

**Модель эволюции нуклеотидной последовательности**

*Любецкий В. А., Жижина Е. А., Горбунов К. Ю.,  
Селиверстов А. В.*

lyubetsk@iitp.ru

Москва, Институт проблем передачи информации РАН

Задача эволюции предковой последовательности вдоль данного эволюционного дерева  $G$  широко изучается и относится к числу важнейших проблем биоинформатики. Эволюционирующая последовательность может быть, например, нуклеотидной, т.е. в алфавите из четырех букв  $\{A, C, T, G\}$ . Предполагается, что эти буквы заменяются друг на друга, и, кроме того, возможны еще два события: в случайное место последовательности *вставляется* некоторый участок в том же алфавите или *удаляется* такой участок. Таким образом, длина  $n$  эволюционирующей последовательности переменная.

В дереве  $G$  каждому  $j$ -му ребру приписана его длина  $t_j$ , которая интерпретируется как время эволюции вдоль этого ребра. Каждой позиции  $i$  от 1 до  $n$  сопоставляется скорость эволюции  $r_i$  в позиции  $i$ . Значение  $r_i$  обычно находится из гамма-распределения с последующим усреднением результата моделирования. Из биологии известны несколько вариантов конкретной матрицы  $R$ , которая управляет заменой одних букв на другие. Тогда эволюция последовательности букв задается простым правилом: буква в любой  $i$ -й позиции последовательности  $\sigma_j$ , сопоставленной началу  $j$ -го ребра, преобразуется в букву в той же позиции последовательности  $\sigma'_j$  в конце  $j$ -го ребра вероятностно как  $\exp(R \cdot r_i t_j)$ , где  $R$  — известная матрица соответствующего размера. Сама эволюционирующая последовательность называется еще *первичной структурой*, и в ней образуется так называемая *вторичная структура* — множество спиралей, где каждая спираль — это некоторое спаривание нуклеотидов по правилу  $G$  с  $C$  и  $T$  с  $A$ , см. [1]–[4].

*Фундаментальная проблема* биоинформатики состоит в моделировании эволюции вдоль данного дерева последовательности вместе с вторичной структурой в ней. Ниже предложена модель для описания такой эволюции. Как приложение этой модели, мы рассматриваем задачу построения *множественного выравнивания последовательностей с учетом их вторичной структуры*. А именно, в обозначениях, которые будут даны ниже, вторичная структура в концевых последовательностях  $\{\sigma_m\}$ , индуцированная минимальной конфигурацией  $\sigma^*$ , позволяет для исходных данных  $\{\theta_m\}$ , в которых вторичная структура заранее не была определена, получить множественное выравнивание с учетом этой индуцированной эволюционным процессом вторичной структуры. Соответствующий алгоритм будет изложен.

### Описание модели эволюции последовательности вместе с ее вторичной структурой

Дано дерево  $G$  эволюции нуклеотидной последовательности с заданными длинами ребер, и концевым вершинам дерева приписаны *современные последовательности*  $\{\theta_m\}$ . *Конфигурацией* называется некоторое произвольное сопоставление всем вершинам, включая концевые, последовательностей переменной длины  $n$ . Последовательности из данной конфигурации, которые сопоставляются концевым вершинам дерева, назовем *концевыми последовательностями*, и обозначим их  $\{\sigma_m\}$ . Ниже предлагается функционал  $H(\sigma)$ , аргументом которого является конфигурация  $\sigma$ , и минимум которого при значении аргумента  $\sigma^*$  соответствует эволюции предковой последовательности вдоль дерева вместе с вторичной структурой в ней.

Функционал выражает три условия на искомую конфигурацию  $\sigma$ :

- 1) для каждой последовательности  $\sigma_j$  и в каждой позиции  $i = 1, \dots, n$  любой последовательности из конфигурации происходит независимая замена букв вдоль любого ребра  $j$  согласно матрице замен  $R$ , как указано выше, и также вставка/удаление участков — слагаемое  $H_1(\sigma)$  ниже;
- 2) значения концевых последовательностей конфигурации близки к соответствующим современным последовательностям — слагаемое  $H_2(\sigma)$ ;
- 3) последовательности из конфигурации по возможности сохраняют вторичную структуру от начала ребра к его концу и вдоль целого пути в дереве, т. е. в течение многих поколений; при этом функционал меньше, если такие пути длиннее и их больше — слагаемое  $H_3(\sigma)$ .

Уточним, что *путем*  $\varphi$  называется путь в дереве  $G$ , вдоль ребер которого сохраняется высокая близость вторичных структур (как везде «близость» понимается в смысле некоторого фиксированного порога). *Длиной*  $l(\varphi)$  такого пути  $\varphi$  назовем число ребер в нем. *Временем*  $t(\varphi)$  пути  $\varphi$  назовем сумму длин  $t_j$  ребер вдоль  $\varphi$ . Далее  $\varphi$  везде *обозначает* путь в этом смысле.

Поскольку последовательности  $\sigma_j$  и  $\sigma'_j$ , приписанные соответственно началу и концу ребра  $j$ , имеют, вообще говоря, разные длины, то дальше предполагается, что для каждой конфигурации  $\sigma$  и каждого ее ребра  $j$  выполнено стандартное выравнивание последовательностей  $\sigma_j$  и  $\sigma'_j$ . Выровненные последовательности, которые теперь включают знак пробела, будем обозначать соответственно  $\delta_j$  и  $\delta'_j$ .

Нами предложен такой функционал  $H(\sigma)$  и стохастический алгоритм для поиска его глобального минимума, основанный на идее аннилинга.

А именно, положим  $H(\sigma) = (H_1(\sigma) + H_3(\sigma)) + H_2(\sigma)$ , где

$$H_1(\sigma) = \sum_j -\ln P(\sigma_j, \sigma'_j, t_j),$$

$j$  пробегает все ребра дерева  $G$ ;  $t_j$  — время, приписанное ребру  $j$ ;

$$P(\sigma_j, \sigma'_j, t_j) = \prod_i \exp(R \cdot r_i t_j)(\sigma_j, \sigma'_j) \cdot \prod_k \exp(-\varepsilon \ln(l_k + 1)),$$

где  $i$  пробегает все столбцы, в которых буква соответствует при выравнивании букве,  $k$  пробегает все максимальной длины участки в этом выравнивании типа, содержащие с одной из сторон только пробелы, и тогда  $l_k$  — длина такого участка,  $\varepsilon$  — параметр, отвечающий за значимость пробелов по сравнению с заменами букв. Далее

$$H_2(\sigma, \theta) = -\lambda \sum_{\sigma_m} \delta(\sigma_m, \theta),$$

где  $\lambda$  — параметр, отвечающий за соотношение двух слагаемых ( $H_1(\sigma) + H_3(\sigma)$ ) и  $H_2(\sigma)$ ;  $\sigma_m$  пробегает все концевые последовательности конфигурации  $\sigma$ , а  $\theta$  — современные последовательности; функция  $\delta$  начисляет некоторый штраф за расхождение по каждой позиции у последовательностей  $\sigma_m$  и  $\theta$ . Заметим, что вторичная структура в  $\theta$  не предполагается известной. Наиболее трудным является вопрос о том, как правильно записать слагаемое  $H_3(\sigma)$ . Представим его в виде суммы

$$H_3(\sigma) = -\left( H_0(\sigma) + k \sum_j U(Q(\sigma_j, \sigma'_j), t_j) \right).$$

Здесь первое слагаемое  $H_0(\sigma)$  — сумма по всем ребрам  $j$  сумм энергий всех спиралей в  $\sigma_j$  со значениями ниже некоторого порога. Это слагаемое отражает тот факт, что в искомой конфигурации  $\sigma$ , описывающей эволюцию, многие  $\sigma_j$  предполагаются со спиральями, имеющими низкую энергию. Второе слагаемое отражает сохранение вторичной структуры и, как следствие, длину и количество путей  $\varphi$ , о которых говорилось выше. Предполагается, что минимизация функционала с указанным выше  $H_3(\sigma)$  влечет минимизацию того же функционала с более сложным слагаемым

$$H'_3(\sigma) = -\left( H_0(\sigma) + k \sum_{\varphi} l(\varphi) \sum_{j \in \varphi} U(Q(\sigma_j, \sigma'_j), t_j) \right).$$

Сохранение вторичной структуры при переходе от  $\sigma_j$  к  $\sigma'_j$ , т. е. величина  $Q$ , описывается в [5]. Принимается  $U(Q(\sigma, \sigma'), t) = Q - \ln(1 + t/\mu)$ . В качестве алгоритма поиска глобального минимума была принята схема аннилинга для алгоритма Метрополиса.

### Обоснование алгоритма

В качестве алгоритма поиска глобального минимума  $\sigma^*$  указанного функционала  $H(\sigma)$  мы рассматриваем схему аннилинга на основе алгоритма Метрополиса-Хастингса, которая представляет собой марковскую цепь на пространстве всех конфигураций модели. Марковская цепь имеет стационарную гиббсовскую меру

$$P_m(\sigma) = \frac{\exp(-\beta_m H(\sigma))}{\sum_{\sigma} \exp(-\beta_m H(\sigma))}$$

при каждом фиксированном  $\beta_m$ .

**Теорема 1.** Если для параметра  $\beta_m$  выполняется условие  $\lim_{m \rightarrow \infty} \frac{\log m}{\beta_m} > \text{const}$ , то описанный выше алгоритм обладает следующим свойством:  $\lim_{m \rightarrow \infty} P\{\sigma(m) \in E_{\min}\} = 1$ , где  $E_{\min}$  — множество глобальных минимумов нашего функционала,  $\sigma(m)$  — конфигурация, возникшая к  $m$ -й итерации,  $P\{\cdot\}$  — распределение этой марковской цепи в момент времени  $m$ .

Работа поддержана РФФИ, проект № 07-01-00445, и МНТЦ 2766.

### Литература

- [1] Seliverstov A. V., Lyubetsky V. A. Translation regulation of intron containing genes in chloroplasts // Journal of Bioinformatics and Computational Biology. — 2006. — V. 4, № 4. — P. 783–793.
- [2] Lyubetsky V., Pirogov S., Rubanov L., Seliverstov A. Modeling classic attenuation regulation of gene expression in bacteria // Journal of Bioinformatics and Computational Biology. — 2007. — V. 5, № 1. — P. 155–180.
- [3] Vitreschak A. G., Mironov A. A., Lyubetsky V. A., Gelfand M. S. Functional and evolutionary analysis of the T-box regulon in bacteria // Genome Biology. — 2007. — in print.
- [4] Горбунов К. Ю., Любецкий В. А. Эволюция предковых регуляторных сигналов вдоль дерева эволюции фактора транскрипции // Молекулярная биология. — 2007. — Т. 41, в печати.
- [5] Горбунов К. Ю., Миронов А. А., Любецкий В. А. Поиск консервативных вторичных структур РНК // Молекулярная биология. — 2003. — Т. 37, № 5. — С. 850–860.