Ученые записки Таврического национального университета им. В.И. Вернадского
Серия «Биология, химия» Том 18 (57). 2005. № 3. С. 180-186.

УДК 571

ПРЕДСКАЗАНИЕ ТРЕТИЧНОЙ СТРУКТУРЫ БЕЛКА С ПОМОЩЬЮ КОМПЬЮТЕРА

Терещенко Ф.А.

Понимание процесса и механизма укладки белковой молекулы – одна из важных, но пока неразрешенных проблем биологии. До сих пор не удается объяснить, не только чем определяется расположение атомов белка в пространстве, но и предсказать это расположение исходя только из последовательности аминокислот.

Методы, уверенно и верно предсказывающие третичную структуру белка, послужили бы большим подспорьем и для чисто научного понимания биологического процесса сворачивания белка, и для практических целей, таких как понимание механизмов катализа, регуляторных каскадов, регуляции экспрессии генов, разработка лекарственных препаратов, и новых генно-инженерных методов лечения.

Сложность проблемы предсказания третичной структуры состоит в том, что белковая молекула даже небольшой длины может, теоретически, принимать астрономическое количество конформаций. Известно, что в природе молекула белка следует определенному пути сворачивания [1], иначе, для образования третичной структуры путем перебора всех возможных конформаций молекула белка в 100 аминокислот сворачивалась бы в худшем случае около 10^{25} лет, что больше, чем время существования Вселенной (парадокс Левенталя [2]). Аналогично, для предсказания структуры даже такого короткого белка методом полного перебора всех конформаций понадобился бы компьютер мощнее теоретически возможного (предел Бремерманна [3]). Таким образом, существует необходимость разработки методов предсказания, дающих оптимальный или околооптимальный результат за приемлемое время.

Следует упомянуть, что основным и *de facto* стандартным депозитарием третичных структур белков является база данных PDB (Protein Data Bank) [4], в которую заносятся как структуры белка установленные физико-химическими методами, так и, в последнее время, компьютерные предсказания.

Таким образом, современные методы предсказания третичной структуры белка основываются на компромиссных подходах, не дают общего решения, имеют ограниченную область применения и свои плюсы и минусы.

В связи с этим целью данной работы явилось проведение краткого обзора методов предсказания третичной структуры белка и анализ основных подходов в моделировании *ab initio* по распознаванию способа сворачивания (т.н. «протягивание» - «threading») белка.

Критерии качества предсказания.

Основным критерием качества предсказания является среднеквадратическая разница (σ) расстояний между всеми атомами предсказанного белка и атомами этого же белка, структура которого определена физическими методами: рентгеновской кристаллографией или ядерным магнитным резонансом [5]. Часто используются компромиссные оценки. Например, замеряется расстояние не между всеми парами атомов, а только между атомами белкового остова.

Моделирование *ab initio*.

Попытка предсказания третичной структуры белка, исходя только из его первичной аминокислотной последовательности, была впервые предпринята проф. Харольдом Шерагой и коллегами [6]. Этот метод основан на применении знаний о физических взаимодействиях как внутри белковой молекулы, так и о влиянии окружающей среды (раствора). Основная посылка этого метода – утверждение, что белок в растворе принимает конформацию, соответствующую его минимальной энергии. Для задачи предсказания структуры энергетический ландшафт определяется набором формул, происходящих из физических измерений и эмпирических заключений. Например, самыми очевидными компонентами формулы энергии являются ограничения на длину валентной связи, валентных и торзионных углов. Количество независимых переменных функции энергии растёт экспоненциально с увеличением количества атомов в молекуле. Сложным этапом является подбор параметров энергетической функции. Чаще всего пользуются общепризнанными стандартами – силовыми полями AMBER [7] и CHARMM [8]. Они представляют собой набор параметров взаимодействий, полученных либо экспериментальным путём, либо эмпирически. Функции энергии имеют обычно огромное количество независимых переменных. Все значения такой функции называются энергетическим ландшафтом. Для функции с тремя переменными такой ландшафт будет трёхмерным, для функции с N переменными – ландшафт будет N-мерным.

Задачей-максимумом метода *ab initio* является нахождение глобального минимума для данного энергетического ландшафта. Если функция энергии составлена верно, все параметры взаимодействия учтены, – глобальный минимум будет совпадать с естественной трёхмерной структурой белка. К сожалению, подбор «правильной» функции и определение «правильных» параметров является трудноразрешимой задачей из-за неполноты наших знаний о физике белковых молекул и их взаимодействия с раствором и между собой.

Основной трудностью нахождения минимума энергии белковой молекулы является сложность многомерного энергетического ландшафта со множеством локальных минимумов энергии. Поскольку полный перебор всех энергетических состояний для сложной молекулы невозможен, для поиска глобального минимума используются приблизительные методы. Наиболее распространен метод Монте-Карло [9]. Вероятность нахождения глобального минимума прямо пропорциональна вычислительной мощности компьютера и количеству времени, потраченному на вычисления.

Современные методы предсказания *ab initio* белков в среднем дают качество предсказания порядка 10 Å, и только для очень коротких белков - около 1-2 Å, пригодное для использования в компьютерной фармакологии и дизайне лекарств. Неоспоримым преимуществом этой группы методов является то, что предсказание не требует знаний структур никаких схожих или гомологических белков.

Одни из лучших результатов предсказания дают алгоритмы, разработанные в Корнельском университете в США проф. Шерагой. Описанный метод позволил получить предсказание с среднеквадратичным отклонением меньше 6 Å для фрагментов белков длиной около 60-70 аминокислотных остатков с преобладанием α-спиралей. В то же время серьёзные топологические ошибки были допущены для α-β белков. Для целых α-β белков метод дал предсказание с σ=7.3 Å, что является лучшим достижением в данной области в настоящий момент [10].

Гомологическое моделирование.

Быстрый рост базы данных трёхмерных структур белков позволяет использовать в предсказаниях информацию о сходных, гомологичных, эволюционно близких белках. Одним из самых распространенных методов предсказания структуры белка является моделирование по гомологии. Известно, что эволюционно близкие белки или их домены имеют схожую третичную структуру. Консервация функции обуславливает селекцию против мутаций, приводящих к значительному изменению аминокислотной последовательности белка. В то же время небольшие изменения могут не привести к неправильному сворачиванию или полному разрушению третичной структуры. Впервые устойчивость белков к аминокислотным заменам и относительную похожесть – гомологию аминокислот - исследовала статистическими методами Др. Маргарет Дэйхофф [11]. Ею были составлены первые матрицы аминокислотных замен. Их усовершенствованные и более селективные варианты широко используются в настоящее время как метрики при сравнении (alignment) двух или более аминокислотных последовательностей белков.

Основным условием применимости метода является наличие в базе данных белка с гомологичной аминокислотной последовательностью. В случае наличия более чем одного белка-гомолога задача предсказания существенно упрощается, т.к. появляется возможность избежать ошибок выборки, часто возникающих при сравнении только двух белков. Выбор наилучшего гомолога или их серии из базы данных безусловно определяет качество предсказания, но очень часто является интуитивным процессом, опирающимся только на опыт исследователя. Сам поиск гомологов осуществляется широко распространенными алгоритмами семейства BLAST [12, 13]. Выбор конкретного алгоритма и метода поиска зависит от степени похожести белков и конкретной задачи моделирования. В последнее время появились методы массового гомологического моделирования, основанного на автоматизированном выборе гомологов [14]. Этот метод, как и все методы типа high-throughput не позволяют достичь оптимального качества предсказаний.

Следующий этап предсказания методом гомологического моделирования – предсказание вторичной структуры белка. Существует большое количество

программ, позволяющих делать такие предсказания. Одним из старейших является основанный на нейронных сетях метод PHD [15].

Сопоставление (алаймент) известного белка с предсказываемым осуществляется с помощью широко распространённых программ, например CLUSTALW [16], при этом ограничения на разрывы накладываются таким образом, чтобы они не приходились на непрерывные α-спирали и β-слои. В случае необходимости введения очень больших промежутков белки могут быть разбиты на фрагменты и предсказаны раздельно по доменам.

Дальнейший комплекс алгоритмов можно разделить на три группы: перенос координат, соответствующих непрерывным α- и β-структурам от гомолога к предсказуемому белку; предсказание петель и др. структур, отличающихся по длине; расстановка боковых цепей аминокислот.

Перенос координат α- и β-структур от гомолога к предсказываемому белку является тривиальным процессом. Намного более сложную проблему представляет собой замыкание петель между этими структурами. Используемые методы варьируют от предсказания, сходного с *ab initio*, до попыток даже аналитического решения задачи [17].

Боковые цепи аминокислот либо не предсказываются вообще, либо вычисляются методами, схожими с *ab initio*. В последнее время популярность приобрели так называемые библиотеки ротамеров. Для создания такой библиотеки вычисляется статистическое распределение торсионных углов ψ_1 и ψ_2 боковых цепей аминокислот как функция от углов φ и ψ данной аминокислоты в уже предсказанном белковом остове [18]. Углы выше ψ_2, там, где они есть, расставляются как функция от углов ψ_1 и ψ_2.

После грубого гомологического моделирования часто возникают стерические ошибки. Предсказание может улучшаться методами *ab initio*.

Наиболее популярным, но далеко не самым доступным и простым в использовании является пакет программ MODELLER, разработанных в Рокфеллеровском Университете Андреем Шали [19].

Основными недостатками метода гомологического моделирования является фактическая неприменимость его в области низких гомологий (меньше 20-25 %) и полная невозможность предсказания в случае отсутствия гомологов. С другой стороны, метод относительно нетруден в использовании даже для непрофессионалов, и даёт надёжные предсказания в области высоких гомологий (часто ε<1Å) [20].

Моделирование методом threading

Одним из методов предсказания в области низких гомологий является метод, получивший название распознавание способа сворачивания (fold recognition) или протягивание (threading). Очень часто даже эволюционно далекие белки сворачиваются аналогично в результате эволюционной конвергенции. Существует, правда пока недоказанное мнение, что белки могут принимать только ограниченное количество суб-третичных структур – «фолдов» [21].

Для предсказания данным методом сначала должна быть сгенерирована библиотека фолдов. Чаще всего в виде профайлов. Профайл создаётся из

сопоставления нескольких гомологичных белков данного фолда. При этом учитывается не только то, какая аминокислота находится в данной позиции (т.н. консенсусная последовательность), но и окружающие её аминокислоты [22]. Основу для создания профайлов могут представлять базы данных фолдов, например SCOP [23].

Следующим этапом, является сравнение предсказываемого белка с базой данных профайлов [24]. Очень часто распознавание фолдов заканчивается, если найден профайл, удовлетворяющий условию поиска. Исследователю часто бывает достаточно установления белковой семьи, к которой принадлежит данный белок. Если же необходимо именно предсказание третичной структуры – используются методы аналогичные *ab initio* или моделированию по гомологии.

Основным недостатком метода является полная невозможность предсказания белка с дотоле неизвестным фолдом. Метод может дать хорошие предсказания в области, где моделирование по гомологии неприменимо, но в то же время уступает по качеству гомологическому моделированию в области высоких гомологий (выше 25-30 %). Часто используемыми методами распознавания фолда являются разработки проф. Манфреда Зиппла в Зальцбургском Университете [25] и проф. Джефри Скольника в Институте Скриппса [26].

В последнее время распространение получил метод mini-threading [27]. Основное отличие его в том, что профайлы составляются для небольших блоков – участков белков длиной в несколько аминокислот. Сборка всего белка происходит с помощью специальных функций энергии методом минимизации. Mini-threading позволяет достаточно уверенно предсказывать белки недоступные для гомологического моделирования.

Проверка качества предсказания

Проверка качества методов предсказания и сравнения нескольких методов была упрощена организацией симпозиумов CASP (критическая оценка предсказания структуры). Суть эксперимента состоит в том, что организаторы симпозиума в сотрудничестве с кристаллографами и специалистами по ЯМР определяют набор белков, структура которых будет в скором времени разрешена. За несколько месяцев публикуются только аминокислотные последовательности этих белков, без указания их функции, класса, и т.д. После получения реальных структур белков, полученные предсказания оцениваются независимыми экспертами, которые не принимают участия в эксперименте. Качество предсказаний оценивается по множеству критериев, в том числе и по среднеквадратичному отклонению от экспериментально разрешенной третичной структуры. Результаты анализируются и публикуются как на веб-странице CASP'а [28], так и в журнале Protein Structure [29].

ВЫВОДЫ

1. Описанные методы предсказания третичной структуры белка: *ab initio*, гомологическое моделирование и *threading* имеют свои достоинства и недостатки. При этом ни один из них не позволяет получить надежное предсказание структуры любого неизвестного белка.

2. Основной задачей вычислительной биологии, занимающейся предсказанием третичной структуры белка, является как развитие и совершенствование вышеописанных методов, так и разработка новых, до получения в перспективе надёжной технологии предсказания структуры любого белка вне зависимости от его молекулярной массы или наличия гомологий среди уже известных структур.

Список литературы

1. Callender, R., Gilmanshin, R., Dyer, B., Woodruff, W. The Primary Processes of Protein Folding // Ann. Rev. Phys. Chem. – 1998. – V. 49. – P.173-202.
2. Honig, B. Protein folding: from the Leventhal paradox to structure prediction // J. Mol. Biol. - 1999. – V. 293. – P. 283–293.
3. Bremermann, H. J. Minimum energy requirements of information transfer and computing // Int. J. Theor. Phys. - 1982. – V. 21. – P. 203-217.
4. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. The Protein Data Bank // Nucleic Acids Research. – 2000. – V. 28. – P.235-242.
5. Cohen, P.E., Sternberg, M.J.E. On the prediction of protein structure: the significance of the root mean square deviation // J. Mol. Biol. – 1980. – V. 138. – P. 321-333.
6. Leach, S.J., Nemethy, G., Scheraga, H.A. Computation of the sterically allowed conformations of peptides //Biopolymers. - 1966 – V. 4. – P. 369-407.
7. Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S., Cheatham, T.E.III, DeBoult, S., Ferguson, D., Seibel, G., Kollman P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate structural and energetic properties of molecules // Comp. Phys. Commun. - 1995. – V. 91. – P. 1-41.
8. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., Karplus. M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations // J.Comp. Chem.- 1983. – V. 4. – P. 187-217.
9. Liu, J.S. Monte Carlo strategies in scientific computing. - Springer Verlag, 2001. - 352 p.
10. Ołdziej, S., Czaplewski, C., Liwo, A., Chinchio, M., Nanias, M., Vila, J.A., Khalili, M., Arnautova, Y.A., Jagielska, A., Makowski, M., Schafroth, H.D., Kaźmierkiewicz, R., Ripoll, D.R., Pillardy, J., Saunders, J.A., Kang, Y.K., Gibson, K.D., Scheraga, H.A. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests // Proc. Natl. Acad. Sci. USA. - 2005. –V. 102. – P. 7547-7552.
11. Dayhoff, M., Barker, W.C., Hunt, L.T. Establishing homologies in protein sequences // Methods Enzymol. - 1983. – V. 91. – P.524-545.
12. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. Basic local alignment search tool // J. Mol. Biol. - 1990. – V. 213. – P. 403-410.
13. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs // Nucleic Acids Res. - 1997. – V. 25. – P.3389-3402.
14. Schwede, T., Kopp, J., Guex, N., Peitsch, M.C. SWISS-MODEL: An automated protein homology-modeling server // Nucleic Acids Res. – 2003. – V. 31. – P. 3381-3385.
15. Rost, B., Sander, C., Schneider, R. PHD--an automatic mail server for protein secondary structure prediction // Comput Appl Biosci. – 1994. – V. 10. – P. 53-60.
16. Higgins, D.G., Sharp, P.M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer // Gene. - 1988. – V. 73. – P.237-44.
17. Wedemeyer, W.J., Sheraga, H.A. Exact analytical loop closure in proteins using polynomial equations // J.Comp.Chem. – 1999. – V. 20. – P. 819-844.
18. Dunbrack, R.L. Jr. Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL // Proteins. Suppl. – 1999. – V. 3. – P.81-87.
19. Šali, A., Potterton, L., Yuan, F., van Vlijmen, H., Karplus, M. Evaluation of comparative protein modeling by MODELLER // Proteins. – 1995. – V. 23. – P. 318-326.

20. Tereshchenko, F.A., Daraselia, N.D. A homology modeling algorithm for protein structure prediction / Proceedings of CASP4 Conference. - 2000. – P. 50-51.
21. Murzin, A.G., Bateman, A. Distant homology recognition using structural classification of proteins // Proteins. Suppl. – 1997. – V. 1. – P.105-112.
22. Shi, J., Blundell, T.L., Mizuguchi, K.. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure- dependent gap penalties // J. Mol. Biol. – 2001. – V. 310. – P. 243-257
23. Murzin, A., Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., Murzin, A.G. SCOP database in 2004: refinements integrate structure and sequence family data // Nucleic Acids Res. - 2004. – V. 32. – P. D226-229.
24. Kelley, L.A., MacCallum, R.M., Sternberg, M.J.E. Enhanced genome annotation using structural profiles in the program 3D-PSSM // J. Mol. Biol. - 2000. – V. 299. – P. 499-520.
25. Koppensteiner, W.A., Lackner, P., Wiederstein, M., Sippl, M.J. Characterization of novel proteins based on known protein structures // J Mol Biol. - 2000. – V. 296. – P. 1139-1152.
26. Zhang, Y., Skolnick, J. Tertiary structure predictions on a comprehensive benchmark of medium to large size proteins // Biophys J. – 2004. – V. 87. – P.2647-2655.
27. Bradley, P., Misura, K.M., Baker, D. Toward high-resolution de novo structure prediction for small proteins // Science. - 2005. – V. 309. – P.1868-1871.
28. http:www.predictioncenter.org.
29. Venclovas, Č., Zemla, A., Fidelis, K., Moult, J. Assessment of progress over the CASP experiments // Proteins. – 2003. - Suppl 6. – P. 585-595.

Поступила в редакцию 22.09.2005 г.