Задачи распознавания структурной организации последовательностей биополимеров

Назипова Н. Н.
nnn@impb.ru
Пущино, Институт математических проблем биологии РАН

Молекулы биополимеров в основе своей имеют линейные цепочки мономерных звеньев, которые называются в случае молекулы ДНК нуклеотидами, в случае белковых молекул — аминокислотами. Такое представление молекул биологических полимеров называют их первичной структурой. В настоящее время благодаря успешному развитию методов расшифровки первичных структур молекул биополимеров появилось большое количество геномов.

Геномом называется наследственный материал клетки. Физически — это ДНК, содержащаяся в одном наборе хромосом организма, логически — это совокупность всех генов, последовательностей, регулирующих их экспрессию, а также фрагментов с неизвестной пока функцией. Геном представляется в виде непрерывной последовательности нуклеотидов. Геном участвует во всех генетических процессах клетки. Наследственность есть результат дупликации генома, мутации основаны на выпадениях, вставках и заменах отдельных нуклеотидов или небольших участков, действие генов — на синтезе РНК и белков.

Самые первые большие задачи, которые были поставлены при изучении структурной организации геномов, была проблема поиска различных известных коротких подпоследовательностей, которые играют важную роль в жизни молекулы, а также задача поиска генов. Эти задачи лежали на поверхности, их решением уже больше десятка лет занимаются многие коллективы ученых.

Нас заинтересовала проблема реконструкции эволюции геномов. Предполагается, что основными путями эволюционного развития является мутационная активность, когда геномы постоянно находятся в состоянии дуплицирования фрагментов, точечных замен, вставок и выпадений символов.

Периодически повторяющиеся фрагменты символов самой разной длины — так называемые тандемные повторы, в последовательностях ДНК и белков встречаются достаточно часто. Иногда с ними связаны конкретные функции (связывание субстрата, межбелковые взаимодействия), пространственная структура, или характерные свойства, например, такие как гибкость, белков или отдельных участков ДНК. Короткие тандемные повторы от 2 до 6 нуклеотидов — микросателлиты используются в регуляции генов или формируют районы полиморфизма длины в геноме, на основании которых происходит идентификация лично-

сти или родства. Более длинные повторы, получившие название VNTRs (variable number of tandem repeats) представляют потенциальные сайты рекомбинации. Тандемные повторы можно отнести к структурной периодичности символьных последовательностей. Структурная периодичность в ДНК и белках подвержена эволюционной дивергенции, и со временем теряет вид совершенных повторов. Так что нельзя бывает однозначно выделить причины возникновения несовершенной периодичности: были ли это дупликации отдельных фрагментов или конвергенция последовательностей, направляемая отбором на достижение лучших функциональных свойств.

Анализ распределения в полных геномах организмов различного вида повторов (совершенных и несовершенных, прямых и инвертированных, тандемных и разнесенных) даст ответы на вопросы, связанные с путями эволюции геномов. Предполагается, что у каждого типа повторов свои механизмы эволюции. Нами разработаны различные методы исследования повторяющихся структур в генетических последовательностях. Один из них [1] основан на статистических подходах, этот метод может быть использован для выявления районов однородности геномов.

Метод поиска тандемных повторов [2] основан на спектральном разложении специализированного функционала, определенного на генетической последовательности, этот метод может использоваться для более детального исследования выявленных статистическим методом участков генома и нахождения на нем несовершенных тандемных повторов (т. н. сателлитов), определения их точных характеристик, таких, как длина паттерна и кратность повтора. Одним из выгодных отличий нашего метода от существующих является отсутствие ограничений на длины исследуемых структур.

Кроме того, разрабатываются методы для поиска разнесенных и инвертированных повторов, а также поиска шпилечных структур. Последняя задача является началом новой работы по распознаванию на геномах эукариот нового класса генов — участков кодирования микроРНК.

Нами создано соответствующее программное обеспечение, с помощью которого уже получены интересные данные по ряду организмов. По результатам работы создаются базы данных, где будут представлены результаты обработки геномов круглого червя Caenorhabditis elegans, ряда видов болезнетворной бактерии Staphylococcus, а также генома человека. Эти базы данных будут выставлены в Интернете для бесплатного доступа.

Работа поддержана РФФИ, проекты № 06-07-89274, № 06-01-08039.

Литература

[1] *Чалей М. Б., Назипова Н. Н., Кутыркин В. А.* Совместное использование различных критериев проверки однородности для выявления скрытой периодичности в биологических последовательностях // Математическая биология и биоинформатика. — 2007. — Т. 2, № 1. — С. 20–35. `www.matbio.org/downloads/Chaley2007(2\_20).pdf`.

[2] *Дедус Ф. Ф., Куликова Л. И., Махортых С. А., Назипова Н. Н., Панкратов А. Н., Тетуев Р. К.* Аналитические методы распознавания повторяющихся структур в геномах // Доклады РАН. — 2006. — Т. 411, № 5. — С. 599–602.