

Использование нейросетевого алгоритма для предсказания аффинности связывания в белок-белковых комплексах ACE2-RBD

Богданова Е.^{1*}, Чернухин А.², Шайтан К.¹, Новоселецкий В.³

¹ Московский государственный университет им. М.В. Ломоносова, Москва, Россия

² Российский химико-технологический университет им. Д.И. Менделеева, Москва, Россия

³ Университет МГУ-ППИ в Шэньчжэне, биологический факультет, Шэньчжэнь, Китай

* eabogdanova.bioinf@gmail.com

Ключевые слова: белок-белковые комплексы; аффинность связывания; коронавирус; нейронные сети

Мотивация и цель: Широкое распространение SARS-CoV-2 по сравнению с его предшественниками обусловлено более высокой аффинностью связывания S-белка с ACE2 [1]. Осознание данной закономерности повлекло за собой большое количество исследований интерфейса связывания RBD-ACE2, а также аминокислотных замен, влияющих на аффинность связывания между данными молекулами [2]. Физическое взаимодействие между молекулами белков имеет давнюю историю изучения с использованием многочисленных экспериментальных и вычислительных методов [3, 4], в том числе методов биоинформатики [5]. Одной из главных характеристик взаимодействия является константа диссоциации комплексов белок-белок (Kd). Создание алгоритмов аккуратной оценки этой энергии позволило бы более эффективно проводить направленный мутагенез взаимодействующих белков, что имеет существенное значение для создания медицинских препаратов белковой природы, включая антитела. В данной работе мы провели тестирование ряда ранее предложенных алгоритмов оценки энергии взаимодействия в комплексах «белок-белок» (Prodigy [6], FoldX [7], DFIRE [8], ROSETTADOCK [9]), а также разработанного нами алгоритма ProBAN [10].

Методы и алгоритмы: Для анализа аффинности связывания были отобраны комплексы RBD-ACE2 с известными пространственными структурами и значениями константы диссоциации (Kd). Итоговый набор данных содержит 48 комплексов ACE2 и RBD вирусов SARS CoV (3 структуры) и SARS CoV-2 (45 структур).

Предсказание аффинности связывания осуществлялось алгоритмом ProBAN, основанным на сверточной нейронной сети. Модель была обучена и протестирована на наборе белок-белковых комплексов с известными значениями константы диссоциации в базе данных PDBBind v.2020. Исследуемые комплексы RBD-ACE2 не содержатся в данной версии базы данных, что позволяет проводить предсказания на обученной модели без возможности «утечки» данных из обучающей выборки. Анализ полученных результатов, а также предсказание алгоритмом ProBAN осуществлялось с использованием Python дистрибутива Anaconda.

Результаты: Для анализа предсказания характеристик связывания в белок-белковых комплексах алгоритмом ProBAN было проведено его сравнение с предсказаниями, полученными веб-сервисом Prodigy. Данный метод осуществляет оценку аффинности связывания функцией, основанной на межмолекулярных контактах и признаках непосредственно на интерфейсе и полученных из анализа поверхности, не относящейся к интерфейсу взаимодействия. ProBAN (MAE = 0.5, корреляция Пирсона = 0.56) показывает более высокое качество предсказания по сравнению с Prodigy (MAE = 0.9, корреляция Пирсона = -0.38). Мы полагаем, что причиной этого является использование как более полной информации о взаимодействиях между атомами, так и большего порогового значения расстояния между атомами (10 Å), которое классифицирует пары атомов на взаимодействующие и нет. Используемое в Prodigy аналогичное пороговое значение расстояния между атомами (5.5 Å), по-видимому, отсеивает часть важных атомов, вносящих вклад в связывание.

Большое число алгоритмов, используемых для оценки аффинности связывания в белок-белковых комплексах, предсказывают не значение константы диссоциации, а свободную энергию Гиббса связывания. Для оценки работы этих алгоритмов (FoldX, DFIRE, ROSETTADOCK) на исследуемом наборе данных из полученных нами значений K_d мы рассчитали ΔG и сравнили с ΔG , полученной с использованием перечисленных алгоритмов. Рассчитанные значения метрик качества свидетельствуют о наибольшей эффективности ProBAN (MAE = 0.7 ккал/моль) среди всех проанализированных алгоритмов (для Prodigy 1.2 ккал/моль, у остальных выше).

Среди с комплексов с наибольшей абсолютной ошибкой (больше 1) подавляющая часть имела разрешение хуже 3 Å (6cs2, 7wk6, 7tex), и один имел разрешение 2.85 Å (8asy). Наблюдаемая закономерность свидетельствует о негативном вкладе нечетко разрешенного положения атомов в качество предсказания аффинности связывания, так как искажается информация об межатомных расстояниях, играющих ключевую роль во взаимодействии между белковыми молекулами.

Выводы: Собран набор данных, содержащий структуры комплексов ACE2-RBD с известными константами связывания, который может использоваться для изучения факторов, влияющих на связывание между RBD и ACE2, а также для обучения и тестирования алгоритмов, осуществляющих предсказание аффинности связывания. Использование нейросетевых методов, в частности ProBAN, для оценки константы диссоциации и свободной энергии Гиббса возможно не только между нативными RBD и ACE2, но и между другими мутантными формами. Полученные метрики качества свидетельствуют о недостаточной стабильности предсказаний для комплексов с высокой аффинностью связывания, что указывает на необходимость улучшения предикторов при увеличении набора доступных данных для анализа. Собранный набор данных, а также результаты предсказания выложены в онлайн-репозитории, доступном по ссылке: https://github.com/EABogdanova/ProBAN_RBD-ACE2.

Финансирование: Исследование поддержано Российским научным фондом (№ 19-74-30003).

Using a neural network algorithm to predict binding affinity in ACE2-RBD protein-protein complexes

Bogdanova E.^{1*}, Chernukhin A.², Shaitan K.¹, Novoseletsky V.³

¹ Lomonosov Moscow State University, Moscow, Russia

² Mendeleev University of Chemical Technology of Russia, Moscow, Russia

³ Shenzhen MSU-BIT University, Shenzhen, China

* eabogdanova.bioinf@gmail.com

Key words: protein-protein complexes; binding affinity; coronavirus; neural networks

Motivation and Aim: The widespread distribution of SARS-CoV-2 compared to its predecessors was associated with the higher binding affinity of the S protein to ACE2 [1]. Awareness of this pattern led to many studies of the RBD-ACE2 binding interface, as well as amino acid substitutions that affect the binding affinity between these molecules [2]. The physical interaction between protein molecules has a long history of being studied using numerous experimental and computational methods [3, 4], including bioinformatics methods [5]. One of the main characteristics of the interaction is the dissociation constant of protein-protein complexes (K_d). The creation of algorithms for accurately assessing this energy would make it possible to carry out targeted mutagenesis of interacting proteins more effectively, which is essential for the creation of protein-based drugs, including antibodies. In this work, we tested several previously proposed algorithms for estimating the interaction energy in protein-protein complexes (Prodigy [6], FoldX [7], DFIRE [8], ROSETTADOCK [9]), as well as the ProBAN [10] algorithm we developed.

Methods and Algorithms: RBD-ACE2 complexes with known spatial structures and dissociation constant (K_d) values were selected for binding affinity analysis. The resulting data set contains 48 ACE2 and RBD complexes of the SARS-CoV (3 structures) and SARS-CoV-2 (45 structures) viruses.

Binding affinity prediction was performed by the ProBAN algorithm based on a convolutional neural network. This model was trained and tested on a set of protein-protein complexes with known dissociation constant values in the PDBBind v.2020 database. The RBD-ACE2 complexes under study are not contained in this version of the database, which allows predictions to be made on a trained model without the possibility of data “leakage” from the training set. Analysis of the results obtained, as well as prediction by the ProBAN algorithm, was carried out using the Anaconda Python distribution.

Results: To analyze the prediction of binding characteristics in protein-protein complexes by the ProBAN algorithm, it was compared with predictions obtained by the Prodigy web service. This method evaluates binding affinity by a function based on intermolecular contacts and features directly at the interface and obtained from analysis of surfaces not related to the interaction interface. ProBAN (MAE = 0.5, Pearson correlation = 0.56) shows superior prediction performance compared to Prodigy (MAE = 0.9, Pearson correlation = -0.38). We believe that the reason for this is the use of both more complete information about the interactions between atoms and a larger threshold value for the distance between atoms (10 Å), which classifies pairs of atoms

as those that interact or not. The similar interatomic distance threshold used in Prodigy (5.5 Å) appears to screen out some of the important atoms that contribute to binding. Many algorithms used to estimate binding affinities in protein-protein complexes do not predict the value of the dissociation constant, but rather the Gibbs free energy of binding. To evaluate the performance of these algorithms (FoldX, DFIRE, ROSETTADOCK) on the data set under study, from the K_d values we obtained, we calculated ΔG and compared it with ΔG obtained using these algorithms. The calculated values of quality metrics for different algorithms indicate the highest efficiency of ProBAN (MAE = 0.7 kcal/mol) among all analyzed algorithms (for Prodigy 1.2 kcal/mol, for others higher).

Among the c complexes with the largest absolute error (more than 1), the overwhelming majority had a resolution worse than 3 Å (6cs2, 7wk6, 7tex) and one had a resolution of 2.85 Å (8asy). The observed pattern indicates a negative contribution of unclearly resolved atomic positions to the quality of binding affinity prediction, since information about interatomic distances, which play a key role in the interaction between protein molecules, is distorted.

Conclusions: A dataset containing the structures of ACE2-RBD complexes with known binding constants has been collected, which can be used to study factors influencing the binding between RBD and ACE2, as well as to train and test algorithms that predict binding affinities. The use of neural network methods, particularly ProBAN, to estimate the dissociation constant and Gibbs free energy is possible not only between native RBD and ACE2, but also other mutant forms. The obtained quality metrics indicate a lack of stability in predictions for complexes with high binding affinity, indicating the need to improve predictors as the data set available for analysis increases. The collected data set, as well as the prediction results, are posted in an online repository, available at the link: https://github.com/EABogdanova/ProBAN_RBD-ACE2.

Funding: The study is supported by the Russian Science Foundation (No. 19-74-30003).

Список литературы/References

1. Choudhary S., Sreenivasulu K., Mitra P. et al. Role of Genetic Variants and Gene Expression in the Susceptibility and Severity of COVID-19. *Ann Lab Med*. 2021;41(2):129-138
2. Vogel M., Augusto G., Chang X. et al. Molecular definition of severe acute respiratory syndrome coronavirus 2 receptor-binding domain. *Allergy*. 2022;77(1):143-149
3. Chothia C., Janin J. Principles of protein – protein recognition. *Nature*. 1975;256:705-708
4. Archakov I., Govorun V.M., Dubanov A.V. et al. Protein-protein interactions as a target for drugs in proteomics. *Proteomics*. 2003;3(4):380-391
5. Shi T.L., Li Y.X., Cai Y.D. et al. Computational Methods for Protein – Protein Interaction and Their Application. *Curr Protein Pept Sci*. 2005;6(5):443-449
6. Xue L.C., Rodrigues J.P., Kastiris P.L. et al. PRODIGY: a web server for predicting the binding affinity of protein-protein complexes. *Bioinformatics*. 2016;32(23):3676-3678
7. Schymkowitz J., Borg J., Stricher F. et al. The FoldX web server: an online force field. *Nucleic Acids Res*. 2005;33:W382-W388
8. Zhang C., Liu S., Zhou Y. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. *Protein Sci*. 2004;13(2):391-399
9. Lyskov S., Gray J.J. The RosettaDock Server for Local Protein-Protein Docking. *Nucleic Acids Res*. 2008;36:W233-W238
10. Bogdanova E.A., Novoseletsky V.N. ProBAN: Neural network algorithm for predicting binding affinity in protein–protein complexes. *Proteins*. 2024. doi 10.1002/prot.26700