Задачи оптимизации, сочетающей классификацию и регрессию, для оценки энергии связывания белка и маленьких молекул

Noskova E. S., Kachkov S. S., Sidorenko A. A.

Задача оценки качества белка актуальна на сегодняшний день благодаря своей применимости в прогнозировании структуры белка - фундаментальной и все же открытой проблемы в структурной биоинформатике. Данная статья посвящена расширению метода оценки качества белка(SBROD), использующему для своих вычислений только форму белкового скелета и переобучение его на более надежных метриках CAD и LDDT. А также улучшению метода посредством изменения функции потерь, используя Z-score.

Введение

Белки играют важную роль в таких фундаментальных биологических процессах, как биологический перенос, образование новых молекул или клеточная защита. Именно поэтому задача оценки качества моделей белковой структуры так важна на сегодняшний день. Существует множество методов для решения данной задачи, но не все из них достаточно оптимальны. Интересным является метод SBROD, предложенный учеными Михаилом Карасиковым, Гийом Пажем и Сергеем Грудининым, который производит оценку качества белковой модели, основываясь только на геометрии белка. В нашей работе мы хотим улучшить представленный метод путем переобучения модели на метриках CAD и lDDT. CAD метрика количественно определяет различия между физическими контактами в модели и в эталонной структуре. В ней используется понятие разности контактных площадей остаток-остаток, введенное Абагяном и Тотровым. Контактные области, лежащие в основе оценки, получены с использованием тесселяции структуры белка Вороного. Локальный тест разности расстояний (lDDT)-это метрика без суперпозиции, которая оценивает локальные разности расстояний всех атомов в модели, включая проверку стереохимической правдоподобности. LDDT хорошо подходит для оценки качества локальной модели, сохраняя хорошую корреляцию с глобальными метрика ми. Выбор этих метрик обуславливается их приемуществом перед другими, ранее используемыми. Также предполагается улучшение функции потерь для оптимального выбора лучшей модели. Для этого будет применена оптимизация Z-score, которая представляет из себя числовое измерение отношения значения к среднему значению в группе значений. Если Z-оценка равна 0, она представляет собой оценку, идентичную средней. Z-оценки также могут быть положительными или отрицательными, с положительным значением, указывающим, что оценка выше среднего, а отрицательный показатель, указывающий, что он ниже среднего.

Литература

[1] Yifeng Yang. Scoring functions in predicting protein structure and protein-protein interaction, January 01 2010.

[2] Alexander Sasse, Sjoerd De Vries , christina Schindler , Isaure Chauvot de Beauchêne, and Martin Zacharias. Rapid design of knowledge-based scoring potentials for enrichment of near-native geometries in protein-protein docking. January 2017.

[3] Dorota Latek and Andrzej Kolinski. Contact prediction in protein modeling: Scoring, folding and refinement of coarse-grained models. *BMC Structural Biology*, 8(36), August 11 2008.

[4] Andrew J. Bordner. Orientation-dependent backbone-only residue pair scoring functions for fixed backbone protein design. *BMC Bioinformatics*, 11:192, 2010.

[5] Rocke. A hybrid scoring function for protein multiple alignment. In *WABI: International Workshop on Algorithms in Bioinformatics, WABI, LNCS*, 2002.

[6] S. F. Altschul. A protein alignment scoring system sensitive to all evolutionary distances. *J. Mol. Evol.*, 36:290–300, 1993.

[7] Yungki Park. Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences. *BMC Bioinformatics*, 10:419, 2009.

[8] Daniel Carbajo and Anna Tramontano. A resource for benchmarking the usefulness of protein structure models. August 02 2012.

[9] Pralay Mitra. Algorithmic approaches for protein-protein docking and quarternary structure inference, July 2011.

[10] John Moult, Krzysztof Fidelis, Burkhard Rost, Tim Hubbard, and Anna Tramontano. Proteins: Structure, function, and bioinformatics suppl 7:3–7 (2005) critical assessment of methods of protein structure prediction (casp)—round 6. August 12 2013.

[11] L. Zhang and J. Skolnick. What should the z-score of native protein structures be?

[12] Davide Salvatore Mare, Fernando Moreira, and Roberto Rossi. Nonstationary z-score measures. *European Journal of Operational Research*, 260(1):348–358, 2017.

[13] Ilia Parshakov, Craig A. Coburn, and Karl Staenz. Z-score distance: A spectral matching technique for automatic class labelling in unsupervised classification. In *IGARSS*, pages 1793–1796. IEEE, 2014.

[14] Harry M. Sneed. Test metrics. *Metrics News, Journal of GI-Interest Group on Software Metrics*, 12(1):41–51, 2007.

[15] Robert R. Hoffman, Peter A. Hancock, and Jeffrey M. Bradshaw. Metrics, metrics, metrics, part 2: Universal metrics? *IEEE Intelligent Systems*, 25(6):93–97, 2010.