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 новому методу оценки качества белка(SBROD), использующему только форму белкового скелета. Предлагаемый метод выводит свою скоринговую функцию, основываясь на учебном наборе белковых моделей. Функция подсчета SBROD состоит из четырех терминов, связанных с различными структурными особенностями белка.

Введение

Белки играют важную роль в таких фундаментальных биологических процессах, как биологический перенос, образование новых молекул или клеточная защита. Это значение вызвало широкое исследование их свойств, которое требует дорогостоящих экспериментов. Потенциально они могут быть заменены более дешевыми и более быстрыми численными методами моделирования. Большинство предложенных методов для прогнозирования структуры белка сначала генерируют набор правдоподобных белковых моделей, а затем ранжирует их с использованием определенного QA метода. Обычно эти методы основаны на скоринговых функциях, которые предсказывают сходство между белковыми моделями и целевыми структурами в таких показателях сходства, как RMSD, GDT-TS и TM-score. В частности, RMSD измеряет среднее расстояние между атомами двух наложенных конформаций белка. GDT-TS и TM-score разработаны для оценки качества белковых моделей, являющихся независимыми от размера белка, и устойчивые к локальным структурным ошибкам. Существует два основных метода оценки качества. Модель консенсуса определяют качество отдельных моделей белка на основе их статистики в оценочном наборе. Напротив, одномодельные методы рассматривают только атомы оцениваемого белка без дополнительной информации о других моделях, и, следовательно, их можно использовать для конформационной выборки. Среди недавно предложенных одномодельных QA методов существуют два основных подхода к разработке оценочной функции: основанный на физической модели и основанный на данных. Построенные на физической модели функции подсчета очков используют некоторую информацию о взаимодействиях в системе, например, принцип минимизации энергии Гиббса. Однако точная оценка свободной энергии Гиббса требует выборки огромного количества конформационных состояний, которые в большинстве практических случаев является трудноразрешимым. Данные методы направлены на построение скоринговых функций, которые аппроксимируют энтальпическую часть свободной энергии Гиббса, раскладывая её на сумму добавочных членов (вкладов), которые представляют собой растяжение связей или углов, диэдральные потенциалы, электростатические и Ван дер Ваальсовые взаимодействия и т. д. Наряду с физическими подходами, существуют подходы, основанные на данных, которые выводят энергию молекулярных взаимодействий из баз данных, предполагающих определенное распределение конформаций или минимизируя определенную функцию потерь. Соответствующие функции подсчета очков обычно производятся либо путем машинного обучения,

либо путем оценки вероятности некоторых конформаций (статистические методы) с использованием статистических данных об определенных белковых структурах из структурных баз данных. В данной работе изучается новый метод оценки качества белка, Smooth Backbone-Reliant Orientation-Dependent (SBROD). SBROD является одномодельным методом QA, который оценивает белок модели, используя геометрические структурные особенности. Он требует только координат белковой основы и, следовательно, нечувствителен к конформациям боковых цепей. Кроме того, SBROD функция скоринга непрерывна по отношению к координатам атомов белка, что делает её также потенциально применимой для использования в молекулярной механике.

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.