3. *Burns P. A., Gordon A. J. E., Glickman B. W.* Influence of neighbouring base sequence on N-methyl-N′-nitro-N-nitrosoguanidine mutagenesis in the *lacI* gene *Escherichia coli* // J. Mol. Biol.— 1987.— **194**, N 4.— P. 385—390
4. *Golding G. B., Glickman B. W.* Sequence-directed mutagenesis: evidence from a phylogenetic history of human α-interferon genes // Proc. Nat. Acad. Sci. USA.— 1985.— **82**. N 12.— P. 8577—8581.
5. *Drake J. W., Baltz R. H.* The biochemistry of mutagenesis // Annu. Rev. Biochem.— 1976.— **45**.— P. 11—37.
6. *Индукция* повторяющихся нуклеотидных последовательностей. Вероятные механизмы эволюции генома и генной конверсии / Р. И. Салганик, А. В. Мазин, Г. Л. Дианов, Л. П. Овчинникова // Генетика.— 1984.— **20**, № 8.— С. 1244—1252.
7. *Льюин Б.* Гены.— М. : Мир, 1987.— 398 с.
8. *Уотсон Дж.* Молекулярная биология гена.— М. : Мир, 1978.— 467 с.
9. *Meselson M., Radding C.* A general model for genetic recombination // Proc. Nat. Acad. Sci. USA.— 1975.— **72**, N 5.— P. 358—361.
10. *Kunkel T. A.* The mutational specificity of DNA polymerases-α and -γ during in *vitro* DNA synthesis // J. Biol. Chem.— 1985.— **260**, N 26.— P. 12866—12879.
11. *Topal M. D., Eadie J. S., Conrad M.* O^6-Methylguanine mutation and repair is nonuniform // Ibid.— 1986.— **261**, N 21.— P. 9879—9885.
12. *Gearhart P. J., Bogenhagen D. F.* Clusters of point mutations are found exclusively around rearranged antibody variable genes // Proc. Nat. Acad. Sci. USA.— 1983.— **80**, N 16.— P. 3439—3443.
13. *A hyperconversion* mechanism generates the chicken light chain preimmune repertoire / C.-A. Reynaud, V. Anquez, H. Grimal, J.-C. Weill // Cell.— 1987.— **48**, N 2.— P. 379—388.
14. *Рогозин И. Б., Соловьев В. В., Колчанов Н. А.* Контекстная предетерминированность мутационного процесса (соматические, спонтанные и индуцированные точковые мутации.— Новосибирск, 1988.— 54 с.— (Препринт / Сиб. отд-ние АН СССР, Ин-т цитологии и генетики, № 08031).
15. *Berek C., Griffiths G. M., Milstain C.* Molecular events during maturation of the immune response to oxazoline // Nature.— 1985.— **316**.— P. 412—418.
16. *Rabbits T. N., Hamlyn P. H., Baer R.* Altered nucleotide sequences of a translocated *c-myc* gene in Burkitt lymphoma // Ibid.— 1983.— **306**.— P. 760—765.
17. *Hauser I., Levine A. S., Dixon K.* Unique pattern of point mutations arising after gene transfer into mammalian cells // EMBO J.— 1987.— **6**, N 1.— P. 63—67.

Ин-т цитологии и генетики Сиб. отд-ния АН СССР,
Новосибирск

Получено 06.04.90

УДК 577.323.435

А. П. Гультяев, Ю. Н. Монаков

МЕТОД ПОСТРОЕНИЯ ВТОРИЧНОЙ СТРУКТУРЫ РНК НА ОСНОВЕ ПРИНЦИПОВ САМООРГАНИЗАЦИИ

Описан метод компьютерного построения вторичной структуры РНК на основе последовательности нуклеотидов. Алгоритм использует метод Монте-Карло для моделирования самоорганизации РНК при последовательном образовании новых спиральных участков. Метод может быть использован для расчетов структур достаточно длинных РНК (вплоть до 5000 нуклеотидов) на персональном компьютере без использования больших объемов памяти.

Введение. Известно, что молекулы РНК имеют тенденцию к образованию двунитчатых участков с уотсон-криковскими парами нуклеотидов, связывая друг с другом отдельные фрагменты молекулы. Образующаяся структура именуется вторичной структурой РНК (этот термин используется также для описания молекулярной геометрии РНК и ДНК, т. е. A, B, C-структур и т. д., но в нашем случае мы под вторичной структурой будем понимать только наличие определенных двунитчатых участков в данной молекуле РНК). Вторичная структура РНК играет важную роль во многих процессах [1].

Зачастую экспериментальное определение вторичной структуры РНК затруднено, поэтому большое значение имеют теоретические мето-

ды построения вторичной структуры по известной нуклеотидной последовательности. В настоящее время разработан ряд алгоритмов и программ, решающих эту задачу, однако до сих пор не удается найти такого способа построения, который давал бы результаты, полностью соответствующие известным экспериментальным данным.

Наиболее приемлемыми по энергетике являются алгоритмы, использующие метод динамического программирования и его модификации [2, 3], при котором проводится поиск одной оптимальной структуры с минимальной свободной энергией. Такая детерминированность является и существенным ограничением этих методов. Предсказываемая структура может быть лишенной биологического смысла. К этому можно добавить, что при исследовании последовательностей различной длины время анализа растет пропорционально N^3 (N — число нуклеотидов). Последнее обстоятельство может создать большие практические трудности. Для длинных молекул РНК был предложен ряд методов расчета вторичной структуры, в которых поиск минимума энергии не является главным. В их основе всегда лежит принцип, позволяющий устанавливать приоритеты при отборе возможных элементов во вторичную структуру. Возникающая при этом свобода выбора относительно конечного результата, с одной стороны, может быть оправдана при сравнении расчета с данными экспериментов, а с другой — вполне компенсируется сокращением времени анализа до величины, пропорциональной N^2. Методы такого типа основаны на контекстном анализе исследуемых последовательностей и на различных принципах самоорганизации РНК [4—7]. Для длинных (более 500 нуклеотидов) РНК они представляются довольно перспективными, так как, по-видимому, молекула РНК при формировании своей структуры и стремлении к минимуму свободной энергии, соответствующему состоянию термодинамического равновесия, будет проходить через ряд состояний, имеющих локальные энергетические минимумы, причем время перехода к равновесному может быть весьма значительным.

В данной работе предложен метод построения вторичной структуры РНК, опирающийся на принцип последовательной самоорганизации молекулы РНК.

Он учитывает как энергетические параметры элементов вторичной структуры, так и топологические эффекты конкурирующих за образование стэкинг-областей потенциальных двунитчатых участков. Процесс самоорганизации моделируется методом Монте-Карло.

Методы. Описание алгоритма. Основой алгоритма является процесс самоорганизации молекулы РНК, состоящей в последовательности этапов, на каждом из которых образуются определенные двунитчатые участки (спирали). Фактически этот процесс заключается в некоторой конкуренции потенциальных спиралей за право быть включенными во вторичную структуру. Построение всех потенциальных спиралей, которые могут быть реализованы на данной последовательности, не представляет трудности, однако построение вторичной структуры РНК, собственно говоря, и состоит в выборе небольшого числа спиралей из огромного множества. Таким образом, мы приходим к вопросу о критерии, которым следует руководствоваться при выборе. Как уже отмечалось выше, поиск глобального минимума свободной энергии не всегда дает удовлетворительные результаты. Ясно, что в процессе самоорганизации при альтернативном выборе из всего набора конкретной спирали для включения ее во вторичную структуру, кроме величины ее энергии (с учетом дестабилизирующей энергии образующихся петель), принципиально важной является характеристика конкурентоспособности данной спирали по отношению ко всем остальным за включение во вторичную структуру на данном этапе анализа. Ясно также, что преимуществом обладают спирали, наименее препятствующие образованию других. Физически это можно представить как образование и нарушение (возможно, частичное) неких структур из данного набора спиралей, в результате которого образуются спирали, наиболее часто встречающиеся в этих динамических структурах. Такой процесс можно смоделировать методом Монте-Карло. В рамках определенного набора спиралей генерируется некоторое число случайных структур и после этого вычисляется частота встречаемости каждой спирали f. Эту величи-

ну вместе со свободной энергией спирали можно использовать при выработке правил отбора спиралей для включения во вторичную структуру. В настоящей работе в качестве критерия выбран параметр $f \cdot \Delta G$. При расчетах использовали значения энергий стэкинг-взаимодействий и петель, приведенные в работе [8]. Существенным моментом построения вторичной структуры является то, что изначально рассматривается неполный набор возможных спиралей. На первых этапах строили наборы наиболее энергетичных спиралей, а затем — с меньшей энергией, учитывая, что структура РНК уже частично сформирована, т. е. некоторым образом моделировали процесс самоорганизации РНК, состоящий из отдельных этапов.

Основная схема алгоритма приведена на рис. 1. В соответствии с вышеизложенным он разбивается на несколько главных блоков.

1. Построение набора спиралей. На этом этапе строятся все спирали, стэкинг-энергия которых меньше определенной величины. Очевидно, что эта величина является функцией длины последовательности. В данной работе эту функцию определяли эмпирически по наилучшим результатам для известных структур. Спирали строятся с дефектами, т. е. допускается наличие внутренних петель. Для этого вначале строится бездефектная спираль, затем (если это возможно) она удлиняется

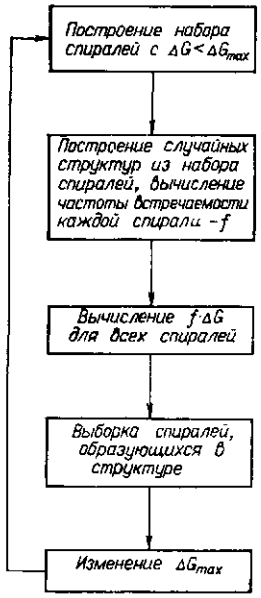

Рис. 1. Основные этапы расчета вторичной структуры
Fig. 1. The main stages of RNA secondary structure calculation.

с двух сторон при наличии внутренних петель с максимальным размером два нуклеотида по каждой нити, если это приносит выигрыш в энергии. На этом вычисляют энергию каждой спирали с учетом всех петель, которые образуются или разрушаются в уже имеющейся после предыдущих этапов структуре РНК.

2. Построение случайных структур из набора спиралей. Эти структуры строятся следующим образом: из данного набора спиралей с помощью метода Монте-Карло выбирают первую спираль данной структуры, отбрасывают все спирали, с ней несовместимые, и из оставшихся выбирают случайную вторую и так до тех пор, пока есть совместимые с уже отобранными. Таким образом, строится определенное число структур, и для каждой спирали вычисляется ее частота встречаемости в этих структурах.

Далее спирали, встречающиеся менее чем в половине построенных случайных структур, из рассмотрения отбрасывают. Затем вычисляют произведение частоты встречаемости данной спирали на ее свободную энергию.

3. Величину произведения частоты и энергии использовали для отбора спиралей, которые образуются на данном этапе. Для этого выбирают спираль с максимальной величиной, отбрасывают несовместимые с ней, из оставшихся выбирают имеющую максимальную величину и т. д.

Этот процесс заканчивается тогда, когда либо исчерпывается количество спиралей, имеющих частоту, бо́льшую половины числа случайных структур, либо в том случае, когда количество отобранных спиралей достигает определенной величины, являющейся функцией числа спиралей в исходном наборе.

Эти блоки — основные части алгоритма, и после их выполнения в структуру строящейся РНК включают новые спирали. Затем предел энергии, определяющий набор спиралей, увеличивается (по абсолютной величине уменьшается), строится новый набор спиралей (естественно, совместимых с уже построенными) и выполняется новый цикл.

Общее количество спиралей является, в принципе, также функцией длины последовательности.

Программа, реализующая приведенный алгоритм, написана на языке СИ. Вычисления выполняли на ПК «Амстрад 1640».

Результаты и обсуждение. При оценке любого алгоритма расчета вторичной структуры РНК первым является вопрос: насколько результат расчета близок к структуре, наблюдающейся в эксперименте. Для проверки мы выбрали несколько РНК принципиально различной длины, для которых имеется наибольший объем экспериментальных данных. Это известный мидивариант РНК фага Q_β, в стабильности которого важную роль играет вторичная структура, длиной 218 нуклеотидов [9], а также 16S и 23S рибосомные РНК *Escherichia coli* длиной 1542 и 2904 нук-

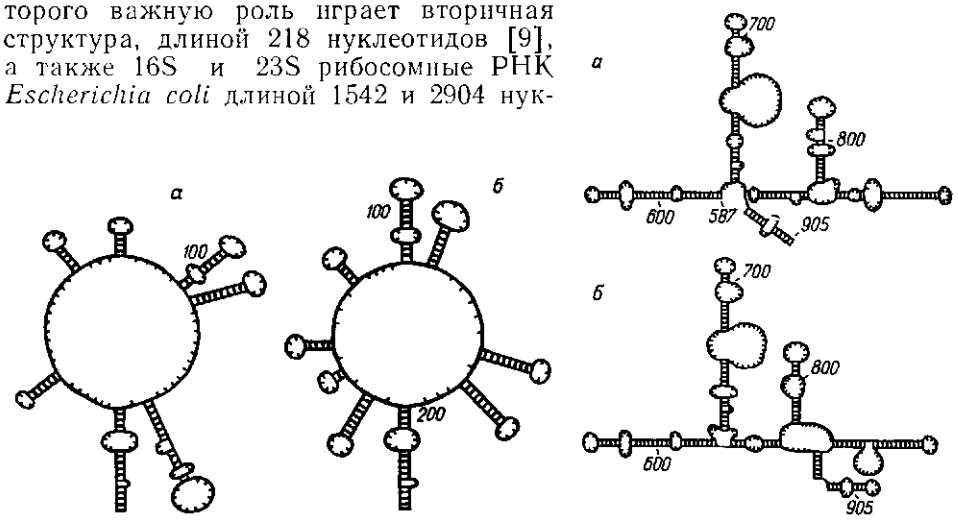

Рис. 2. Вторичная структура мидиварианта РНК фага Q_β, предсказываемая по предложенному алгоритму (*а*), и приведенная в обзоре [10] (*б*)

Fig. 2. The secondary structures of *Q* phage midivariant RNA, predicted by presented algorithm (*a*) and from review [10]

Рис. 3. Вторичная структура фрагмента 587—905 16S рРНК *E. coli* по предсказанию (*а*) и по модели [11] (*б*)

Fig. 3. The secondary structures of fragment 587—905 from *E. coli* 16S-rRNA: predicted (*a*) and in model [11] (*б*).

леотида соответственно. Следует сразу отметить, что модели вторичных структур этих молекул все же допускают некоторые разночтения.

Мидивариант РНК фага Q_β, по-видимому, может существовать в двух формах: закрытой, т. е. со спариванием 5′- и 3′-концов, и открытой [9]. Алгоритм, предложенный в нашей работе, с достаточно большой точностью дает закрытую форму (рис. 2). Видно, что за исключением пары спиралей, соединенных в одну, расчет соответствует данным, приведенным в литературе [9, 10].

Естественно, при увеличении длины исследуемой РНК точность расчета падает. Так, соответствие расчета вторичной структуры 16S рРНК *E. coli* уже составляет 55—60 % (в зависимости от модели, с которой проводится сравнение). Однако все же эта величина достаточно велика по сравнению с другими методами, которые проверяли на этой последовательности. Так, расчет, выполненный при работе с фрагментом 16S рРНК размером 574 нуклеотида, дает около 50 % соответствия даже при введении в программу данных экспериментов [2]. В нашем же случае программа, работающая с полной последовательностью, для достаточно протяженных участков дает результат, практически полностью совпадающий с моделями, выведенными из экспериментальных и филогенетических данных. Для примера на рис. 3 приведены структуры фрагмента 587—905 16S рРНК *E. coli* по модели [11] и по расчету.

Обращают на себя внимание некоторые особенности расчета по описываемому алгоритму. Программа практически полностью предсказывает спиральные участки, образующие шпилечные петли, и спирали,

образующие внутренние петли с первыми. Спирали, образующие разветвленные (сложные) петли, предсказываются хуже. Спирали, образующие шпилечные петли, при проведении расчета образуются раньше, и в том случае, если не задаваться целью построить полную вторичную структуру РНК, а выявить наиболее стабильные участки структуры, можно ограничиться только этими фрагментами. Точность такого расчета будет достигать 70—80 %.

При сравнении модели вторичной структуры 16S рРНК с результатом расчета видно также, что целый ряд участков, образующих спирали в модели, в расчете также образуют спирали, но другие. Не исключено, что этот результат демонстрирует некоторые альтернативные формы вторичной структуры [12], однако это может быть только предположением.

При расчете более длинных молекул точность падает и составляет для 23S рРНК *E. coli* [13] (2904 нуклеотида) 40—45 %. Здесь также более точно предсказываются структуры фрагментов, не образующие сложных петель, и если ограничиваться первым 25—30 спиралями в расчете, то точность составляет около 75 %.

Интересно, что точность расчета для длинных (более 2000 нуклеотидов) молекул РНК можно повысить, предположив, что сворачивание вторичной структуры начинается еще в процессе синтеза РНК. Программа по уже описанному алгоритму, но построенная таким образом, что определенное количество спиралей вычисляется из последовательно увеличивающихся по длине фрагментов общей РНК вплоть до полной длины, дает более высокую точность. Этот вопрос, по-видимому, требует дальнейшего специального исследования.

THE METHOD FOR RNA SECONDARY STRUCTURE CONSTRUCTION ON THE BASIS OF SELF-ORGANIZATION PRINCIPLES

A. P. Gultyaev, Yu. N. Monakov

All-Union Institute of Influenza,
Ministry of Public Health of the USSR Leningrad

Summary

The method for computer calculation of RNA secondary structure is described. The algorithm is based on successive formation of helices in the RNA self-organization process with the generation of random structure sets by the Monte-Carlo method. This way permits constructing RNA secondary structures for rather long sequences (up to 5000 nucleotides) by means of the personal computer without using large memory volumes.

СПИСОК ЛИТЕРАТУРЫ

1. *Wada A., Suyama A.* Local stability of DNA and RNA secondary structure and its relation to biological function // Progr. Biophys. and Mol. Biol.— 1986.— **47**.— P. 113—157.
2. *Zuker M., Stiegler P.* Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information // Nucl. Acids Res.— 1981.— **9**, N 1.— P. 133—148.
3. *Some* simple computational methods to improve the folding of large RNAs / A. B. Jacobson, L. Good, J. Simonetti, M. Zuker // Ibid.— 1984.— **12**, N 1.— P. 45—51.
4. *Колчанов Н. А., Соловьев В. В., Жарких А. А.* Контекстные методы теоретического анализа генетических макромолекул (ДНК, РНК и белков) // Итоги науки и техники.— М. : ВИНИТИ, 1985.— С. 6—37 (Молекуляр. биология; Т. 21).
5. *Martinez H. M.* An RNA folding rule // Nucl. Acids Res.— 1984.— **12**, N 1.— P. 323—334.
6. *Миронов А. А., Дьяконов Л. П., Кистер А. Э.* Предсказание ансамблей вторичных структур РНК. Кинетический анализ самоорганизации // Молекуляр. биология.— 1984.— **18**, № 6.— С. 1686—1694.
7. *Nussinov R., Pieczenik G.* Structural and combinatorial constraints on base pairing in large nucleotide sequences // J. Theor. Biol.— 1984.— **106**, N 3.— P. 245—259.
8. *Improved* free-energy parameters for predictions of RNA duplex stability / S. M. Freier, R. Kierzek, J. A. Jaeger et al. // Proc. Nat. Acad. Sci. USA.— 1986.— **83**, N 12.— P. 9373—9377.

9. *Mills D. R., Kramer F. R., Spiegelman S.* Complete nucleotide sequence of a replicating RNA molecule // Science.— 1973.— **180**, N 4089.— P. 916—927.
10. *Эйген М., Шустер П.* Гиперцикл. Принципы самоорганизации молекул.— М. : Мир, 1982.— 270 с.
11. *Detailed* analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids / C. R. Woese, R. Gutell, R. Gupta, H. F. Noller // Microbiol Rev.— 1983.— **47**, N 4.— P. 621—669.
12. *Williams A. L., Tinoco I.* A dynamic programming algorithm for finding alternative RNA secondary structures // Nucl. Acids Res.— 1986.— **14**, N 1.— P. 299—315.
13. *Gutell R. R., Fox G. E.* A compilation of large subunit RNA sequences presented in a structural format // Ibid.— 1988.— **16**, suppl.— P. 175—185.

ВНИИ гриппа МЗ СССР, Ленинград Получено 29.05.90

УДК 577.112

И. А. Жилкин, А. М. Ерошкин

МЕТОД ПОИСКА ФУНКЦИОНАЛЬНО ВАЖНЫХ ОБЛАСТЕЙ БЕЛКОВ И ПЕПТИДОВ ПО АМИНОКИСЛОТНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

Предлагается эмпирический метод предсказания функционально важных областей белков и пептидов. Рассматриваются только такие функционально важные участки, остатки которых располагаются близко в последовательности (линейные или непрерывные функционально важные участки). В основе метода лежит ранее обнаруженная корреляция между локализацией известных непрерывных функционально важных областей в последовательностях белков и низкими значениями профилей сходства этих последовательностей с последовательностями белков человека [1]. Применение предлагаемого метода к большому набору белков позволяет правильно предсказать более половины из известных непрерывных функциональных центров.

Введение. В работе [1] было предложено для анализа полипептидных последовательностей использовать профиль сходства последовательности исследуемого белка или пептида с последовательностями белков человека. Профиль строится следующим образом. Для исследуемого белка и каждого белка из выборки белков человека рассчитывается карта

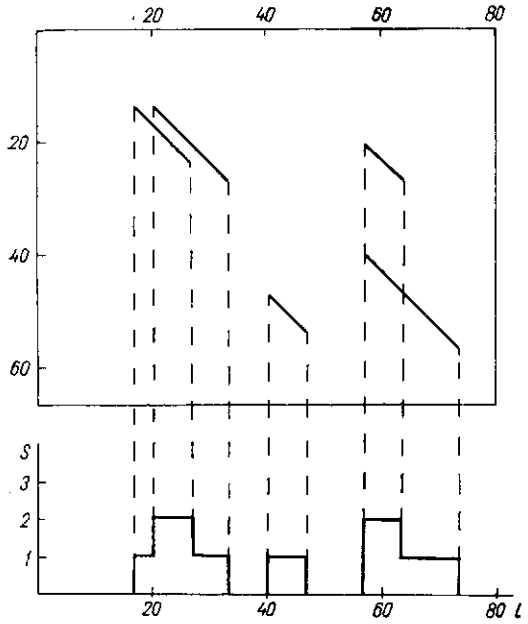

Рис. 1. Схема построения профиля сходства исследуемого белка с одним из белков выборки белков человека по карте локального сходства (горизонтальная ось соответствует исследуемому белку, вертикальная — белку выборки). Отрезки на карте сходства — совпадающие участки; внизу — соответствующий профиль сходства S, l — его длина

Fig. 1. Construction scheme of resemblance profile for investigated protein versus one protein from set of human proteins as based on the map of local resemblance (across — investigated protein down — protein from human protein set). Lines on the resemblance map — coinciding regions. Under the map is the resemblance profile S; l — investigated protein sequence length.