Имя журнала

Решение задачи оптимизации, сочетающей классификацию и регрессию, в применении к молекулярному докингу

Биоинформатика

Бишук Антон [1]*, **Кадукова Мария** [2,3]†, **Грудинин Сергей** [2]‡

1 Факультет Управления Прикладной Математики и Информатики, Московский Физико-Технический Институт

2 Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France

3 Центр исследований молекулярных механизмов старения и возрастных заболеваний, Московский Физико-Технический Институт

Аннотация: В работе рассматривается задача предсказания нативной конформации белка и лиганда. Решение задачи происходит путем поиска минимальной свободной энергии связи молекулы белка с лигандом. Для этого оптимизируется скоринг–функция, включающая в себя классификацию, базирующуюся на методе опорных векторов(SVM), и регрессию с различными функциями потерь. Проверка тезисов будет осуществляться на данных, состоящих из комплексов белков и лигандов, для которых необходимо определить наилучшую позу лиганда или предсказать свободную энергию связывания.

MSC: XXXXX, YYYYY

Ключевые слова: Лиганд – белок • скоринговая функция • свободная энергия связывания • классификация • регрессия

1. Введение

Растущая потребность в открытии более эффективных лекарственных средств стимулирует развитие новых подходов в молекулярном моделировании. [1]. Молекулярный докинг является одной важных задач молекулярного моделирования. Он заключается в предсказании взаимной ориентации молекул, наиболее выгодной для образования устойчивого комплекса [2]. Молекулярный докинг – это метод молекулярного моделирования, который позволяет предсказать наиболее выгодную для образования устойчивого комплекса ориентацию молекул. Он порождает огромное пространство признаков за счет большого числа степеней

* *E-mail: bishuk.ayu@phystech.edu*
† *E-mail: mn.kadukova@gmail.com*
‡ *E-mail: sergei.grudinin@gmail.com*

свободы системы. Для решения данной задачи необходимо проводить анализ и обработку большого объема данных, в связи с чем задача является вычислительно сложной, а значит требует решений с высокой производительностью.

Образование комплекса «белок – лиганд» является термодинамическим событием, описываемым константой связывания. Она напрямую зависит от свободной энергией связывания. Так, минимуму энергии связывания соответствует нативная конформация. Свободная энергия связывания зависит от множества факторов, строгий подсчет которых требовал бы сэмплирования всего конфигурационного пространства, что в свою очередь является вычислительно сложной задачей в силу высокой размерности пространства[3]. Для решения задачи нахождения минимальной энергии в последние годы был предложен ряд аппроксимирующих алгоритмов, базирующихся на скоринговых функциях[4] для оценки энергии связывания белка и лиганда.

Convex-PL обучается через задачу классификации, суть которой в том, чтобы разделить нативную и ненативную конформацию системы, состоящей из белка и лиганда. Энергия, предсказанная Convex-PL, должна быть минимальна для нативной конформации.

Одним из таких методов является Convex-PL[3]. Он классифицирует конформации системы, состоящей из белка и лиганда, на нативные и ненативные. Происходит это в том предположении, что нативной конформации соответствует минимальная энергия связывания. Основная идея метода Convex-PL состоит в представлении белка и лиганда в виде конечного набора точек и последующего подсчета функционала, которым в данном случае является свободная энергия связывания, на всевозможных комбинациях различных пар атомов. Однако этот метод имеет недостаток в лице недостаточной корреляции между реальной энергией связывания и полученной из решения оптимизационной задачи [5]. Для устранения этого недостатка предлагается использовать метод, совмещающий в себе как бинарную классификацию, так и регрессию.

2. Модель взаимодействий

Пусть имеется P нативных комплексов белков-лигандов $\{C_{i1}\}_{i=1}^{P}$. Применив к лигандам изометрические преобразования, сгенерируем для каждой нативной позы $D-1$ ненативных поз $\{C_{ij}\}_{i=1,j=2}^{\mathrm{P,D}}$. Таким образом, для каждого из P комплексов имеем D конформаций: одну нативную и $D-1$ ненативных. Требуется найти скоринговый функционал E, который удовлетворяет следующим неравенствам:

$$E(C_{i1}) < E(C_{ij})\ \forall i \in \{1,\dots,P\},\ \forall j \in \{2,\dots,D\}. \tag{1}$$

В качестве такого функционала будем рассматривать свободную энергию связывания белков с лигандами, определенную для всех возможных комплексов. Сделаем ряд допущений для упрощения формы функционала.

Во-первых, будем рассматривать комплекс «белок-лиганд» как набор атомов, каждый из которых имеет

некоторый тип. Тип каждого атома зависит от его химических свойств, таких как номер элемента в периодической таблице, принадлежность к ароматической функциональной группе, гибридизация, полярность и т.д. Пусть M_1 – число типов атомов белка, а M_2 – число типов атомов лиганда. Тогда всего получим $M_1 \times M_2$ различных атомных взаимодействий.

Во-вторых, будем считать, что скоринговый функционал E определяется только взаимодействиями между парами атомов рассматриваемого комплекса. При этом в каждой паре первый атом является атомом лиганда, а второй – атомом белка. Кроме того, будем рассматривать только те пары, в которых расстояние между атомами не превышает некоторой пороговой величины $r_{\max}$. В качестве $r_{\max}$ возьмем значение 7Å, что отличается от значения в 10Å, как это было сделано в других работах [6–11]. Такое решение было принято для уменьшения размерности данных, однако в дальнейшем, эту условность можно опустить, тогда следует ожидается улучшение результатов.

В-третьих, будем считать, что E зависит только от распределения расстояний между взаимодействующими атомами.

И, наконец, предположим, что E является линейным функционалом и имеет следующий вид:

$$E(n(r)) = \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} \int_0^{r_{max}} n^{kl}(r) f^{kl}(r) dr, \tag{2}$$

здесь $f^{kl}(r)$ – неизвестные функции взаимодействия между атомами типов k и l. Будем называть их *скоринговыми потенциалами*,а функции $n^{kl}(r)$ – численными плотностями распределений пар атомов типов k и l по расстоянию r между ними:

$$n^{kl}(r) = \sum_{i,j} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(r - r_{ij})^2}{2\sigma^2}\right], \tag{3}$$

где σ^2 – стандартное отклонение (константа). Сумма берется по всем парам (i, j) атомов с типами k и l соответственно, у которых расстояние между атомами не превышает пороговой величины $r_{\max}$, атом i принадлежит лиганду, а атом j – белку.

Разложим неизвестные скоринговые потенциалы $f^{kl}(r)$ и плотности $n^{kl}(r)$ по полиномиальному базису:

$$\begin{aligned} f^{kl}(r) &= \sum_q w_q^{kl} \psi_q(r), \\ n^{kl}(r) &= \sum_q x_q^{kl} \psi_q(r), \end{aligned} \tag{4}$$

где $\psi_q(r)$ – ортогональные базисные функции на интервале $[0, r_{\max}]$, а w_q^{kl} и x_q^{kl} – коэффициенты разложения функций $f^{kl}(r)$ и $n^{kl}(r)$ соответственно. Поскольку базисные функции ортогональны, справедливо следующее соотношение:

$$\int_0^{r_{\max}} \psi_i(r)\psi_j(r)\Omega(r)dr = \delta_{ij}, \ r \in [0, r_{\max}], \tag{5}$$

где $\Omega(r)$ – некоторая неотрицательная весовая функция на $[0, r_{\max}]$, δ_{ij} – символ Кронекера. Из данного условия (5) ортогональности базисных функций могут быть найдены коэффициенты раложения w_q^{kl} и x_q^{kl}:

$$\begin{aligned} w_q^{kl} &= \int_0^{r_{\max}} f_{kl}(r)\psi_q(r)dr, \\ x_q^{kl} &= \int_0^{r_{\max}} n_{kl}(r)\psi_q(r)dr, \end{aligned} \tag{6}$$

Таким образом, функционал E можно записать в следующем виде:

$$E(n(r)) = \sum_{k=1}^{M_1}\sum_{l=1}^{M_2}\sum_{pq}^{\infty} w_q^{kl} x_p^{kl} \int_0^{r_{\max}} \psi_q(r)\psi_p(r)\Omega(r)dr. \tag{7}$$

Учитывая условие ортогональности (5) и ограничиваясь порядком разложения Q, получаем приближенное выражение скорингового функционала:

$$\begin{gathered} E(n(r)) \approx \sum_{k=1}^{M_1}\sum_{l=1}^{M_2}\sum_{q=0}^{Q} w_q^{kl} x_q^{kl} = \langle \mathbf{w}, \mathbf{x} \rangle, \\ \mathbf{w}, \mathbf{x} \in \mathbf{R}^{Q \times M_1 \times M_2}. \end{gathered} \tag{8}$$

Неизвестный вектор $\mathbf{w}$ будем называть *скоринговым вектором*, а вектор $\mathbf{x}$ – *структурным вектором*. Структурный вектор можно получить из кристаллографических данных. Для оценки энергии связывания в работе [12] использовался порядок разложения $Q = 10$, и рассматривались $M_1 = 23$ и $M_2 = 40$ типов атомов.

3. Постановка задачи

С точки зрения биоинформатики, задача заключается в оценке свободной энергии связывания белка с маленькой молекулой (лигандом): лиганд, наиболее подходящий для связи с белком, в своем наилучшем положении имеет **наименьшую свободную энергию** взаимодействия с молекулой белка. Для нахождения такого лиганда и его энергии связывания будем решать две задачи: бинарную классификацию (SVM) и регрессию.

Заранее обозначим $\mathbf{X}_i$ – **i**-ая строка матрицы признаков $\mathbf{X}$, матрица $\overset{nat}{\mathbf{X}}$ – матрица признаков для комплексов в натиных конформациях. Важно отметить, что для каждого комплекса существует единственная «нативная» поза с меткой $y = 1$ и D − 1 «ненативная» с меткой $y = -1$.

3.1. Классификация поз на нативные и ненативные

Для поиска вектора $\mathbf{w}$, имеющего значение энергии взаимодействия, решается следующая задача классификации:

$$\begin{aligned} \underset{\mathbf{w},b_i,\xi_i}{\text{Minimize:}} \quad & \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i \\ & y_i[\mathbf{w}^T\mathbf{X}_i - b_j] - 1 + \xi_i \geq 0 \\ \text{Subject to:} \quad & \xi_i \geq 0, \end{aligned} \tag{9}$$

где: $\mathbf{w}$, b_j – переменные невязки; ξ_i – оптимизируемые параметры модели; C – некоторый коэффициент регуляризации; P – число комплексов белок-лиганд; D – число взаимных положений (поз) белка и лиганда в каждом из комплексов; $j = 1, \ldots, \mathrm{P}$; $i = 1, \ldots, \mathrm{P} \cdot \mathrm{D}$.

3.2. Предсказание энергии связывания

Для каждого комплекса известно значение s_i, соответствующее энергии связывания i-ой нативной конформации, полученной в эксперименте. Его также можно предсказывать, используя для обучения матрицу признаков $\overset{nat}{\mathbf{X}}$ нативных поз. Тем самым получаем задачу регрессии предсказания энергии связывания:

$$\underset{\mathbf{w}}{\text{Minimize:}} \quad \sum_i \|\mathbf{w}^T\overset{nat}{\mathbf{X}}_i - s_i\|^2 + \alpha\|\mathbf{w}\|^2, \tag{10}$$

где: α – коэффициент регуляризации ridge-регрессии.

3.3. Итоговая задача

Таким образом, после объединение задач **??**SVM) и **??**Reg) требуется решить следующую оптимизационную задачу:

$$\begin{aligned} \underset{\mathbf{w},b_i,\xi_i}{\text{Minimize:}} \quad & \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i + C_r\sum_i f(\overset{nat}{\mathbf{X}}_i, \mathbf{w}, s_i) \\ & y_i[\mathbf{w}^T\mathbf{X}_i - b_j] - 1 + \xi_i \geq 0 \\ \text{Subject to:} \quad & \xi_i \geq 0 \end{aligned} \tag{11}$$

Здесь $f(\overset{nat}{\mathbf{X}}_i, \mathbf{w}, s_i)$ – функция потерь регрессии, например, *Mean Squared Error*, C_r – коэффициент регуляризации для функции потерь регрессии.

4. Теоретическое обоснование

В уравнении 11 будем принимать за функцию потерь регресии MSE:

$$f\left(\overset{nat}{\mathbf{X}}_i, \mathbf{w}, s_i\right) = \left(\langle\overset{nat}{\mathbf{X}}_i, w\rangle - s_i\right)^2, \qquad i = 1, \ldots, P. \tag{12}$$

Тогда проведем ряд преобразований для нативных конформаций:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C_r \sum_{i=1}^{P} f\left(\overset{nat}{\mathbf{X}}_i, \mathbf{w}, s_i\right) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_r \left\|\overset{nat}{\mathbf{X}}\,\mathbf{w} - \mathbf{s}\right\|^2 =$$

$$= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_r\mathbf{w}^T\overset{nat}{\mathbf{X}}^T\overset{nat}{\mathbf{X}}\,\mathbf{w} - 2C_r\mathbf{w}^T\overset{nat}{\mathbf{X}}^T\mathbf{s} + C_r\mathbf{s}^T\mathbf{s} = \frac{1}{2}\mathbf{w}^T\left(\mathbf{I} + 2C_r\overset{nat}{\mathbf{X}}^T\overset{nat}{\mathbf{X}}\right)\mathbf{w} - 2C_r\mathbf{w}^T\overset{nat}{\mathbf{X}}^T\mathbf{s} + C_r\mathbf{s}^T\mathbf{s} =$$

$$= \frac{1}{2}\left\|\left[\mathbf{I} + 2C_r\overset{nat}{\mathbf{X}}^T\overset{nat}{\mathbf{X}}\right]^{\frac{1}{2}}\mathbf{w}\right\|^2 - 2C_r\left(\mathbf{w}^T\left[\mathbf{I} + 2C_r\overset{nat}{\mathbf{X}}^T\overset{nat}{\mathbf{X}}\right]^{\frac{1}{2}}\right)\left(\left[\mathbf{I} + 2C_r\overset{nat}{\mathbf{X}}^T\overset{nat}{\mathbf{X}}\right]^{-\frac{1}{2}}\overset{nat}{\mathbf{X}}^T\mathbf{s}\right) + C_r\mathbf{s}^T\mathbf{s} =$$

$$= \frac{1}{2}\left(\left[\mathbf{I} + 2C_r\overset{nat}{\mathbf{X}}^T\overset{nat}{\mathbf{X}}\right]^{\frac{1}{2}}\mathbf{w} - 2C_r\left[\mathbf{I} + 2C_r\overset{nat}{\mathbf{X}}^T\overset{nat}{\mathbf{X}}\right]^{-\frac{1}{2}}\overset{nat}{\mathbf{X}}^T\mathbf{s}\right)^2 + C_r\mathbf{s}^T\mathbf{s} - 2C_r^2\left\|\left[\mathbf{I} + 2C_r\overset{nat}{\mathbf{X}}^T\overset{nat}{\mathbf{X}}\right]^{-\frac{1}{2}}\overset{nat}{\mathbf{X}}^T\mathbf{s}\right\|^2.$$

Введем следующую замену:

$$\begin{aligned}
&\mathbf{A} = \left[\mathbf{I} + 2C_r\overset{nat}{\mathbf{X}}^T\overset{nat}{\mathbf{X}}\right]^{\frac{1}{2}} \\
&\mathbf{d} = 2C_r\left[\mathbf{I} + 2C_r\overset{nat}{\mathbf{X}}^T\overset{nat}{\mathbf{X}}\right]^{-\frac{1}{2}}\overset{nat}{\mathbf{X}}^T\mathbf{s} = 2C_rA^{-1}\overset{nat}{\mathbf{X}}^T\mathbf{s} \\
&\mathbf{w}' = \mathbf{A}\mathbf{w} - \mathbf{d}
\end{aligned} \tag{13}$$

Величина, стоящая вне квадрата является константой, а потому не влияет на оптимизацию:

$$C_r\mathbf{s}^T\mathbf{s} - 2C_r^2\left\|\left[\mathbf{I} + 2C_r\overset{nat}{\mathbf{X}}^T\overset{nat}{\mathbf{X}}\right]^{-\frac{1}{2}}\overset{nat}{\mathbf{X}}^T\mathbf{s}\right\|^2 = \text{const}$$

После замены получим следующую задачу оптимизации:

$$\begin{aligned}
&\underset{\mathbf{w},b_i,\xi_i}{\text{Minimize:}} && \tfrac{1}{2}\|\mathbf{w}'\|^2 + C\sum_i \xi_i \\
& && y_i[(\mathbf{A}^{-1}(\mathbf{w}' + \mathbf{d}))^T\mathbf{X}_i - b_j] - 1 + \xi_i \geq 0 \\
&\text{Subject to:} && \xi_i \geq 0
\end{aligned} \tag{14}$$

Важно отметить, что $\mathbf{X}_i$ – i-ую строку матрицы $\mathbf{X}$ – в терминах линейной алгебры следует воспринимать как столбец, а не как строку.

Построим двойственную задачу:

$$\mathcal{L}(\mathbf{w}', \mathbf{d}, \xi, \lambda, r) = \frac{1}{2}\|\mathbf{w}'\|^2 + C\sum_{i=1}^{\mathrm{P\cdot D}}\xi_i - \sum_{\substack{j=0\\ j=nD}}^{\mathrm{D(P-1)}}\sum_{i=1}^{\mathrm{D}}\lambda_i\left(y_i[\langle\mathbf{A}^{-1}\left(\mathbf{w}'+\mathbf{d}\right), \mathbf{X}_i\rangle - b_j] - 1 + \xi_i\right) - \sum_{i=1}^{\mathrm{P\cdot D}} r_i\xi_i,$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}'} = \mathbf{w}' - \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i y_i\langle\mathbf{A}^{-1}\left(\mathbf{w}'+\mathbf{d}\right), \mathbf{X}_i\rangle'_{w'} = 0 \to \underline{\mathbf{w}' = \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i y_i\mathbf{A}^{-1}\mathbf{X}_i}$$

$$\frac{\partial\mathcal{L}}{\partial b_k} = \underline{\sum_{j=1}^{D}\lambda_{k+j}y_{k+j} = 0}, \qquad k \in \{0, \mathrm{D}, 2\mathrm{D}, \ldots, \mathrm{D(P-1)}\}, \tag{15}$$

$$\frac{\partial\mathcal{L}}{\partial\xi_i} = C - \lambda_i - r_i = 0 \to \underline{\lambda_i + r_i = C}, \qquad \forall i \in \{1, \ldots \mathrm{P\cdot D}\}.$$

Подставим оптимальные значения параметров в лагранжиан.

$$\begin{aligned}
&\mathcal{L}(\lambda, r) = \frac{1}{2}\langle\sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i y_i\mathbf{A}^{-1}\mathbf{X}_i, \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i y_i\mathbf{A}^{-1}\mathbf{X}_i\rangle + C\sum_{i=1}^{\mathrm{P\cdot D}}\xi_i - \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i y_i\langle\mathbf{A}^{-1}\sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i y_i\mathbf{A}^{-1}\mathbf{X}_i, \mathbf{X}_i\rangle - \\
&- \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i y_i\langle\mathbf{A}^{-1}\mathbf{d}, \mathbf{X}_i\rangle + \sum_{\substack{j=0\\ j=nD}}^{\mathrm{D(P-1)}}\sum_{i=1}^{\mathrm{D}}\lambda_{i+j}y_{i+j}b_j + \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i - \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i\xi_i - \sum_{i=1}^{\mathrm{P\cdot D}} r_i\xi_i = \\
&= \frac{1}{2}\sum_{i=1}^{\mathrm{P\cdot D}}\sum_{j=1}^{\mathrm{P\cdot D}}\lambda_i\lambda_j y_i y_j\langle\mathbf{A}^{-1}\mathbf{X}_i, \mathbf{A}^{-1}\mathbf{X}_j\rangle + \sum_{i=1}^{\mathrm{P\cdot D}}\xi_i(C - \lambda_i - r_i) - \sum_{i=1}^{\mathrm{P\cdot D}}\sum_{j=1}^{\mathrm{P\cdot D}}\lambda_i\lambda_j y_i y_j\langle\mathbf{A}^{-2}\mathbf{X}_i, \mathbf{X}_j\rangle - \\
&- \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i y_i\langle\mathbf{A}^{-1}\mathbf{d}, \mathbf{X}_i\rangle + \sum_{\substack{j=0\\ j=nD}}^{\mathrm{D(P-1)}} b_j\sum_{i=1}^{\mathrm{D}}\lambda_{i+j}y_{i+j} + \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i
\end{aligned} \tag{16}$$

Учитывая следующие замечания, получим выражение для лагранжиана и ограничений в задаче условной минимизации:

1. $\sum\limits_{i=1}^{\mathrm{P\cdot D}}\xi_i(C - \lambda_i - r_i) = 0$ из ограничений;
2. $\sum\limits_{i=1}^{P} b_i\sum\limits_{j=1}^{D}\lambda_{i+j}y_{i+j} = 0$ из ограничений;
3. $\langle\mathbf{A}^{-2}\mathbf{X}_i, \mathbf{X}_i\rangle = \langle\mathbf{A}^{-1}\mathbf{X}_i, \mathbf{A}^{-1}\mathbf{X}_i\rangle$, т.к. $\mathbf{X}_i^\top\mathbf{A}^{-2}\mathbf{X}_i = \mathbf{X}_i^\top\mathbf{A}^{-1}\mathbf{A}^{-1}\mathbf{X}_i = (\mathbf{A}^{-1}\mathbf{X}_i)^\top\mathbf{A}^{-1}\mathbf{X}_i$;
4. $\widehat{\mathbf{X}} \stackrel{\mathrm{def}}{=} \mathbf{X}\mathbf{A}^{-1}$.

$$\begin{aligned}
\mathcal{L}(\lambda) &= -\frac{1}{2}\sum_{i=1}^{\mathrm{P\cdot D}}\sum_{j=1}^{\mathrm{P\cdot D}}\lambda_i\lambda_j y_i y_j\langle\widehat{\mathbf{X}}_i, \widehat{\mathbf{X}}_j\rangle - \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i y_i\langle\mathbf{A}^{-1}\mathbf{d}, \mathbf{X}_i\rangle + \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i = \\
&= -\frac{1}{2}\sum_{i=1}^{\mathrm{P\cdot D}}\sum_{j=1}^{\mathrm{P\cdot D}}\lambda_i\lambda_j y_i y_j\langle\widehat{\mathbf{X}}_i, \widehat{\mathbf{X}}_j\rangle + \sum_{i=1}^{\mathrm{P\cdot D}}\lambda_i\left(1 - y_i\langle\mathbf{d}, \widehat{\mathbf{X}}_i\rangle\right)
\end{aligned} \tag{17}$$

Двойственной задачей является $\underset{\lambda}{\operatorname{argmax}}\,\mathcal{L}(\lambda)$.

Значит, исходная задача по теореме Каруша-Куна-Таккера эквивалентна следующей двойственной:

$$\begin{aligned} \underset{\lambda_i}{\text{minimize:}} \quad & \frac{1}{2}\sum_{(i,j)=1}^{\mathrm{P\cdot D}} \lambda_i\lambda_j y_i y_j \langle \widehat{\mathbf{X}}_i, \widehat{\mathbf{X}}_j\rangle + \sum_{i=1}^{\mathrm{P\cdot D}} \lambda_i \left(y_i\langle \mathbf{d}, \widehat{\mathbf{X}}_i\rangle - 1\right) \\ \text{subject to:} \quad & 0 \le \lambda_i \le C, \\ & \sum_{j=1}^{D} \lambda_{k+j} y_{k+j} = 0, \\ & k \in \{0, \mathrm{D}, 2\mathrm{D}, \ldots, \mathrm{D}(\mathrm{P}-1)\}. \end{aligned} \tag{18}$$

5. Данные

Данные представляют из себя размеченную матрицу признаков для 12,000 комплексов белков с лигандами, для каждого из которых известна одна нативная и 18 ненативных конформаций (декоев). Декои были сгенерированы путем поворотов и трансляций относительно нативной конформации с сохранением постоянного RMSD, равного 1 Å . Матрица признаков записана так, что кластеры из конформаций одного комплекса следуют друг за другом. В кластере всегда на первом месте стоит нативное положение, затем все ненативные.

Данные для бинарной классификации

Основными признаками объектов являются гистограммы распределений расстояний между парами атомов белка и лиганда различных типов. Размерность вектора признаков составляет 6,440. Так же в данных есть вектор меток $\mathbf{y}$, у которого на месте нативной конформации стоит единица, а на месте ненативной конформации -1. Метки были присвоены конформациям в ходе экспериментов.

Данные для регрессии Для каждого из представленных комплексов известно значение величины, которую можно интерпретировать как энергию связывания. Это значение определено только для нативной позы лиганда в комплексе. Для задачи регрессии мы используем 379 дополнительных признаков для учета взаимодействий комплексов с растворителем и учета энтропийных потерь из-за ограничения подвижности лиганда. 65 из 379 признаков соответствуют площади поверхности, доступной растворителю, для каждого типа атома взаимодействующих атомов белка и лиганда. Следующие 315 признаков являются гистограммами распределений расстояний между взаимодействующими атомами белка, лиганда и пробными "атомами растворителя которые мы расставляем по сетке вокруг комплекса. Энтропийные потери описываются "подвижностью"лиганда, рассчитанной как логарифм числа его стабильных конформаций.

6. Эксперимент

Для решения оптимизацонной задачи было решено использовать Оксфордский солвер OSQP[13, 14], базирующийся на алгоритме ADMM[15].

Первым этапом работы стала проверка того, что численное решение задачи при $C_r = 0$ соответствует решению, полученному для задачи SVM. Это условие могло не выполняться из-за накопления численных ошибок, вызванных неточностью операций над числами с плавающей точкой при поиске квадратного корня и обратной матрицы в (13) и (14).

Далее важно было понять, действительно ли дополнительные признаки, которые были добавлены для задачи регрессии, улучшают корреляцию предсказанных энергий с экспериментальными значениями. Чтобы разделить проверку гипотез, было решено использовать такой набор данных, который не ухудшал классификацию, то при этом увеличивал корреляцию энергии. Поэтому в первую очередь было решено добавить всего один признак, который имеет одно значение для всех положений в комплексе. Таким образом влияние новых полученных данных на классификацию должно стать минимальным. Таким образом, мы произвели вычисления для матрицы, состоящей только из признаков для классификации с добавлением признака, описывающего подвижность(flexibility). Добавление flexibility привело к улучшению корреляции.

Так же было замечено, что классификация не будет ухудшаться, если будет выполнено следующее условие: $\frac{C_r}{C} \gg 1$.

Затем необходимо было подобрать параметры модели и проверить ее работу на кросс-валидации и тестовых данных.

TODO1: Провести обработку данных, связанных с регрессии(скалирование, PCA, корреляции), что позволит убрать признаки, не подходящие с физической точки зрения.

TODO2: Проверить, что корреляция в Convex-PL хуже, чем в этом алгоритме. для этого необходимо найти лучшее предсказание нашего алгоритма.

TODO3: Найти подходящее скалирование для параметров C и C_r.

7. Вывод

Ожидается, что модель позволит получать энергию связывания, которая в равной степени хорошо предсказывает нативное положение и коррелирует с реальной энергией связывания.

Список литературы

[1] Raimund Mannhold, Hugo Kubinyi, and Gerd Folkers. *Virtual screening: principles, challenges, and practical guidelines*, volume 48. John Wiley & Sons, 2011.

[2] Thomas Lengauer and Matthias Rarey. Computational methods for biomolecular docking. *Current opinion in structural biology*, 6(3):402–406, 1996.

[3] Maria Kadukova and Sergei Grudinin. Convex-pl: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization. *Journal of computer-aided molecular design*, 31(10):943–958, 2017.

[4] Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157, 2011.

[5] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[6] Sheng-You Huang and Xiaoqin Zou. An iterative knowledge-based scoring function for protein–protein recognition. *Proteins: Structure, Function, and Bioinformatics*, 72(2):557–579, 2008.

[7] Gwo-Yu Chuang, Dima Kozakov, Ryan Brenke, Stephen R Comeau, and Sandor Vajda. Dars (decoys as the reference state) potentials for protein-protein docking. *Biophysical journal*, 95(9):4217–4227, 2008.

[8] Vladimir N Maiorov and Gordon M Grippen. Contact potential that recognizes the correct folding of globular proteins. 1992.

[9] Jian Qiu and Ron Elber. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins: Structure, Function, and Bioinformatics*, 61(1):44–55, 2005.

[10] Dror Tobi and Ivet Bahar. Optimal design of protein docking potentials: efficiency and limitations. *Proteins: Structure, Function, and Bioinformatics*, 62(4):970–981, 2006.

[11] Myong-Ho Chae, Florian Krull, Stephan Lorenzen, and Ernst-Walter Knapp. Predicting protein complex geometries with a neural network. *Proteins: Structure, Function, and Bioinformatics*, 78(4):1026–1039, 2010.

[12] Sergei Grudinin, Maria Kadukova, Andreas Eisenbarth, Simon Marillet, and 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. *Journal of computer-aided molecular design*, 30(9):791–804, 2016.

[13] Bartolomeo Stellato, Goran Banjac, Paul Goulart, Alberto Bemporad, and Stephen Boyd. OSQP: An operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 2020.

[14] G. Banjac, B. Stellato, N. Moehle, P. Goulart, A. Bemporad, and S. Boyd. Embedded code generation using the OSQP solver. In *IEEE Conference on Decision and Control (CDC)*, 2017.

[15] G. Banjac, P. Goulart, B. Stellato, and S. Boyd. Infeasibility detection in the alternating direction method of

multipliers for convex optimization. *Journal of Optimization Theory and Applications*, 183(2):490–519, 2019.