

Формулировка и решение задачи оптимизации, сочетающей классификацию и регрессию, для оценки энергии связывания белка и маленьких молекул

Илья Игашов

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам
(практика, В.В. Стрижов)

Группа 594, весна 2018

Консультанты: Сергей Грудинин, Мария Кадукова

Цель работы

Определение нативных поз и предсказание свободной энергии связывания для комплексов "белок-лиганд" методами машинного обучения.

Проблемы

- Задача вычислительно сложная.
- Существующие методы не дают желаемого качества предсказаний.

Метод решения

Объединение существующих методов классификации и регрессии в одну оптимизационную задачу.

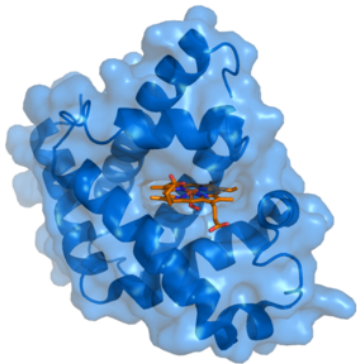
Молекулярный докинг и скоринговые функции

- 1 Lengauer T, Rarey M (Jun 1996). "Computational methods for biomolecular docking". Current Opinion in Structural Biology. 6 (3): 402–6
- 2 Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. Molecular Docking: "A powerful approach for structure-based drug discovery"

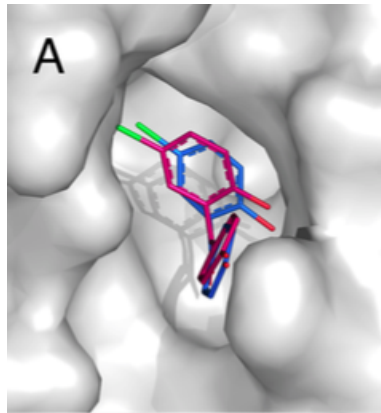
Convex-PL. Классификация и регрессия

- 1 Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization. Maria Kadukova, Sergei Grudinin
- 2 Predicting binding poses and affinities for protein-ligand complexes in the 2015 D3R Grand Challenge using a physical model with a statistical parameter estimation. Sergei Grudinin, Maria Kadukova, Andreas Eisenbarth, Simon Marillet, Frédéric Cazals

Постановка задачи: модель взаимодействия



(a) Комплекс "белок-лиганд"



(b) Сайт связывания лиганда с белком

Постановка задачи: модель взаимодействия

- M_1 типов атомов лигандов
- M_2 типов атомов белков
- Нативной позе соответствует минимум энергии связывания
- Целевая функция:

$$E(n(r)) = \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} \int_0^{r_{\max}} n^{kl}(r) f^{kl}(r) dr$$

- Численные плотности распределений пар атомов по расстоянию между ними:

$$n^{kl}(r) = \sum_{i,j} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(r-r_{ij})^2}{2\sigma^2}}$$

- $f^{kl}(r)$ – неизвестные "скоринговые потенциалы"
- После разложения функций $n^{kl}(r)$ и $f^{kl}(r)$ по полиномиальному базису $E(n(r))$ принимает вид:

$$E(n(r)) \approx \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} \sum_{q=0}^Q w_q^{kl} x_q^{kl} = (\mathbf{w}, \mathbf{x})$$

$$\mathbf{w}, \mathbf{x} \in \mathbb{R}^{Q \times M_1 \times M_2}$$

Постановка задачи: оптимизационная задача

Дано

Множество троек (\mathbf{x}, y_i, s_i) , $i = 1, \dots, N$, где $\mathbf{x} \in \mathbb{R}^{Q \times M_1 \times M_2}$ – структурный вектор (вектор признаков i -го комплекса), $y_i \in \{-1, 1\}$ – поза i -го комплекса (нативная или ненативная), s_i – значение энергии связывания i -го комплекса.

Задача оптимизации

$$\underset{\mathbf{w}, \xi_i}{\text{minimize:}} \quad \frac{1}{2} \|\mathbf{w}\|^2 + \sum_i C_i \xi_i + C_r \sum_i f(\mathbf{x}_i, \mathbf{w}, s_i)$$

$$\text{subject to:} \quad y_i(\mathbf{w}, \mathbf{x}_i) - 1 + \xi_i \geq 0, \\ \xi_i \geq 0,$$

где ξ_i – оптимизируемые параметры модели, $f(\mathbf{x}_i, \mathbf{w}, s_i)$ – некоторая функция потерь регрессии (Mean Square Error), C_i , C_r – коэффициенты регуляризации.

Постановка задачи: уточнения

- Функцию потерь классификации "hinge-loss" заменим на "log-loss":

$$V(\mathbf{w}, \mathbf{x}_i, y_i) = \frac{1}{\ln 2} \ln \left(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i} \right)$$

- Введем замену переменных:

$$\mathbf{w}' = \mathbf{A}\mathbf{w} - \mathbf{B},$$

$$\mathbf{A} = \left[\frac{1}{2} \mathbf{I} + C_r \mathbf{X} \mathbf{X}^T \right]^{\frac{1}{2}},$$

$$\mathbf{B} = C_r \left[\frac{1}{2} \mathbf{I} + C_r \mathbf{X} \mathbf{X}^T \right]^{-\frac{1}{2}} \mathbf{X} \mathbf{s}$$

Задача оптимизации

$$\underset{\mathbf{w}'}{\text{minimize:}} \quad \|\mathbf{w}'\|^2 + \sum_i C_i \ln \left(1 + \exp \left\{ -y_i \left[\mathbf{A}^{-1} (\mathbf{w}' + \mathbf{B}) \right]^T \mathbf{x}_i \right\} \right)$$

- База PDDBind: содержит экспериментально определенные комплексы белков и лигандов и соответствующие им измеренные значения энергии связывания.
- "General dataset" содержит 14,620 комплексов белков и лигандов, их позы, а также значения энергии связывания данных комплексов.
- "Refined dataset" содержит отобранные из "general" данные лучшего качества.
- Для каждого комплекса в базе представлены:
 - 19 конформаций (1 нативная и 18 ненативных),
 - 23 типа белковых атомов и 40 типов атомов лигандов,
 - Значение свободной энергии связывания для нативной позы.

Эксперимент: план работы

- ✓ Прочитать и подготовить данные.
- ✓ Произвести замену переменных.
- ✓ Подать данные на вход солверу (на данный момент работаем с `scipy.optimize.minimize`).
- Получить результаты, оценить качество предсказаний с помощью коэффициента корреляции (для значений свободной энергии).
- Подобрать оптимальные значения для коэффициентов регуляризации.
- Вернуться к hinge-loss и провести эксперимент с этой функцией потерь.
- Возможно, перейти на более быстрый солвер (например, `CVXPY`).

На данный момент:

- Сформулирована оптимизационная задача.
- Запущен базовый алгоритм.

Предстоит:

- Оптимизировать вычисления.
- Получить осмысленные результаты.
- Поработать с параметрами модели и функциями потерь.