Построение матрицы попарных потенциалов для решения задачи обратного фолдинга

Рубинштейн Александр

Московский физико-технический институт

Курс: Автоматизация научных исследований в машинном обучении (практика, В. В. Стрижов)/ весна 2019

Задача обратного фолдинга:

Цель

предсказание последовательностей аминокислот, которые сворачиваются в заданную пространственную структуру - белок.

Проблемы

- Огромная размерность признакового пространства
- NP-трудная задача дискретной оптимизации
- Необходимость лаборатрных экспериментов для проверки качества

Подход к решению

- Построение ранжирующей функции для оценивания третичной структуры белка
- Построение матрицы попарных потенциалов
- Постановка задачи оптимизации

Задача структурной биологии

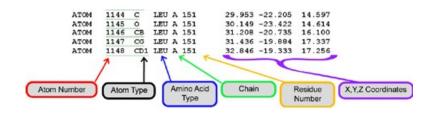
Задан набор аминокислот:

 $C = \{ALA, ARG, ...TYR, VAL\}$

 $\mathbf{a}=(a_1,\ldots,a_N)$ последовательность остатков белка, где $a_i\in\mathcal{C}.$

Заданы координаты атомов каждой аминокислоты:

(**b** - основная часть, скелет)[N, CA, C, H, O] + (боковая цепь)[HA, CB, HB1, HB2, SY, HY]



Литература

Сведение к задаче оптимизации

 Andrii Riazanov, Mikhail Karasikov, and Sergei Grudinin Inverse protein folding problem via quadratic programming, 2016.

Извлечение признаков и создание скоринговой функции

 Карасиков М.Е., Стрижов В.В
 Построение ранжирующей функции для прогнозирования третичной структуры белка, 2017.

Постановка задачи

 $E(\mathbf{a},\mathbf{b})$ - попарный потенциал межатомных взаимодействий

$$\phi_{\mathbf{b}}(a) := E(\mathbf{a}, \mathbf{b}) = \sum_{j=1}^{N} \sum_{i=1}^{N} E_{ij}(a_i, a_j)$$
 (1)

Каждой оптимальной последовательности аминокислот соответствует скелет \mathbf{b}_0 :

$$\mathbf{b}_0 = \arg\min_{\mathbf{b}} E(\mathbf{a}^*, \mathbf{b}). \tag{2}$$

Можно сформулировать задачу обратного фолдинга как оптимизационную:

$$\|\mathbf{b}_0 - \arg\min_{\mathbf{b}} E(\mathbf{a}, \mathbf{b})\| \to \min_{\mathbf{a}}.$$
 (3)

Структура матрицы попарных потенциалов

Оптимизационную задачу можно привести к следующему виду:

$$\min_{\mathbf{x}=[\mathbf{x}_1,\dots,\mathbf{x}_N]} \mathbf{x} \mathbf{Q} \mathbf{x} \tag{4}$$

s.t.
$$\mathbf{x}_k \in \{0, 1\}^{20}, \|\mathbf{x}_k\|_0 = 1$$
 (5)

где

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

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

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

Этапы работы

Ранжирующая функция

- Извлечение признаков, основанных на попарном взаимодействии атомов
- Подбор адекватных методов нормализации для признаков разных типов
- Обучение модели

Матрица попарных потенциалов

- Составление попарных потенциалов на основе обученных весов модели
- Построение матрицы свободной энергии для всех рассматриваемых пар аминокислот
- Постановка задачи оптимизации

Метод построения матрицы

- А Блоки соответствуют парам аминокислот белка
- В Мы варьируем типы этих аминокислот
- С Каждому типу признаков соответствует вектор-гистограмма
- Гистограммы нормализуются так же, как в обученной модели
- E Обученные веса домножаются на элементы гистограммы и соотносятся с ячейками матрицы

Заключение и результаты

- Рассмотрена функция, ранжирующая 3D структуры белка
- Построена матрица попарных потенциалов
- Сформулирована задача оптимизации для полученной матрицы
- Актуальность: результаты исследования способствуют определению молекул, обладающих необходимыми свойствами