Секция «Биоинформатика»

Разработка чувствительного метода поиска ортологов длинных некодирующих РНК

Научный руководитель – Миронов Андрей Александрович

Мыларщиков Дмитрий Евгеньевич
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия
E-mail: dmitrymyl@gmail.com

При анализе некодирующих РНК (англ. ncRNA) конкретного организма, особенно длинных некодирующих РНК (англ. lncRNA), нередко возникает задача поиска ортологов этих РНК в других организмах. Ввиду того, что последовательности генов ncRNA эволюционируют быстрее последовательностей белок-кодирующих генов [6], поиск ортологов прямым выравниванием ncRNA и генома другого организма с расслабленными параметрами зачастую даёт много ложноположительных результатов. Для повышения отношения сигнала к шуму используется дополнительная информация: вторичная структура РНК[4], факт экспрессии[5] или косвенные свидетельства этого и синтения[2]. Каждый из этих подходов позволяет не проводить большое количество выравниваний, но обладает своими недостатками. Так, далеко не для всех lncRNA вторичная структура может играть функциональную роль[3], а значит, неконсервативна; нередко экспрессия lncRNA тканеспецифична[5] и нет доступа к информации об экспрессии генов в разных тканях другого организма; может пропасть сигнал синтении[11]. Для исследования *de novo* определённых РНК кажется логичным применять поиск по синтении, так как он не требует дополнительной информации о самих РНК и их возможных ортологах.

Мы разработали программный пакет ortho2align (https://github.com/dmitrymyl/ortho2align), который совмещает поиск ортологов по синтении и статистическую оценку значимости блоков выравнивания. Для определения синтеничных областей между геномом запроса и геномом поиска используется liftOver[1] при наличии парного полногеномного выравнивания или якорные гены `[U+2060]`— ортологи белок-кодирующих генов из базы OrthoDB[9]. Попарные выравнивания синтеничных областей осуществляются blastn[10] с расслабленными параметрами для детекции даже слабой схожести последовательности. Каждый блок выравнивания каждой РНК проходит статистическую проверку на случайность согласно фоновой модели случайных выравниваний РНК против генома поиска. Прошедшие порог по q-value блоки выравнивания объединяются в одну цепочку для каждой РНК методом динамического программирования. В случае нескольких ортологов одной РНК выбирается один ортолог с наибольшей суммарной длиной блоков выравнивания.

На подвыборке ортологичных и неортологичных lncRNA между человеком и мышью из [5] алгоритм показал чувствительность до 90% и специфичность до 60% в зависимости от параметров алгоритма. Ложноположительные результаты возникали по двум причинам: 1) наличие коротких участков консервативности; 2) наличие протяжённых участков консервативности от начала и до конца неортологичных lncRNA. Вторая причина может указывать на истинную ортологию, которую не смогли определить ранее (ввиду низкой экспрессии или непрохождения через статистические фильтры).

Применение пакета ortho2align к полувыделяемым РНК[7] позволило предсказать около 200 *de novo* консервативных РНК на уровне Млекопитающих, что может указывать на их консервативную функцию в качестве архитектурых РНК ядерных телец[8].

Работа поддержана грантом РФФИ № 20-04-00459 А.

Источники и литература

1) W. J. Kent и др., «The Human Genome Browser at UCSC», Genome Res., т. 12, вып. 6, сс. 996–1006, янв. 2002.
2) J. Chen и др., «Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs», Genome Biology, т. 17, вып. 1, 2016.
3) E. Rivas, J. Clements, и S. R. Eddy, «A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs», Nature Methods, т. 14, вып. 1, Art. вып. 1, янв. 2017.
4) J. S. Pedersen и др., «Identification and Classification of Conserved RNA Secondary Structures in the Human Genome», PLOS Computational Biology, т. 2, вып. 4, с. e33, апр. 2006.
5) A. Necsulea и др., «The evolution of lncRNA repertoires and expression patterns in tetrapods», Nature, т. 505, вып. 7485, сс. 635–640, 2014.
6) K. C. Pang, M. C. Frith, и J. S. Mattick, «Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function», Trends in Genetics, т. 22, вып. 1, сс. 1–5, янв. 2006.
7) T. Chujo и др., «Unusual semi-extractability as a hallmark of nuclear body-associated architectural noncoding RNAs», The EMBO Journal, т. 36, вып. 10, сс. 1447–1462, 2017.
8) T. Chujo и T. Hirose, «Nuclear Bodies Built on Architectural Long Noncoding RNAs: Unifying Principles of Their Construction and Function», Mol. Cells, т. 40, вып. 12, с. 889, 2017.
9) E. V. Kriventseva и др., «OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs», Nucleic Acids Research, 2019.
10) S. F. Altschul, W. Gish, W. Miller, E. W. Myers, и D. J. Lipman, «Basic local alignment search tool», Journal of Molecular Biology, 1990.
11) H. Hezroni, D. Koppstein, M. G. Schwartz, A. Avrutin, D. P. Bartel, и I. Ulitsky, «Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species.», Cell reports, т. 11, вып. 7, сс. 1110–22, 2015.