenko S., Stepanov E. and Porozov Yu.* Modeling of Protein Conformational Movements Based on Principle of Average Action. ISCB-ECCB 2011, Vienna, 2011, July 15–19.
 4. *Порозов Ю.Б.* Моделирование конформационной подвижности белков XIX Национальный конгресс «Человек и Лекарство 2012». Симпозиум «Биоинформатика и компьютерное конструирование лекарств». Москва, 23–27 апреля 2012.
-