УДК 519.21

А.В. БЫЦЬ

УПРОЩЕННОЕ ПРЕДСТАВЛЕНИЕ ПРОСТРАНСТВЕННОЙ СТРУКТУРЫ БЕЛКА В ВИДЕ КАРТЫ КОНТАКТОВ

Рассмотрено упрощенное представление пространственной структуры белка в виде карты контактов. Изучены преимущества использования карт контактов для предсказания пространственных структур белков и для их классификации. Проведен обзор и сравнительный анализ существующих методов предсказания карты контактов по известной аминокислотной последовательности белка и методов предсказания пространственной структуры белка по известной карте контактов последнего.

Введение. Молекула любого белка представляет собой линейную последовательность аминокислот. Порядок их расположения в последовательности называется первичной структурой белка. Каждая встречающаяся в природе аминокислотная последовательность определенным образом располагается в трехмерном пространстве. Только для небольшой части встречающихся в природе аминокислотных цепочек возможны несколько отличающихся друг от друга пространственных расположений или отсутствие устойчивого пространственного расположения.

Участки разных белков могут иметь одинаковые пространственные расположения. Так называемые вторичные структуры охватывают непрерывные отрезки аминокислотной последовательности и обладают следующим свойством: если отрезок из нескольких аминокислот образует вторичную структуру определенного типа, то расположение этих аминокислот относительно друг друга в пространстве зависит только от типа этой вторичной структуры и практически не зависит от самой аминокислотной последовательности. Наиболее распространенными типами вторичных структур являются β-структуры и правозакрученные α-спирали. Они присутствуют в подавляющем большинстве белков. Остальные типы вторичных структур или встречаются в белках очень редко, или мало отличаются от двух вышеупомянутых.

Пример пространственной конфигурации молекулы белка схематически показан на рис. 1, где 1 – α-спирали; 2 – параллельные тяжи β-структур; 3 – антипараллельные тяжи β-структур.

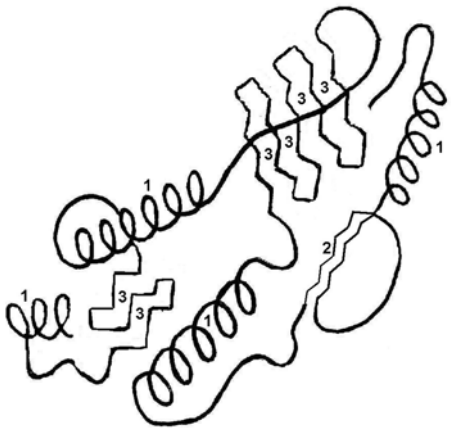

РИС. 1. Пример пространственной конфигурации молекулы белка

Способы представления пространственной структуры белка. Пространственная структура белка может описываться координатами всех его атомов в эвклидовом пространстве. Другое возможное описание – это матрица расстояний – симметричная матрица, содержащая попарные расстояния между всеми атомами. При построении матрицы расстояний для простоты могут браться попарные расстояния не между атомами, а между аминокислотами белка. Главные преимущества матрицы расстояний – ее независимость от расположения начала и осей координат, возможность ее применения, когда известна только часть данных. Задача восстановления координат атомов по матрице расстояний имеет решение за полиномиальное время.

Карта контактов (или контактная карта) белка – это еще более упрощенное по сравнению с матрицей расстояний представление пространственной структуры, представляющее собой булеву симметричную матрицу $C=[c_{ij}]$ размерностью $n \times n$, где n – количество аминокислот в белке; $i, j = 1,2,\ldots, n$ – порядковые номера аминокислот в первичной структуре белка. Элементы матрицы удовлетворяют условиям: $c_{ij} = 1$, если расстояние между i-й и j-й аминокислотами белка меньше некоторого порогового значения, и $c_{ij} = 0$ – в противном случае.

Поскольку каждая аминокислота состоит из определенного количества атомов, расстояние между двумя аминокислотами может определяться как расстояние между какими-то атомами, принадлежащими этим аминокислотам, либо как расстояние между центрами масс аминокислот. Чаще всего для построения контактных карт используются следующие пороговые значения расстояний между двумя аминокислотами:

- расстояние между их C_α атомами с порогом 6–12 Å;
- расстояние между их C_β атомами с порогом 6–12 Å (для аминокислоты глицина, в которой нет C_β атома, в этом случае используется C_α атом, который есть в каждой аминокислоте; C_β атомы есть во всех аминокислотах, кроме глицина);
- наименьшее расстояние между их любыми атомами (кроме атомов водорода) с порогом 4,5–6 Å.

На рис. 2 показан пример контактной карты белка. Контакты (элементы матрицы, равные единице) обозначены квадратами с точкой посредине. Для удобства на рис. 2 отображены элементы только одной из двух полностью симметричных половин матрицы, причем, только те, для которых $|i-j| \geq 4$, где i, j – порядковые номера аминокислот в аминокислотной последовательности белка. В правой части рис. 2 схематически представлена соответствующая этой контактной карте пространственная структура белка (α-спираль обозначена лентой, а β-тяжи – стрелками).

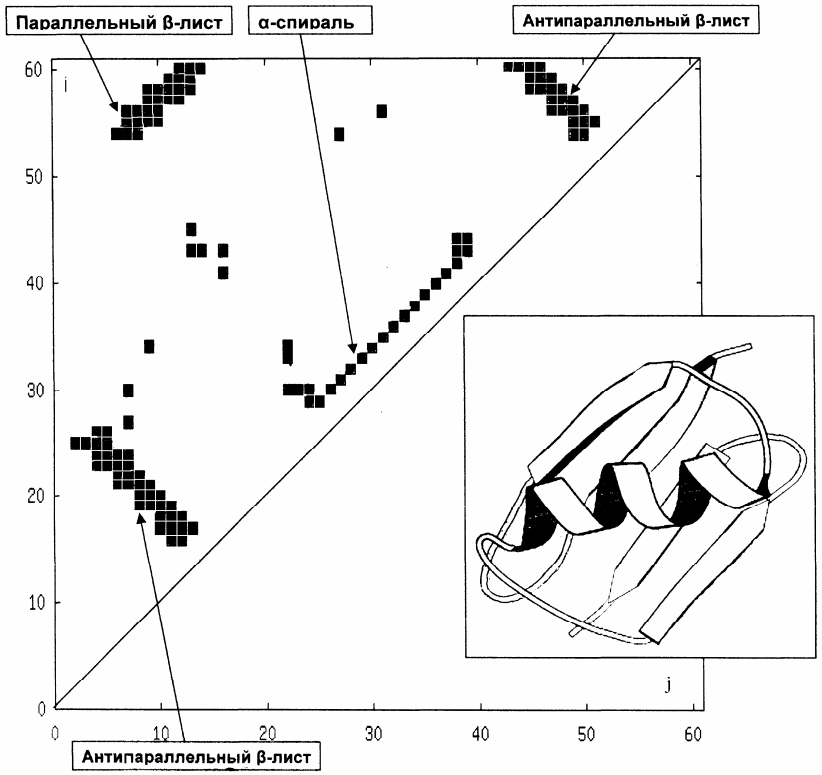

РИС. 2. Пример контактной карты белка, состоящего из 61 аминокислоты

Взаимное расположение контактов на контактной карте подчиняется следующим правилам, причиной которых являются строгие геометрические закономерности строения вторичных структур:

– контакты между аминокислотами одной α-спирали расположены на расстоянии, не большем 4 позиций от главной диагонали карты;
– контакты между аминокислотами двух соседних тяжей антипараллельной β-структуры на карте занимают область толщиной в несколько позиций, вытянутую перпендикулярно главной диагонали карты;

– контакты между аминокислотами двух соседних тяжей параллельной β-структуры занимают область толщиной в несколько позиций, вытянутую параллельно главной диагонали карты (см. рис. 2).

Для аминокислот i, i+1, i+2, j, j+1, j+2 из параллельной β-структуры,

если $C(i, j) = 1$ и $C(i + 2, j+2) = 1$,

то $C(i, j+2) = 0$ и $C(i + 2, j) = 0$;

для аминокислот i, $i + 1$, $i + 2$, j, $j + 1$, $j + 2$ из антипараллельной β-структуры:

если $C(i, j + 2) = 1$ и $C(i + 2, j) = 1$,

то $C(i, j) = 0$ и $C(i + 2, j + 2) = 0$;

для аминокислот i, $i + 1$, $i + 2$, $i + 3$, $i + 4$ из α-спирали,

если $C(i, i + 4) = 1$, $C(i, j) = 1$, и $C(i + 4, j) = 1$,

то $C(i + 2, j) = 0$.

Последние три правила проиллюстрированы на рис. 3, где схематически показаны контакты между некоторыми аминокислотами: а – антипараллельная β-структура; б – параллельная β-структура; в – α-спираль белка.

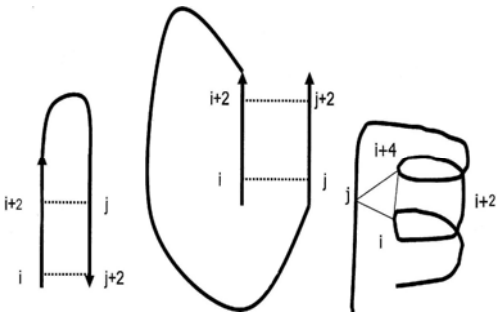

РИС. 3. Контакты между аминокислотами во вторичных структурах

Контактные карты могут использоваться как промежуточное звено в процессе предсказания координат атомов белка в трехмерном пространстве. Целесообразность такого их применения связана с тем, что по известной аминокислотной последовательности контакты предсказываются проще, чем непосредственно координаты атомов белка в трехмерном пространстве. Эти координаты при определенных условиях можно реконструировать по известной контактной карте [1].

Предсказание контактной карты по известной аминокислотной последовательности белка. Существует много различных методов для предсказания контактной карты по известной аминокислотной последовательности белка. Их можно разделить на две взаимо не исключающие категории. К первой относятся статистические подходы, использующие метод коррелирущих мутаций. Они основаны на идее, что поддержание функций белка накладывает ограничения на эволюцию его аминокислотной последовательности, т.е. если мутации подвергается какая-то аминокислота, то вероятно, что и контактирующая с ней аминокислота тоже мутирует, чтобы компенсировать произошедшие изменения химических свойств. Поэтому коррелирующие мутации, наблюдающиеся в похожих аминокислотных последовательностях, могут быть использованы как индикаторы возможного контакта аминокислот.

В методах предсказания контактных карт, относящихся ко второй группе, используются методы машинного обучения, такие как нейронные сети [2, 3], скрытые Марковские модели, метод опорных векторов [4] и генетическое программирование. Могут использоваться также комбинации нескольких методов.

В зависимости от подхода, используемого для предсказания карт контактов, точность предсказания может зависеть от качества автоматизированного сравнительного анализа аминокислотных последовательностей (так называемого множественного выравнивания последовательностей) и от точности предсказания вторичных структур. Кроме того, она всегда связана с долей в белке аминокислот, входящих в β-листы: в β-белках, α+β-белках и α / β-белках контакты между тяжами одного β-листа предсказываются с большей точностью, чем контакты между α-спиралью и β-листом или между разными α-спиралями. Это наглядно демонстрируют данные из табл 1 [4]. В ней в первом столбце указаны классы согласно классификации SCOP [5], к которым относятся пространственные структуры белков. Во втором столбце указано количество белков каждого класса, для которых предсказывались контакты. В остальных столбцах приведена точность (отношение количества правильно предсказанных контактов к количеству всех предсказанных контактов) и «покрытие» (отношение количества правильно предсказанных контактов к количеству действительно существующих контактов) предсказаний контактов в случаях, когда учитываются только те пары, для которых соответственно $|i-j| \geq \geq 6$, $|i-j| \geq 12$ и $|i-j| \geq 24$, где i и j – порядковые номера аминокислот в аминокислотной последовательности.

ТАБЛИЦА. Точность предсказания контактов в белках из 6 классов белковых структур по классификации SCOP

Название класса согласно SCOP	Количество белков	$\|i-j\| \geq 6$		$\|i-j\| \geq 12$		$\|i-j\| \geq 24$	
		Точность	Покрытие	Точность	Покрытие	Точность	Покрытие
alpha	11	0.24	0.24	0.17	0.18	0.11	0.09
beta	10	0.38	0.17	0.32	0.17	0.22	0.17
a + *b*	15	0.45	0.25	0.35	0.25	0.21	0.23
a / *b*	7	0.37	0.19	0.33	0.19	0.28	0.20
Малые	4	0.36	0.18	0.28	0.19	0.11	0.15
coil-coil	1	0.22	0.40	0.03	0.16	0.00	–
Все	48	0.37	0.21	0.30	0.20	0.21	0.19

Однако контакты с участием аминокислоты, не входящей во вторичную структуру, контакты между двумя разными α-спиралями, между α-спиралью и β-структурой, между двумя разными β-структурами, не параллельными и не антипараллельными друг другу, как правило, предсказываются значительно хуже, чем контакты между аминокислотами одной вторичной структуры (рис. 4), где в верхнем правом треугольнике показаны истинные контакты белка 1DZOA, а в нижнем левом треугольнике – предсказанные контакты этого белка. Интересно, что большинство неверно предсказанных контактов расположены поблизости от настоящих контактов [6, 7].

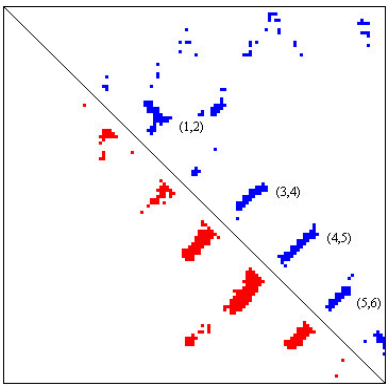

РИС. 4. Карты предсказанных и истинных контактов

Наибольшая точность предсказания контактов по известной аминокислотной последовательности составила 32 % [3]. Эта точность выше точности предсказания координат в трехмерном пространстве по известной аминокислотной последовательности.

Предсказание координат атомов белка по известной карте контактов. Общая задача вычисления координат атомов белка в трехмерном пространстве, совместимых с данной контактной картой, известна как задача о существовании графа дисков единичного радиуса. Доказано, что она NP-трудная. Однако было разработано несколько эмпирических методов (например [1, 8]) для предсказания координат атомов белка по известной карте контактов. Для построенных по эмпирическим данным картам контактов (с пороговым расстоянием для контакта, 9 равным Å) координаты атомов белка могут быть восстановлены со средним квадратичным отклонением 1–2 Å. Однако по предсказанным картам контактов координаты атомов прогнозируются с неудовлетворительно высоким средним квадратичным отклонением более 3 Å.

Заключение. Благодаря относительной простоте алгоритмов предсказания карты контактов белка по известной аминокислотной последовательности и алгоритмов предсказания трехмерной структуры белка по известной карте контактов представляется перспективным использование карт контактов как промежуточного звена в процессе предсказания координат атомов белка в трехмерном пространстве по известной аминокислотной последовательности белка. Однако указанные алгоритмы требуют усовершенствования в связи с недостаточной точностью предсказаний [3, 4, 9–12]. Кроме того, карты контактов как очень простое и наглядное описание пространственной структуры белка (по сравнению с описанием с помощью координат атомов в трехмерном пространстве) представляют большой интерес для эволюционной классификации белков и выявления эволюционного родства различных белков и их частей [12].

1. *Reconstruction* of 3D Structures From Protein Contact Maps / M. Vassura, L. Margara, Di P Lena et al. // EEE / ACM Transactions on Computational Biology and Bioinformatics. – 2008. – **Iss.** 5 (3). – P. 357–367.
2. *Pollastri G., Baldi P.* Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners // Bioinformatics. – 2002. – **18**. – P. 62–70.
3. *Punta M., Rost B*. PROFcon: novel prediction of long-range contacts // Bioinformatics. – 2005. – **21**, N 13. – P. 2960–2968.
4. *Cheng J., Baldi P.* Improved residue contact prediction using support vector machines and a large feature set // BMC Bioinformatics. – 2007. – **8**. – P. 1–9.
5. *Murzin A. G., Brenner S. E., Hubbard T., Chothia C*. SCOP: a structural classification of proteins database for the investigation of sequences and structures // J. Mol. Biol. – 1995. – Iss. 247. – P. 536–540
6. *Vullo A., Walsh I., Pollastri G*. A two-stage approach for improved prediction of residue contact maps // BMC Bioinformatics. – 2006. – **7**. – P. 100–112.
7. *Latek D., Kolinski A*. Contact prediction in protein modeling: scoring, folding and refinement of coarse-grained models // BMC Struct Biol. – 2008. – **8**. – P. 50–61.
8. *Vendruscolo M, Domany E.* Protein folding using contact maps // Vitam Horm. – 2000. – **58**. – P.171–212.
9. *Izarzugaza J.M., Graña O., Tress M.L., Valencia A., Clarke N.D*. Assessment of intramolecular contact predictions for CASP7 // Proteins. – 2007. – **69**. – P. 152–158.
10. CASP6 assessment of contact prediction / O. Graña, D. Baker, MacCallum R.M. et al. // Proteins. – 2005. – **61.** – P. 214–224.
11. *MacCallum R.M*. Striped sheets and protein contact prediction // Bioinformatics. – 2004. – **20**. – P. 1224–1231.
12. *Bartoli L., Capriotti E., Fariselli P., Martelli P.L., Casadio R.* The pros and cons of predicting protein contact maps // Methods Mol Biol. – 2008. – **413**. – P. 199–217.

О.В. Биць

СПРОЩЕНЕ ПОДАННЯ ПРОСТОРОВОЇ СТРУКТУРИ БІЛКА
У ВИГЛЯДІ КАРТИ КОНТАКТІВ

Розглянуто спрощене представлення просторової структури білка у вигляді карти контактів. Вивчені переваги використання карт контактів для передбачення просторових структур білків та для їх класифікації. Проведено огляд і порівняльний аналіз існуючих методів передбачення карти контактів за відомою амінокислотною послідовністю білка та методів передбачення просторової структури білка за відомою картою контактів останнього.

O.V. Byts

SIMPLIFIED REPRESENTATION OF A THREE-DIMENTIONAL PROTEIN STRUCTURE AS CONTACT MAPS

Contact map is considered which is the simplified representation of three-dimentional protein structure. The advantages of using contact maps for three-dimentional protein structure prediction and for protein classification are studied. The review and comparative analysis are made of existing methods for protein contact map prediction when the amino-acidic sequence is known and of methods for three-dimentional protein structure prediction when the protein contact map is known.

Получено 24.04.2009

Об авторе:

Быць Алексей Викторович,

кандидат технических наук, старший научный сотрудник
Института кибернетики имени В.М. Глушкова НАН Украины.
baleks@i.com.ua