

УДК 575.852.112:575.852.113:004.273

КОМПЬЮТЕРНАЯ СИСТЕМА АНАЛИЗА МОЛЕКУЛЯРНОЙ ЭВОЛЮЦИИ ГЕНОВ И БЕЛКОВ: СООТНЕСЕНИЕ МОЛЕКУЛЯРНОЙ ЭВОЛЮЦИИ С ЭВОЛЮЦИЕЙ ФЕНОТИПИЧЕСКИХ ПРИЗНАКОВ ОРГАНИЗМОВ

К.В. Гунбин¹, М.А. Генаев¹, Д.А. Афонников^{1,2}, Н.А. Колчанов^{1,2}

¹Институт цитологии и генетики Сибирского отделения РАН, г. Новосибирск;

²Новосибирский государственный университет, г. Новосибирск;

e-mail: genkvg@bionet.nsc.ru)

Создана компьютерная система анализа молекулярной эволюции генов и белков SAMEM, основанная на анализе отношения частот фиксации радикальных аминокислотных замен к консервативным (K_R/K_C) и исследовании скоростей изменения физико-химических свойств аминокислот в эволюции белков (V_P). Для анализа используются все известные физико-химические свойства аминокислот и производится статистическое соотнесение изменения этих свойств с фенотипическими признаками организмов. Это позволяет облегчить интерпретацию подходов K_R/K_C и V_P . SAMEM доступна по адресу <http://pixie.bionet.nsc.ru/samem/>

Ключевые слова: молекулярная эволюция, радикальные и консервативные замены аминокислот, система клиент-сервер.

В последнее время адекватность использования отношения частот фиксации несинонимических замен к синонимическим для определения режима эволюции генов поставлена под сомнение [1, 2]. В качестве альтернативы предлагаются анализ отношения частот фиксации радикальных аминокислотных замен к консервативным (K_R/K_C) [3] и анализ скоростей фиксации различных аминокислотных замен в эволюции белков (V_P) [4]. Недостатком этих подходов является априорная классификация замен по свойствам или типам аминокислот на радикальные или консервативные. В пакете TREESAAP 3.2 [5] эта проблема частично решается учетом всех известных физико-химических свойств аминокислот. В качестве нулевой гипотезы TREESAAP 3.2 использует предположение о равновероятной частоте всех типов аминокислотных замен, что в эволюции белков часто не выполняется [6]. Наш подход, как и TREESAAP, учитывает все известные физико-химические свойства аминокислот, но учитывает неравенство частот замен аминокислот аналогично методу [4]. Кроме того, для соотнесения эволюционных изменений физико-химических свойств аминокислот с эволюцией фенотипических признаков организмов мы используем статистики Спирмена и Кендалла, что позволяет выявить взаимосвязь между особенностями эволюции генов и фенотипов организмов.

Материалы и методы

Система SAMEM (<http://pixie.bionet.nsc.ru/samem/>) состоит из двух основных конвейеров, анализа эволю-

ции генов и анализа эволюции белков, и одного дополнительного, позволяющего пользователю формировать выборки данных, используя NETBLAST 2.2.18. Конвейеры представляют собой наборы вычислительных модулей (программ), запускаемых в консольном режиме в среде Linux. Основные конвейеры позволяют провести общие этапы анализа (многократное выравнивание и построение эволюционной модели аминокислотных замен, построение филогенетического дерева, реконструкция предковых последовательностей в каждом узле филогенетического дерева), комбинируя различные методы. Многократное выравнивание рассчитывается программами MAFFT 6.717 и KALIGN 2.04, филогения программами FASTTREE 2.1.1 и PHYML 3.0. Эволюционная модель аминокислотных замен оценивается при помощи MODELESTIMATOR 1.1. Предковые последовательности белков реконструируются ANCESCON, FASTML (серверная версия) и CODEML (из пакета PAML 4.4), генов — ANC-GENE, FASTML (серверная версия) и CODEML (из пакета PAML 4.4). Для анализа используются данные о 531 свойстве аминокислот [7]. Общий модуль оценки взаимосвязи изменений физико-химических свойств аминокислот с фенотипическими признаками реализован на языке R с использованием пакета Stats (R 2.10.1).

Конвейер анализа генов использует дополнительно этап трансляции генов в последовательности белков для дальнейшего анализа, что увеличивает точность выравнивания. Пользователь может задать число групп для разделения 20 аминокислот. Для каждого из 531 свойства аминокислоты группиру-

ются методом К-средних (пакет Stats, R 2.10.1). Для анализа режима эволюции генов используется отношение K_R/K_C (программа HON-NEW [3]).

При анализе последовательностей белков существенной особенностью является доработанный нами метод анализа скоростей фиксации различных аминокислотных замен. Он основан на Марковском моделировании эволюции белков пакетом INDELible 1.02 [8] и непараметрическом перестановочном тесте, предложенном нами в работе [9]. Отметим, что INDELible 1.02 позволяет учесть детальные особенности эволюции белков (различие частот встречаемости аминокислот и частот их замен у паралогичных групп внутри одного семейства белков вплоть до учета особенностей эволюции на отдельных ветвях дерева) [8]. Тест с перестановками [9] позволяет выявить для каждой ветви филогенетического дерева характерные типы аминокислотных замен, частота которых статистически значимо превышает частоту в смоделированных последовательностях. Кроме того, в конвейере анализа эволюции белков реализована возможность суммарного анализа эволюционных изменений физико-химических свойств аминокислот. Для этого использованы коррекции эволюционных расстояний Кимуры, Джукса—Кантора и Гамма.

SAMEM создан на основе клиент-серверной конвейерной системы обработки данных BioinfoWF. Серверная часть системы реализована в виде приложения на языке Perl и отвечает за запуск и выполнение конвейера. На вход приложению подается схема конвейера и описание вычислительных модулей. Схема конвейера, форматы данных и взаимодействие вычислительных модулей в ней описывается на языке XML. Описание конвейера состоит из двух файлов. Первый описывает вычислительные модули, второй задает топологию конвейера. Описания модулей состоят из следующих разделов: 1) входные файлы; 2) выходные файлы; 3) параметры и опции (идентификатор, описание, тип параметра, значение по умолчанию, внешний вид поля запроса значения для параметра на странице web-браузера); 4) правила генерации командной строки; 5) правила действия пользовательского интерфейса, позволяющие динамически изменять web-интерфейс модуля в процессе редактирования конвейера. Второй файл задает топологию конвейера. В нем описываются имена входных/выходных файлов для каждого модуля и значения для каждого параметра или опции. Приложение запускает конвейер, создавая файл с отчетом о статусе выполнения каждого узла в формате XML. Серверная часть поддерживает параллельный запуск узлов конвейера и позволяет производить удаленный запуск ресурсоемких расчетных модулей на вычислительном кластере. Режим запуска узла (локальный или удаленный) задается в схеме конвейера.

Клиентская часть реализована в виде web-приложения. Пользователю предлагается работать с уже готовыми схемами. При этом доступны базовые опции управления конвейером: 1) установка входных файлов, параметров и опций для каждого вычислительного модуля в конвейере; 2) старт с произвольного узла и остановка на произвольном узле в конвейере; 3) отслеживание статуса выполнения каждого вычислительного модуля в конвейере; 4) просмотр входных/выходных файлов для каждого этапа расчета в конвейере; 5) привязка форматов входных/выходных файлов к различным приложениям для их визуализации. Генерация страниц в формате HTML реализована на Perl с использованием библиотеки HTML::Template. Дополнительно используется JavaScript библиотека jQuery.

Тестирование системы

Тестирование работоспособности системы производилось на выборке генов и белков лизоцимов приматов, представленной в работе [10]. Результаты тестирования доступны через Интернет (<http://pixie.bionet.nsc.ru/samem/test.html>). Результаты согласуются с опубликованными ранее в работе [10] и дополнены многочисленными данными об особенностях изменений отдельных физико-химических свойств аминокислот. Особенности эволюции генов и белков могут быть соотнесены с данными о фенотипах приматов, что позволяет уточнить интерпретацию, предложенную в работе [10]. Тестирование также проведено на белках Notch-каскада сигналов (Serrate, Notch, Delta, Fringe) многоклеточных животных, белках суперсемейства TIR1/AFB растений.

Выводы

Создана компьютерная система SAMEM для анализа режимов эволюции генов и белков. В основе метода лежит учет всех известных физико-химических свойств аминокислот. Уникальной особенностью системы является возможность напрямую соотносить вычисленные особенности молекулярной эволюции с эволюцией заданных фенотипических признаков организмов, и, таким образом, интерпретировать результат.

* * *

Работа поддержана Российским фондом фундаментальных исследований (грант № 09-04-01641-а); Интеграционными проектами СО РАН № 113, 119; Программами РАН № 22 (проект № 8), 23 (проект № 29).

СПИСОК ЛИТЕРАТУРЫ

1. *Drummond D.A., Wilke C.O.* Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution // *Cell*. 2008. Vol. 134. N 2. P. 341—352.
2. *Huang Y., Koonin E.V., Lipman D.J., Przytycka T.M.* Selection for minimization of translational frameshifting errors as a factor in the evolution of codon usage // *Nucleic Acids Res.* 2009. Vol. 37. N 20. P. 6799—6810.
3. *Zhang J.* Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes // *J. Mol. Evol.* 2000. Vol. 50. N 1. P. 56—68.
4. *Pupko T., Sharan R., Hasegawa M., Shamir R., Graur D.* Detecting excess radical replacements in phylogenetic trees // *Gene*. 2003. Vol. 319. P. 127—135.
5. *Woolley S., Johnson J., Smith M.J., Crandall K.A., McClellan D.A.* TreeSAAP: selection on amino acid properties using phylogenetic trees // *Bioinformatics*. 2003. Vol. 19. N 5. P. 671—672.
6. *Henikoff S., Henikoff J.G.* Amino acid substitution matrices from protein blocks // *Proc. Natl. Acad. Sci. U.S.A.* 1992. Vol. 89. N 22. P. 10915—10919.
7. *Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M.* AAIindex: amino acid index database, progress report 2008 // *Nucleic Acids Res.* 2008. Vol. 36. Database issue. P. D202—D205.
8. *Fletcher W., Yang Z.* INDELible: a flexible simulator of biological sequence evolution // *Mol. Biol. Evol.* 2009. Vol. 26. N 8. P. 1879—1888.
9. *Gunbin K.V., Afonnikov D.A., Kolchanov N.A.* Molecular evolution of the hyperthermophilic archaea of the *Pyrococcus* genus: analysis of adaptation to different environmental conditions // *BMC Genomics*. 2009. Vol. 10. P. 639.
10. *Messier W., Stewart C.B.* Episodic adaptive evolution of primate lysozymes // *Nature*. 1997. Vol. 385. N 6612. P. 151—154.

Поступила в редакцию
15.04.2010

**COMPUTER SYSTEM FOR ANALYSIS OF MOLECULAR EVOLUTION
MODES OF PROTEIN-CODING GENES (SAMEM): RELATION OF MOLECULAR
EVOLUTION WITH THE PHENOTYPICAL FEATURES OF ORGANISMS**

K.V. Gunbin, M.A. Genaev, D.A. Afonnikov, N.A. Kolchanov

A web-available computer system for analyzing the molecular evolution modes of genes and proteins was developed (SAMEM: <http://pixie.bionet.nsc.ru/samem/>). SAMEM computations are based on the ratio of radical to conservative amino acid replacements, on the rate of amino acid changes in the course of protein evolution, and on the statistical relations between evolutionary change of all known amino acid properties and user defined phenotypical features of organisms.

Key words: *molecular evolution, radical to conservative amino acid replacements, client-server system.*

Сведения об авторах

Гунбин Константин Владимирович — канд. биол. наук, науч. сотр. Института цитологии и генетики СО РАН. Тел. (383)363-49-23; e-mail: genkvg@bionet.nsc.ru

Генаев Михаил Александрович — аспирант, Институт цитологии и генетики СО РАН. Тел. (383)363-49-22; e-mail: mag@bionet.nsc.ru

Афонников Дмитрий Аркадьевич — канд. биол. наук, ст. науч. сотр., зав. сектором эволюционной биоинформатики Института цитологии и генетики СО РАН, преподаватель НГУ. Тел. (383)363-49-23; e-mail: ada@bionet.nsc.ru

Колчанов Николай Александрович — докт. биол. наук, акад. РАН, директор Института цитологии и генетики СО РАН, проф., зав. кафедрой информационной биологии факультета естественных наук НГУ. Тел. (383)363-49-80; e-mail: kol@bionet.nsc.ru