работка неизвестных морфологическому анализатору склоняемых слов, учет синтаксической модели, описывающей словоизменение распознанной конструкции.

Работа выполнена при поддержке РФФИ, проект № 05-01-00442а.

Литература

[1] *Александровский Д. А., Кормалев Д. А., Кормалева М. С., Куршев Е. П., Сулейманова Е. А., Трофимов И. В.* Развитие средств аналитической обработки текста в системе ИСИДА-Т // КИИ-2006. Труды конференции. — Т. 2. — М.: Физматлит, 2006. — С. 555–563.

[2] *Кормалев Д. А.* Автоматическое построение правил извлечения информации из текста // 1-я межд. конф. «Системный анализ и информационные технологии» САИТ–2005. — Т. 1. — М.: КомКнига, 2005. — С. 205–209.

[3] *Кормалев Д. А., Куршев Е. П.* Развитие языка правил извлечения информации в системе ИСИДА-Т // Межд. конф. «Программные системы: теория и приложения», ИПС РАН, Переславль-Залесский, октябрь 2006 г. — Т. 1. — М.: Физматлит, 2006. — С. 365–377.

[4] *Кормалев Д. А.* Обобщение и специализация при построении правил извлечения информации // Конф. КИИ-2006. — Т. 2.— М.: Физматлит, 2006. — С. 572–579.

[5] *Леонтьева Н. Н., Семенова С. Ю.* Инструменты построения фрейма «ПЕРСОНА» // НТИ, Сер. 2. Информ. процессы и системы. — 2001. — № 8.

[6] *Сулейманова Е. А.* Классификация ресурсов знаний в системе извлечения информации из текста // ММРО-13 (наст. сб.). — 2007. — С. 625–628.

[7] *Appelt D. E.* The Common Pattern Specification Language: Technical report. — SRI International, Artificial Intelligence Center, 1996.

[8] *Appelt D. E., Israel D. J.* Introduction to Information Extraction. Tutorial // 16th Int'l. Joint Conf. on Artificial Intelligence IJCAI'99, Sweden, 1999.

[9] *Grishman R.* TIPSTER Text Architecture Design. Version 3.1. — New York: NYU, 1998.

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 σ_j, сопоставленной началу j-го ребра, преобразуется в букву в той же позиции последовательности σ'_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 σ^*, 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 конфи-

гурация σ, и минимум которого при значении аргумента σ^* соответствует эволюции предковой последовательности вдоль дерева вместе с вторичной структурой в ней.

Функционал выражает три условия на искомую конфигурацию σ:

1) для каждой последовательности σ_j и в каждой позиции $i = 1, \ldots, n$ 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* φ 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 φ назовем число ребер в нем. *Временем* $t(\varphi)$ пути φ назовем сумму длин t_j ребер вдоль φ. Далее φ везде *обозначает* путь в этом смысле.

Поскольку последовательности σ_j и σ'_j, приписанные соответственно началу и концу ребра j, имеют, вообще говоря, разные длины, то дальше предполагается, что для каждой конфигурации σ и каждого ее ребра j выполнено стандартное выравнивание последовательностей σ_j и σ'_j. Выровненные последовательности, которые теперь включают знак пробела, будем обозначать соответственно δ_j и δ'_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 вы-

равнивании типа, содержащие с одной из сторон только пробелы, и тогда l_k — длина такого участка, ε — параметр, отвечающий за значимость пробелов по сравнению с заменами букв. Далее

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

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

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

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

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

Сохранение вторичной структуры при переходе от σ_j к σ'_j, т. е. величина Q, описывается в [5]. Принимается $U\big(Q(\sigma,\sigma'),t\big) = Q - \ln(1+t/\mu)$. 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 собоймарковскую цепь на пространстве всех конфигураций модели. Марковская цепь имеет стационарную гиббсовскую меру

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

при каждом фиксированном β_m.

Теорема 1. *Если для параметра β_m выполняется условие $\lim\limits_{m\to\infty} \frac{\log m}{\beta_m} > \text{const}$, то описанный выше алгоритм обладает следующим свойством: $\lim\limits_{m\to\infty} \mathsf{P}\{\sigma(m) \in E_{\min}\} = 1$, где $E_{\min}$ — множество глобальных минимумов нашего функционала, $\sigma(m)$ — конфигурация, возникшая к m-й итерации, $\mathsf{P}\{\cdot\}$ — 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.

Средства OLAP-моделирования и их применение в задачах здравоохранения

Ноженкова Л. Ф.

expert@icm.krasn.ru

Красноярск, Институт вычислительного моделирования СО РАН

Технология оперативной аналитической обработки многомерных данных OLAP (On-line Analytical Processing) считается одним из разделов интеллектуального анализа данных. Аналитические OLAP-модули все чаще появляются в составе отечественных и зарубежных продуктов и финансово-производственных приложений. Существо аналитической обработки сводится к автоматизированной поддержке формирования аналитических запросов, агрегированию данных, операциям над многомерным кубом данных с использованием плоских представлений — кросс-таблиц, кросс-диаграмм, картограмм. Наибольшее применение технология OLAP