Математические модели эволюции и регуляции экспрессии генов

В.А. Любецкий
Москва

Модель эволюции генов

Модель основана на: построении филогенетического дерева генов, а затем и дерева видов бактерий и вычислениях связанных с ними параметров на основе молекулярных данных; и также на анализе гипотез об эволюционных событиях молекулярного уровня у бактерий.

К упомянутым событиям, прежде всего, относятся: дивергенция генов в процессе дивергенции видов, дупликация генов, потеря и приобретение генов, горизонтальный перенос генов. Молекулярные данные – это наборы белковых последовательностей, объединенных по сходству и по выполнению определенной функции, и, в первую очередь, комплексы ортологических групп белков. Общая схема реконструкции эволюционных событий молекулярного уровня такова: отбираются белковые семейства и решается задача построения их множественных выравниваний, затем задача построения соответствующего филогенетического дерева генов G. Дальнейший анализ основан на использовании сходства (при построении дерева видов S) и различия (при анализе эволюционных событий) между многими филогенетическими деревьями генов $\{G_i\}$ и этим S. А именно, при построении дерева видов S различия в топологиях белковых деревьев G_i подавляются и находится «консенсусное» для них дерево. При анализе эволюционных событий, напротив, ищутся существенные различия в топологии отдельного белкового дерева G (часто G это одно из деревьев того же семейства $\{G_i\}$) и уже построенного дерева видов S. Для объяснения этих различий рассматриваются математические модели эволюции генов, а события эволюционной истории семейства микроорганизмов реконструируются путем оптимизации параметров этих моделей. А именно, в качестве модели эволюции берется так или иначе организованное сравнение деревьев белков и видов, а в качестве параметров – множества вершин этих деревьев и отнесенных к ним эволюционных событий. Оптимизация состоит в том, что ищутся такие значения этих параметров, на которых характеристики эволюции принимают экстремальные значения.

Модель регуляция экспрессии генов

Предлагается модель, в первую очередь, классической РНКовой регуляции экспрессии генов с помощью прерывания (терминации) процесса транскрипции. Модель опирается на представление о макросостоянии вторичной структуры в регуляторной области РНК между рибосомой и полимеразой, на формулы резонансного типа, определяющие

величину замедления РНК-полимеразы набором шпилек в той же области, на представления о процессах посадки и последующего движения рибосомы и полимеразы. Специальное внимание уделяется подбору параметров модели. Для проверки модели проведено компьютерное моделирование и получены, в частности, зависимости вероятности терминации транскрипции от величины концентрации загруженных тРНК и от концентрации аминокислоты в клетке или в культуре для многих регуляторных областей в геномах бактерий (здесь данные приводятся для четырех стрептомицетов) и при различных значениях трех параметров, которые рассматриваются как основные. Полученные зависимости согласуются с доступными экспериментальными данными; в том числе, по форме графиков, относящихся к активности фермента в зависимости от концентрации аминокислоты (например, атранилат синтазы от триптофана в культуре у *S. venezuela*).

Белок-ДНКовая регуляция транскрипции, как и секвестор трансляции, находят отражение в предлагаемой модели, но их подробные разработки будут представлены в другой статье.

В дальнейшем на основе нашей модели предполагается получить предсказания о влиянии точечных «мутаций» в регуляторных областях геномов на результат аттенюаторной регуляции, включая предсказания об эволюционной устойчивости организмов. А затем – включить эту модель в более широкую модель регуляции и метаболизма у бактерий. Другое возможное использование: сейчас аттенюаторная регуляция предсказывается обычно на основе множественного выравнивания, для этого требуется несколько последовательностей; получение с помощью модели на индивидуальной последовательности характерной для аттенюации или ее отсутствия кривой при подходящих параметрах могло бы рассматриваться как аргумент в пользу наличия или отсутствия аттенюации.

При построении этих моделей используются математические методы, в том числе распознавание образов. Указанных модели характеризуются весьма высокой вычислительной сложностью. Поэтому нами были разработаны специальные приемы построения быстрых алгоритмов (полиномиальных невысокой степени вместо естественно возникающих здесь сверхэкспоненциальных), и использовались параллельные вычислительные архитектуры, включая специальную организацию памяти при вычислениях.

Литература

1. Vitreschak A.G., Lyubetskaya E.V., Shirshin M.A., Gelfand M.S., Lyubetsky V.A. Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis // FEMS Microbiol Lett. 2004 May 15; 234(2), p. 357-370.

2. Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria // BMC Microbiology, 2005, accepted in press.

Масштабируемые алгоритмы классификации текстов

А.В. Максаков
(Москва)

Введение

Несмотря на то, что проблема классификации текстов достаточно хорошо изучена на данный момент, достаточно актуальным остается прикладной аспект теории машинного обучения. Известно множество алгоритмов, обеспечивающих сравнительно высокую точность классификации, в частности, метод опорных векторов[1] (SVM). Однако их высокая вычислительная сложность и ресурсоемкость делает проблематичным их использование в задачах, требующих обработки большого количества документов в процессах обучения и классификации. По этой причине актуальным представляется поиск масштабируемых и сравнительно быстрых алгоритмов классификации, точность которых была бы сравнима с вышеуказанными алгоритмами.

Модификация наивного алгоритма Байеса

В последнее время в литературе точность наивного алгоритма Байеса оценивается как одна из самых низких среди рассматриваемых алгоритмов. Тем не менее на практике он используется достаточно часто, причиной этому служит его простота и высокая производительность. У алгоритма есть несколько систематических проблем. В частности, он основан на принципе независимости признаков, также точность классификации существенно падает при наличии неравномощных обучающих выборок.

Вероятность принадлежности документа классу в алгоритме определяется следующим образом:

$$p(C \mid d') = \frac{p(d' \mid C)p(C)}{\sum_{C \in c} p(d' \mid C)p(C)} = \frac{p(C)\prod_{w \in d'} (p_{Cw})^{f_w}}{\sum_{C \in c} p(C)\prod_{w \in d'} (p_{Cw})^{f_w}}$$

где f_w-количество вхождений лексемы w в документ, $p_{Cw} = p(w \mid C)$

Прологарифмировав обе части и убрав общие для всех классов слагаемые получаем следующее правило определения класса для документа: