Скоринговая функция для обратной задачи вычислительного построения белков CPD

Александр Рубинштейн

8 декабря 2018

Отчет

Обзор

- Основные идеи
- Составные части проекта
- Выводы

Основные идеи

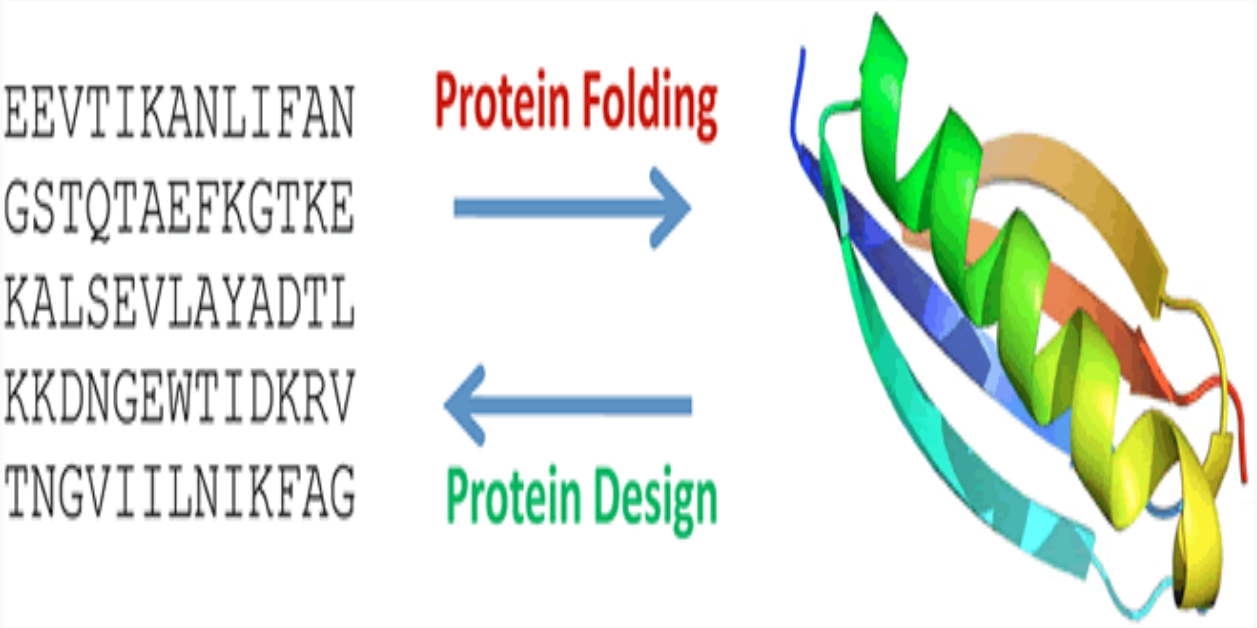

Figure: Задача построения белков по последовательности аминокислот

Источник: http://www.cbms.k.u-tokyo.ac.jp/english/lab/zhang.html

Основные идеи

Использование скоринговой функции:

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

$$E_{ij}(a_i, a_j) = K \cdot (\vec{w}^T \vec{f}_{ij}(a_i, a_j, \vec{b}_0) + C) \tag{2}$$

Основные идеи

Генерация матриц энергии:

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

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

Основные идеи

Решение задачи оптимизации:

$$\begin{aligned} \min_{\vec{x}\in\{0,1\}^{20N}} \quad & \vec{x}^T Q \vec{x} \\ \text{s.t.} \quad & \mathbf{A}\vec{x} = \mathbf{1}_N, \end{aligned} \tag{3}$$

где

$$\mathbf{A} = \begin{bmatrix} \underbrace{1\cdots1}_{20} & \underbrace{0\cdots0}_{20} & \cdots\cdots & \underbrace{0\cdots0}_{20} \\ 0\cdots0 & 1\cdots1 & \cdots\cdots & 0\cdots0 \\ \vdots & \vdots & \ddots & \vdots \\ 0\cdots0 & 0\cdots0 & \cdots\cdots & 1\cdots1 \end{bmatrix}, \quad \mathbf{A} \in \{0,1\}^{N\times 20N}.$$

Базовый алгоритм

Три основные стадии:

- Извлечение признаков из модели белка
- Обучение на CASP и предсказание GDT-TS
- Присвоение энергии (скора) каждому белку

Точность модели

$\mathsf{B}(y_i, \hat{y}_j)$ соответствует скору белка при замене y_i на $\hat{y}_j$ в j-ой позиции последовательности.

$$S(\vec{y}_{\text{nat}}, \vec{y}_{\text{pred}}) = \frac{\sum_{k=1}^{N} \mathsf{B}((\vec{y}_{\text{nat}})_k, (\vec{y}_{\text{pred}})_k)}{\sum_{k=1}^{N} \mathsf{B}((\vec{y}_{\text{nat}})_k, (\vec{y}_{\text{nat}})_k)} \tag{4}$$

Хорошее предсказание соответствует $S > 0$, а лучшие результаты имеют S близкую к 1. Отрицательные S означают предсказания плохого качества.