

Оптимизация параметров модели, 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 сложна из-за высокой размерности пространства признаков.
- Существующие методы не дают желаемого качества предсказаний.

Метод решения

Объединение существующих методов классификации и регрессии в одну оптимизационную задачу.

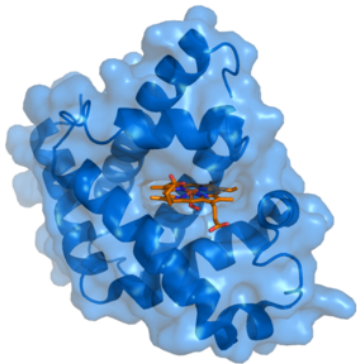
Молекулярный докинг и скоринговые функции

- 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"

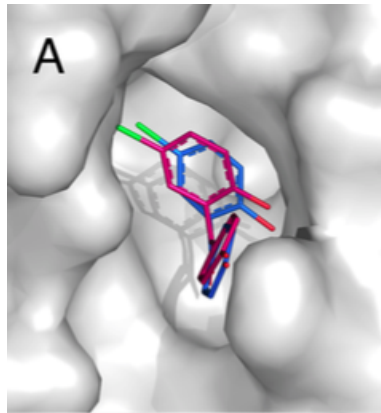
Convex-PL. Классификация и регрессия

- 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

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

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

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

Постановка задачи: уточнения

- Введем замену переменных: $\mathbf{w}' = \mathbf{A}\mathbf{w} - \mathbf{B}$,
 $\mathbf{A} = [\frac{1}{2}\mathbf{I} + C_r\mathbf{X}\mathbf{X}^T]^{\frac{1}{2}}$, $\mathbf{B} = C_r [\frac{1}{2}\mathbf{I} + C_r\mathbf{X}\mathbf{X}^T]^{-\frac{1}{2}} \mathbf{X}\mathbf{s}$

Задача оптимизации

$$\underset{\mathbf{w}', \xi_i}{\text{minimize:}} \quad \frac{1}{2} \|\mathbf{w}'\|^2 + C \sum_i \xi_i$$

$$\text{subject to:} \quad y_i (\mathbf{A}^{-1} (\mathbf{w}' + \mathbf{B}))^T \mathbf{x}_i - 1 + \xi_i \geq 0, \\ \xi_i \geq 0.$$

Двойственная задача

$$\underset{\lambda_i}{\text{minimize:}} \quad - \sum_i \lambda_i + \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \langle \hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j \rangle$$

$$\text{subject to:} \quad 0 \geq \lambda_i \geq C,$$

$$\text{где } \hat{\mathbf{X}} = (\mathbf{A}^{-1})^T \mathbf{X}.$$

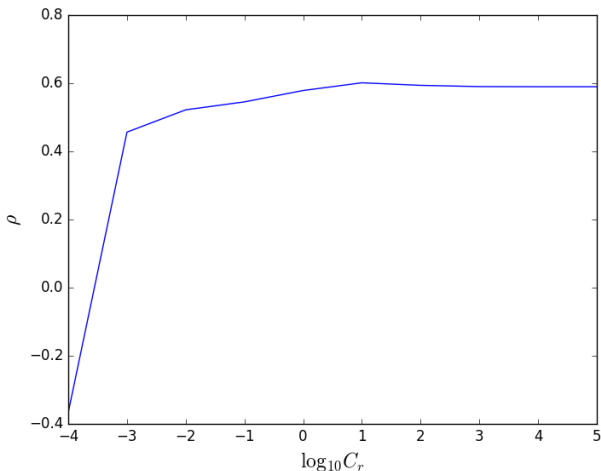
- База PDDBind: содержит экспериментально определенные комплексы белков и лигандов и соответствующие им измеренные значения энергии связывания.
- "General dataset" содержит 14,620 комплексов белков и лигандов, их позы, а также значения энергии связывания данных комплексов.
- "Refined dataset" содержит 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 LIBLINEAR
- Подобрать оптимальные значения для коэффициентов регуляризации
- Получить результаты, оценить качество предсказаний с помощью коэффициента корреляции (для значений свободной энергии)
- Получить результаты на докинг-тестах из бенчмарка CASF

Эксперимент: результаты

Зависимость коэффициента корреляции Пирсона ρ от значения логарифма коэффициента регуляризации C_r :



- Коэффициент корреляции Пирсона $\rho = 0.521182$
- Результаты на докинг-тесте из бенчмарка CASF:

t1 native 28.2 t1q1n 55.9	t1q1 64.1 t1q2n 67.2	t1q2 72.3 t1q3n 73.3	t1q3 77.4
t2 native 46.7 t2q1n 68.7	t2q1 78.5 t2q2n 78.5	t2q2 84.1 t2q3n 85.1	t2q3 88.2
t3 native 56.9 t3q1n 73.3	t3q1 83.1 t3q2n 84.1	t3q2 88.7 t3q3n 90.3	t3q3 92.3

На данный момент:

- Сформулирована оптимизационная задача
- Проведен базовый эксперимент
- Полученные результаты не превосходят того, что было получено ранее

Предстоит:

- Оптимизировать вычисления
- Точнее подобрать параметры модели
- Расширить пространство признаков, добавив дополнительные дескрипторы связей и вращений комплексов