Построение матрицы попарных потенциалов для решения задачи обратного фолдинга

А.Р. Рубинштейн
Руководитель: М.Е. Карасиков

Московский физико-технический институт
Факультет управления и прикладной математики

Москва, 2019

Цели

- **Задача обратного фолдинга:** предсказание последовательностей аминокислот, которые сворачиваются в заданную пространственную структуру - белок
- **Актуальность:** результаты исследования способствуют определению молекул, обладающих необходимыми свойствами

Задачи

1. **Построение ранжирующей функции для оценивания третичной структуры белка**
   - Извлечение признаков, основанных на попарном взаимодействии атомов
   - Подбор адекватных методов нормализации для признаков разных типов
   - Обучение модели
2. **Построение матрицы попарных потенциалов**
   - Составление попарных потенциалов на основе обученных весов модели
   - Построение матрицы свободной энергии для всех рассматриваемых пар аминокислот
   - Постановка задачи оптимизации

Проблемы

- Огромная размерность
- NP-трудная задача дискретной оптимизации
- Необходимость лабораторных экспериментов для проверки качества

Литература

* [Andrii Riazanov, Mikhail Karasikov, and Sergei Grudinin, 2016]
  Inverse protein folding problem via quadratic programming - *Постановка задачи оптимизации*
* [Карасиков М.Е., Стрижов В.В, 2017]
  Построение ранжирующей функции для прогнозирования третичной структуры белка - *Создание скоринговой функции и извлечение признаков*

Задача структурной биологии

Amino Acids$(AA) = \{ALA, ARG, ...TYR, VAL\}$

Atoms of AA :
(основная часть, скелет)$[N, CA, C, H, O] +$
(боковая цепь)$[HA, CB, HB1, HB2, SY, HY]$

Нас интересует отображение подмножества скелетов в подмножество аминокислот. То есть прогнозирование первичной последовательности аминокислот по третичной структуре белка.

Постановка задачи

Пусть белок состоит из N аминокислот. Множество $\mathcal{C} = \{1, 2, \ldots, 20\}$ содержит индексы, которые соответствуют различным типам аминокислот.

Пусть $\vec{a} = (a_1, \ldots, a_N)$ последовательность остатков белка, причем $a_i \in \mathcal{C}$. Третичная структура белка с заданным положением основных атомов (скелет) - $\vec{b}$, а попарный потенциал межатомных взаимодействий - $E(\vec{a}, \vec{b})$.

Постановка задачи

Каждой оптимальной последовательности аминокислот a^* соответствует скелет $\vec{b}_0$:

$$\vec{b}_0 = \arg\min_{\vec{b}} E(\vec{a^*}, \vec{b}). \tag{1}$$

Мы можем сформулировать задачу обратного фолдинга как оптимизационную:

$$\|\vec{b}_0 - \arg\min_{\vec{b}} E(\vec{a}, \vec{b})\| \to \min_a . \tag{2}$$

Положим, что $E_{kl} : \mathcal{C}^2 \to \mathbb{R}$ определяет симметричный попарный потенциал, выражающий взаимодействие между остатками белка k и l.

Тогда задачу можно переформулировать следующим образом:

$$\phi_{\vec{b}}(a) := E(\vec{a}, \vec{b}) = \sum_{j=1}^{N} \sum_{i=1}^{N} E_{ij}(a_i, a_j) \to \min_{a_1, a_2, \ldots, a_N \in \mathcal{C}}. \tag{3}$$

Структура матрицы попарных потенциалов

$$\mathbf{Q} = \begin{bmatrix} [E_{11}] & [E_{12}] & \cdots & [E_{1N}] \\ [E_{21}] & [E_{22}] & \cdots & [E_{2N}] \\ \vdots & \vdots & \ddots & \vdots \\ [E_{N1}] & [E_{N2}] & \cdots & [E_{NN}] \end{bmatrix},$$

$$E_{ij} = \begin{bmatrix} E_{ij}(c_1, c_1) & E_{ij}(c_1, c_2) & \cdots & E_{ij}(c_1, c_{20}) \\ E_{ij}(c_2, c_1) & E_{ij}(c_2, c_2) & \cdots & E_{ij}(c_2, c_{20}) \\ \vdots & \vdots & \ddots & \vdots \\ E_{ij}(c_{20}, c_1) & E_{ij}(c_{20}, c_2) & \cdots & E_{ij}(c_{20}, c_{20}) \end{bmatrix},$$

$$\mathbf{Q} \in \mathbb{R}^{20N \times 20N}, \quad E_{ij} \in \mathbb{R}^{20 \times 20}.$$

Метод построения матрицы

- Получение элементов матрицы для каждого из типов признаков
  - ✓ Блоки соответствуют парам аминокислот белка
  - ✓ Мы варьируем типы этих аминокислот
  - ✓ Каждому типу признаков соответствует вектор-гистограмма
  - ✓ Гистограммы нормализуются так же, как в обученной модели
  - ✓ Обученные веса домножаются на элементы гистограммы и соотносятся с ячейками матрицы
- Построена матрица попарных потенциалов
- Сформулирована задача оптимизации для полученной матрицы

Заключение и результаты

- Рассмотрена функция, ранжирующая 3D структуры белка
  - ✓ **Парно-сепарабельная** скоринговая функция
  - ✓ Использует только основные атомы белка
- Построена матрица попарных потенциалов
- Сформулирована задача оптимизации для полученной матрицы

Спасибо за внимание!