Имя журнала

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

Биоинформатика

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

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

2 Institute of Mathematics, Research Institution, Street, Postal code, City, Country

3 Institute of Mathematics, Research Institution, Street, Postal code, City, Country

Аннотация: В работе представлено решение задачи минимизации свободной энергии связи молекулы белка с лигандом посредством оптимизации скоринговых функций. Оптимизация включает в себя классификацию, базирующуюся на методе опорных векторов(SVM), и регрессию с различными функциями потерь. Такой комплексный подход к оптимизации обусловлен двумя факторами: первый – это недостаточно высокая корреляция предсказаний и экспериментальных значений при решении классической задачи классификации энергий связывания; второй – это переобучение регрессионных моделей при решении задачи нахождения оптимальной энергии. В данной работе представлено построение модели, сочетающей в себе классификацию и регрессию, тем самым решая проблемы отдельных подходов. 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» • Скоринг функции • Свободная энергия связывания • Классификация • Регрессия • Лиганд

1. Введение

Растущая потребность в открытии более эффективных лекарственных средств, стимулирует развитие новых подходов в молекулярном моделировании, что открывает широкие горизонты в области изучения химических соединений. Доказательством этому может служит метод виртуального скрининга, который сегодня активно используется для поиска и анализа химических соединений, обладающих рядом необходимых свойств[1].

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

Одной из основных задач молекулярного моделирования является молекулярный докинг, заключающийся в предсказании взаимной ориентации молекул, наиболее выгодной для образования устойчивого комплекса[2]. 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, описываемое постоянной степени сродства соединения, которая напрямую связана со свободной энергией связывания. Так минимуму энергии связывания соответствует нативная конформация. Свободная энергия связывания зависит от множества факторов, строгий подсчет которых требовал бы семплирования всего конфигурационного пространства, что в свою очередь является вычислительно сложной задачей в силу высокой размерности пространства[3]. Для решения данной задачи, в последние годы, был предложен ряд аппроксимирующих алгоритмов, базирующихся на скоринговых функциях[4] для оценки энергии связывания.

Одной из таких функций является Convex-PL[3], которая дает минимум на нативной конформации системы, состоящей из белка и лиганда. Основная идея построения Convex-PL состоит в представлении белка и лиганда в виде конечного набора точек и последующего подсчета функционала на всевозможных комбинациях различных пар атомов. Однако эта функция имеет недостаток в лице недостаточной корреляции между энергией связывания и полученных результатов, которые будут решены при помощи функций потерь регрессии. Для этого нужно решить ряд задач оптимизации[5].

2. Модель взаимодействий

2.1. Subsections

2.1.1. Subsubsections

3. Постановка задачи

С точки зрения биоинформатики, задача заключается в оценке свободной энергии связывания белка с маленькой молекулой (лигандом): наилучший лиганд в своем наилучшем положении имеет **наименьшую свободную энергию** взаимодействия с белком. Для этого необходимо решить две задачи: бинарную классификацию (SVM) и регрессию.

3.1. Бинарная классификация

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

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

Где:

- $i = 1, \ldots, \mathrm{P} \cdot \mathrm{D}$;
- P – число комплексов белок-лиганд;
- D – число используемых для обучения взаимных положений (поз) белка и лиганда в каждом из комплексов;
- $\mathbf{X}$ – матрица признаков;

! Для каждого комплекса существует одна «нативная» поза с $y = 1$ и $\mathrm{D} - 1$ «ненативная» с $y = -1$.

3.2. Регрессия

Для каждого комплекса известно значение s_i, несущее смысл энергии связывания, полученной в эксперименте. Его также можно предсказывать, используя для обучения матрицу признаков $\overset{\bullet}{\mathbf{X}}$ нативных поз.

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

3.3. Итоговая задача

Требуется решить следующую оптимизационную задачу:

$$\begin{aligned} \underset{\mathbf{w},b_i,\xi_i}{\text{Minimize:}} \quad & \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_i C_i\xi_i + C_r\sum_i f(\overset{\bullet}{\mathbf{X}}_i, \mathbf{w}, s_i) \\ & y_i[\mathbf{w}^T\mathbf{x}_i - b_i] - 1 + \xi_i \geq 0 \\ \text{Subject to:} \quad & \xi_i \geq 0 \end{aligned} \tag{3}$$

Здесь $f(\overset{\bullet}{\mathbf{X}}_i, \mathbf{w}, s_i)$ – функция потерь регрессии, например, *Mean Squared Error* или *Mean Absolute Error*.

3.4. Данные

Данные для бинарной классификации:

Около 12,000 комплексов белков с лигандами: для каждого из них есть 1 нативная поза и 18 ненативных. Основными признаками объектов являются гистограммы распределений расстояний между различными атомами белка и лиганда, размерность вектора признаков порядка 20,000.

Данные для регрессии:

Для каждого из представленных комплексов известно значение величины, которую можно интерпретировать как энергию связывания.

4. Эксперимент

5. Вывод

6. Tables and figures

Таблица 1. **Some caption text.**

Some title			
row 1, column 1	row 1, column 2		
row 2, column 1	row 2, column 2		
row 3, column 1	row 3, column 2		
Another title	Value 1	Value 2	Value 3
row 1	130	30	30
row 2	1025	1	15
row 3	100	1	10
row 4	2925	1	4
row 5	2950	1	2

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.