Оценка энергии связывания белка и маленьких молекул

Анастасия Грачёва

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/Группа 694, весна 2019

Белково-лигандные взаимодействия

Решаемая задача

Поиск маленьких молекул - лигандов, наиболее сильно взаимодействующих с исследуемым белком.

Цель

Повышение качества предсказания нативной позы лиганда.

Существующее решение

Обучив классификатор, получаем скоринговый вектор. Его скалярное произведение со структурным вектором нативной позы минимально, но не обязательно равно реальному значению энергии связывания.

Предложение

Использовать данные о свободной энергии связывания молекул и решать одновременно задачи регрессии и классификации.

Модель взаимодействий

Есть P нативных комплексов белков-лигандов $\{C_{i0}\}_{i=1}^{P}$. Применив к лигандам изометрические преобразования, сгенерируем для каждого комплекса D ненативных поз $\{C_{ij}\}_{j=1}^{D}$.

Требуется найти функцию E, т.ч. $E(C_{i0}) < E(C_{ij})$

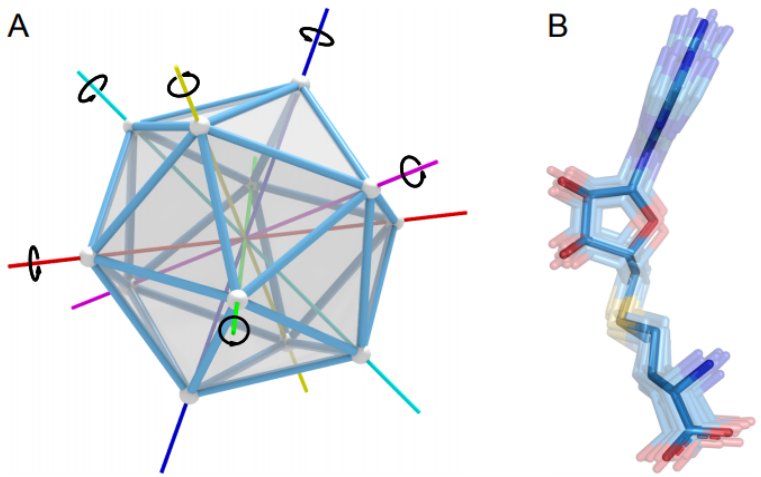


Figure 4: Decoys generation procedure. A : Six icosahedral axes about which we rotate the ligand. B : An example of a native ligand configuration with the corresponding 18 decoys generated with RMSD of 0.5 Å. These are 12 rotational decoys and 6 translational decoys.

Модель взаимодействий

Предположения

1. будем рассматривать комплекс "белок-лиганд" как набор атомов, каждый из которых имеет некоторый тип
2. E определяется только взаимодействиями между парами атомов комплекса
3. E зависит только от распределения расстояний между взаимодействующими атомами

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

$f^{kl}(r)$ – неизвестные функции взаимодействия между атомами
$n^{kl}(r)$ – известные плотности распределений пар атомов по расстоянию между ними:

Модель взаимодействий

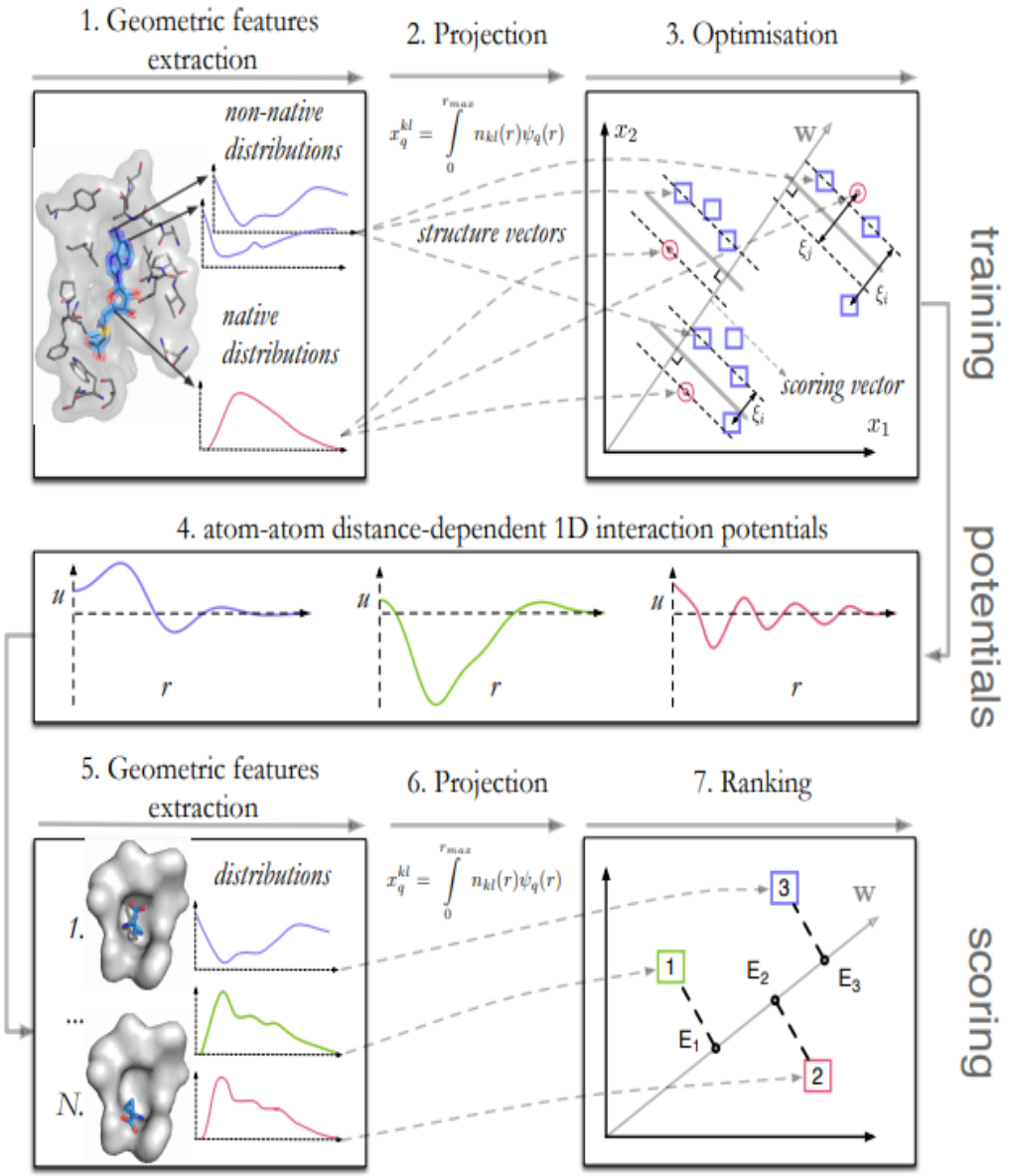
Разложим $f^{kl}(r)$ и $n^{kl}(r)$ по полиномиальному базису до порядка Q:

$$f^{kl}(r) \approx \sum_{q}^{Q} w_q^{kl} \psi_q(r), \; n^{kl}(r) \approx \sum_{q}^{Q} x_q^{kl} \psi_q(r), \tag{2}$$

$$E(n(r)) \approx \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} \sum_{q=0}^{Q} w_q^{kl} x_q^{kl} = \langle \mathbf{w}, \mathbf{x} \rangle, \tag{3}$$

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

Постановка задачи



Постановка задачи

$$
\begin{aligned}
\min_{\mathbf{w},\xi_{ij}}: \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{ij}\xi_{ij} + C_r\sum_{i}\left(\langle \mathbf{w}, \mathbf{x}_{i0}\rangle - s_i\right)^2 \\
\text{s. t.:} \quad & y_{ij}[\langle \mathbf{w}, \mathbf{x}_{ij}\rangle - b_i] - 1 + \xi_{ij} \geq 0, \\
& \xi_{ij} \geq 0, \\
& i \in \{1, \ldots, P\} \\
& j \in \{0, \ldots, D\}
\end{aligned}
\tag{4}
$$

Решение

Приведение к одному квадратичному слагаемому

$$\begin{aligned} \min_{\mathbf{w}', b_i, \xi_{ij}} \quad & \frac{1}{2}\|\mathbf{w}'\|^2 + C\sum_{ij}\xi_{ij} \\ \text{s. t.:} \quad & y_{ij}[\left(\mathbf{A}^{-1}\left(\mathbf{w}' + \mathbf{B}\right)\right)^{\mathsf{T}}\mathbf{x}_{ij} - b_i] - 1 + \xi_{ij} \geq 0, \\ & \xi_{ij} \geq 0 \end{aligned} \tag{5}$$

Поиск двойственной задачи:

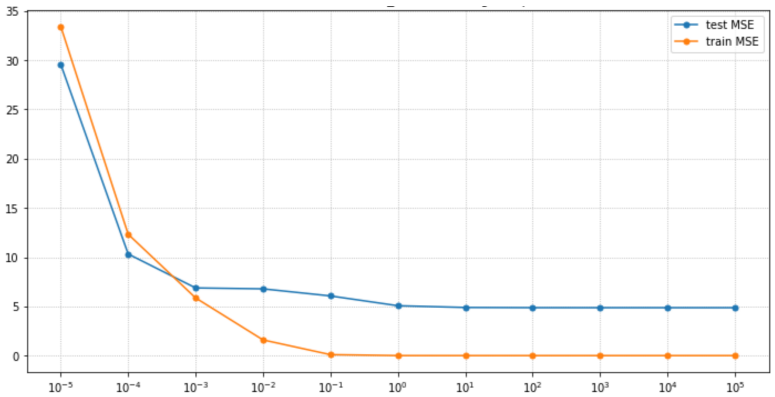
$$\begin{aligned} \min_{\lambda_{ij}} \quad & \frac{1}{2}\sum_{(i,j),(p,q)}\lambda_{ij}\lambda_{pq}y_{ij}y_{pq}\langle\widehat{\mathbf{x}}_{ij}, \widehat{\mathbf{x}}_{pq}\rangle + \sum_{ij}\lambda_{ij}\left(y_{ij}\langle\mathbf{B}, \widehat{\mathbf{x}_{ij}}\rangle - 1\right) \\ \text{s. t.:} \quad & 0 \leq \lambda_{ij} \leq C, \\ & \forall i, \sum_{j}\lambda_{ij}y_{ij} = 0 \end{aligned} \tag{6}$$

Литература

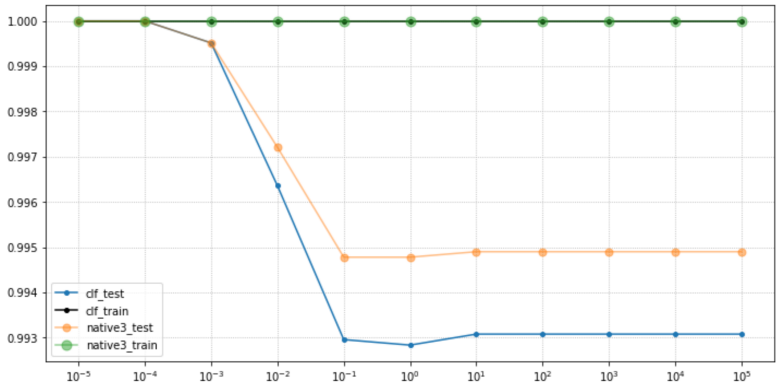
1. Maria Kadukova, Sergei Grudinin. Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization.
2. Sergei Grudinin, Maria Kadukova, Andreas Eisenbarth, Simon Marillet, Frederic Cazals. Predicting binding poses and affinities for protein-ligand complexes in the 2015 D3R Grand Challenge using a physical model with a statistical parameter estimation.
3. Support Vector Machines (CS229 Lecture notes)

Результаты

800 комплексов



(a) MSE



(b) Классификация

Рис.: Регрессия и классификация в зависимости от C_r

Заключение

Предложенный способ имеет высокий потенциал, хотя в данный момент и подвержен переобучению. Тем не менее, даже на неизвестных данных качество классификации в данный момент превышает полученные в базовом решении результаты (от 90 до 94% в зависимости от способа генерации данных).
В будущем можно увеличить количество комплексов в обучающей выборке и оценить работу алгоритма на структурах, сгенерированных более сложными докинг-методами.