АНАЛИЗ СТРУКТУРНЫХ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
2005 год Выпуск 174
УДК 519.95+577.2

ТАКСОНОМИЯ ТРАЕКТОРИЙ И ВОССТАНОВЛЕНИЕ СТРУКТУРЫ ГЕННЫХ СЕТЕЙ[1]

И.А. Борисова, Н.Г. Загоруйко,
А.В. Ратушный, В.А. Тихошвай, Н.А. Колчанов

Введение

Генная сеть представляет собой набор скоординированно работающих и взаимно регулируемых генов, обеспечивающих выполнение определенной функции в организме. К числу структурных компонентов генных сетей относятся сами гены, кодируемые ими РНК и белки, пути передачи сигналов, ведущие от клеточных рецепторов к регуляторным районам генов и т.д.

Состояние генной сети в фиксированный момент времени может быть описано величинами концентраций химических веществ, входящих в нее. Портрет динамики поведения этой системы в период активности может быть представлен в виде набора одномерных кривых, описывающих изменение концентраций этих элементов. Сейчас существуют методики, позволяющие проводить соответствующие измерения и получать описание

[1]Работа выполнена при поддержке гранта фонда «Научный потенциал» 23 03 9; при финансовой поддержке Минобразования России, грант А03-2.8 85; грантов РФФИ 02 04 48802, 02-07-90359, 03 04 48506, 03-07-96833; грантов ОМН РАН 2003 1.4.3.; СО РАН (Интеграционный проект 119); ФХБ РАН 10.4; Национального Института Здоровья США № 2 R01 HG-01539-04A2.

генной сети в виде набора траекторий, отражающих изменения во времени концентраций химических веществ, являющихся ее структурными компонентами.

Потенциально в таком описании заложен большой объем информации о взаимодействии различных элементов генной сети. Было бы интересно, например, обнаружить группы элементов с приблизительно одинаковой динамикой поведения во времени. Решить эту задачу можно методами таксономии, приспособленными для таксономии множества одномерных динамических траекторий. Получаемые таксоны будут объединять такие элементы, у которых совпадают периоды подъема и спада их активности. Эти результаты будут полезны при решении ряда задач молекулярной биологии: построение структуры генной сети, понимание механизма ее функционирования, обнаружение мутаций и распознавание типа этих мутаций.

В данной работе объектом таксономического анализа являются траектории, отражающие динамику работы элементов эритроидной генной сети. Данные получены на машинной модели, имитирующей работу этой сети и представляющей собой систему из 90 дифференциальных уравнений [1].

1. Методы таксономии динамических траекторий

Существует несколько подходов к задаче таксономии динамических траекторий. В самом простом случае траектории, наблюдаемые в N последовательных моментов времени, рассматриваются как обычные N-мерные вектора. Расстояния между ними вычисляются при помощи обычных метрик (например, Евклидовой), и к ним применяются стандартные методы таксономии.

Однако в данной задаче этот подход оказывается неприемлемым, потому что он ориентируется на абсолютные значения концентраций, которые для разных веществ отличаются друг от друга на девять порядков. Так как интерес представляют не величины концентраций, а характер изменения их динамики, то в начале делается линейная нормировка величин концентраций, которая приводит данные о каждом элементе к диапазону от 0 до 1. В результате этого после проведения таксономии в одной

группе окажутся траектории с похожим поведением динамик, а не с близкими абсолютными значениями в каждый момент времени.

Имеется еще третий подход, при котором методом динамического программирования процессы сдвигаются и растягиваются друг относительно друга с целью получения максимального сходства [2]. Он позволяет выделить группы динамик, порождаемых одинаковым процессом, но протекающим в зависимости от обстоятельств с разной скоростью в разные моменты времени. В данном случае предполагается, что все элементы генной сети работают в одинаковых для них стационарных условиях, что делает применение динамического программирования нецелесообразным.

2. Таксономия динамических траекторий для эритроидной генной сети

В этой работе исследовались данные о концентрациях 34 веществ, измеренных на протяжении 100 часов работы эритроидной генной сети в норме и при возникновении 20 типов мутаций различной сложности — от одной до четырех мутаций одновременно. Под мутацией здесь понимается некоторое внешнее или внутреннее воздействие, нарушающее нормальное функционирование некоторого элемента генной сети и, как следствие, меняющее работу всей сети в целом.

Предварительный анализ показал, что среди траекторий изменения концентраций различных веществ можно увидеть такие, которые похожи по форме, но сильно отличаются своими средними значениями. Есть траектории с похожими средними значениями, но различные по форме. В связи с этим делалось три варианта таксономии: группировка веществ по похожести их формы; группировка по похожести средних значений; группировка по взвешенной похожести средних значений и формы.

Таксономия делалась с помощью алгоритмов семейства FOREL [2]. Программы были адаптированы к особенностям таксономии траекторий. Наибольший интерес представляли результаты таксономии, при которой объединялись кривые с похожей динамикой поведения во времени (т.е., таксономия по форме при

нормировке концентраций по их максимальным значениям). Данные для таксономии представляли собой таблицу, состоящую из 714 строк (34 вещества*21 тип мутаций) и 100 столбцов (концентрации веществ в 100 моментов времени). Мера похожести одного объекта на другой определяется через Евклидово расстояние между их 100 мерными векторами.

Количество таксонов зависит от задаваемого радиуса таксона. Задача выбора оптимального числа таксонов в таксономии сама по себе не имеет идеального решения. Так что, в каждой конкретной ситуации приходится рассматривать несколько вариантов, которые иногда в дальнейшем используются параллельно. На данной таблице при малых радиусах получается много таксонов, большинство из которых содержит по одному или два объекта. Назовем такие мелкие таксоны "единичными". В более крупных таксонах можно выделить типичные объекты, т.е. такие, сумма расстояний от которых до всех остальных объектов данного таксона минимальна. Количественные характеристики некоторых вариантов таксономии показаны в таблице. Здесь же указаны типичные представители каждого из не "единичных" таксонов — название вещества и (в скобках) номер мутации.

Радиус таксономии R=3 Число таксонов = 30 Крупных таксонов 11			Радиус таксономии R=5 Число таксонов = 19 Крупных таксонов 9		
№	Кол-во объектов	Типичный объект	№	Кол-во объектов	Типичный объект
1	40	EPORJak2(8)	1	45	EPROJak2(8)
2	11	AG(15)	2	23	BG(2)
3	21	Hemoglob(15)	3	21	Hemoglob(15)
4	11	TfRTfout(2)	4	550	mRNAGATA(8)
5	7	Fe_in(15)	5	12	TfRTfout(15)
6	547	RNAGATA(8)	6	7	Fe_in(15)
7	25	Proto_IX(20)	7	28	Proto_IX(11)
10	7	Fe_in(17)	10	5	Fe_in(17)
11	9	AG(9)	11	9	Proto_IX(6)
12	8	Proto_IX(6)			
28	3	AG(19)			

На рис.1 и 2 представлены формы траекторий типичных представителей.

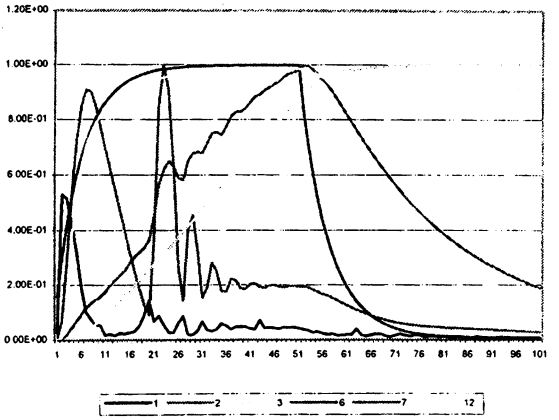

Рис.1. Формы типичных траекторий для разных таксонов при радиусе $R = 3$.

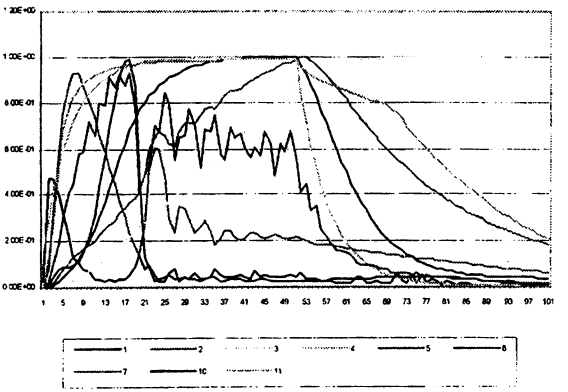

Рис.2. Формы типичных траекторий для различных таксонов при радиусе $R = 5$.

3. Результаты таксономии

Можно предположить, что в случаях, когда при очень больших радиусах почти все вещества попадают в один общий таксон, а несколько веществ (один-два) составляют "единичные" таксоны, то эти вещества либо непосредственно подвержены мутации, либо напрямую связаны с веществом, подвергшимся мутации. Так, при радиусе $R = 5$ для первой мутации такими веществами оказались AG и TfRTfout; для четвертой — Heme, AG, BG, Fe_in и Proto_IX; для седьмой — Fe_in; для восьмой — BG и EPOR; для одиннадцатой — Fe_in; для тринадцатой — Heme; для девятнадцатой — AG; для двадцатой — BG.

Можно видеть, что для четвертой мутации, в которой веществом-мутантом было Fe_in, среди четырех "единичных" таксонов оказались такие вещества:

1) Heme,
2) AG, BG,
3) Fe_in,
4) Proto_IX,

что подтверждает нашу гипотезу. Она подтвердилась также и для девятнадцатой и двадцатой мутаций: вещества-мутанты оказались в составе "единичных" таксонов. При других (главным образом, двойных) мутациях в составе "единичных" таксонов веществ-мутантов не было. Вещества, оказавшиеся в таких "единичных" таксонах, вероятно, находятся в сильной зависимости от веществ-мутантов.

Аналогичное предположение можно выдвинуть и по поводу веществ, которые в большинстве случаев принадлежат одному таксону и лишь при определенной мутации — к другому. Такое поведение наблюдается у EPOR и EPORJak2 во второй мутации, у mRNAALAS и ALAS2 в третьей мутации, у IRP_ALAS в десятой, у EPOR, EKLFactr, mRNAEPOR, mRNAEKLF в восьмой и у mRNAPPO в восемнадцатой. Высказанное предположение подтвердилось для второй, восьмой, десятой и восемнадцатой мутаций. При малых радиусах таксонов ($R = 0.01$) выделилось 135 таксонов: 25 из них содержали больше двух элементов и максимальный размер таксона — 247 элементов. На основе этой, более чувствительной таксономии, была предпринята попытка

определить вещества, имеющие сходную природу и(или) механизмы функционирования. Для этого выделялись вещества, которые в большинстве мутаций совместно попадали в одни и те же таксоны (были близки в пространстве результатов таксономии). Ими оказались такие группы веществ:

1) AG и BG;

2) ALAS2 и mRNAALAS;

3) CPO, FCH, mRNAALAD, mRNAPBGD, mRNACPO, mRNAFCH, mRNATAL1, mRNAHOXB;

4) mRNAEPOR и mRNAEKLF;

5) mRNAAG и mRNABG;

6) HOXBtrfc и EKLFactr;

7) mRNAIRP, mRNAUROS, mRNAUROD, mRNAPPO.

Но в большинстве случаев в один таксон попадало одно и тоже вещество при различных мутациях. Об этих веществах можно предположить, что они слабо зависят от исследованных типов мутаций: между элементами генной сети, отвечающими за выработку данного вещества и элементами, подвергнутыми мутации, существенной связи нет.

Есть вещества, которые при разных мутациях входят в составы разных таксонов. К таким веществам относятся, в частности, те, которые были выбраны в качестве наиболее информативных при распознавании типов мутаций [3].

4. Таксономия 34-х веществ

Следующий вариант таксономии был сделан на таблице данных, состоящей из 34 строк (объектов) и 2100 столбцов (признаков). Эти объекты были получены путем "склеивания" траекторий одного и того же вещества при каждом из 21 состояния генной сети. При радиусе таксонов равном 7 выделилось 25 таксонов, среди которых более одного объекта имело 5 таксонов. Состав этих таксонов был таким:

1) TfRTfout и Fe_in;

2) ALAS2 и mRNAALAS:

3) EKLFactr, mRNAEPOR и mRNAEKLF;

4) mRNAALAD, mRNACPO, mRNAFCH, mRNATAL1 и mRNAHOXS;

5) mRNAUROS и mRNAUROD.

При радиусе таксонов равном 14 выделились 20 таксонов, среди которых было 7 не единичных.

Множество объектов (траекторий) в пространстве своих характеристик обнаруживает некоторую классификационную структуру. При определенных радиусах удается выделить несколько сгустков (таксонов) похожих друг на друга траекторий и выделить типичных представителей этих таксонов.

5. Восстановление структуры связей между элементами генной сети

В настоящее время достаточно полное описание создано для небольшого числа генных сетей. В большинстве случаев имеется лишь информация о генах, входящих в состав сети, но данных о структурных связях между ними нет. В данной работе делается попытка решить задачу восстановления структуры генной сети по наблюдениям за концентрациями веществ, вырабатываемых отдельными генами. При этом используются данные наблюдений за работой генной сети в норме и при нокаутировании отдельных генов.

В этой задаче исследуются взаимодействия только между генами, и соответственно наблюдение ведется только за концентрациями веществ (МРНК и белков), характеризующих уровень активности этих генов. Поочередно каждый ген нокаутируется, в результате чего прекращается выработка соответствующего МРНК и белка, что приводит к изменениям в работе всей генной сети.

Предварительные исследования возможности решения задачи в такой постановке проводились на модели генной сети по дифференцировке эритроцитов. В рамках этой сети рассматривались 20 генов, среди которых было выделено 4 гена, и требовалось выявить: для каких остальных генов выделенные 4 гена играют роль транскрипционных факторов. Анализируемая информация представляла собой траектории концентраций 20 МРНК и 20 белков, измеренных на протяжении 100 часов с интервалом в один час. Траектории нормировались по своим максимальным значениям, после чего проводилась таксономия этих

траекторий. В результате все траектории делились на группы (таксоны) в зависимости от характера их динамики. Кроме номера таксона каждой траектории приписывалась ее нормировочная величина, характеризующая среднюю активность гена на рассматриваемом временном интервале. Из большого числа вариантов таксономии траекторий, полученных при разных значениях диаметра таксонов, был выбран вариант, обладающей наибольшей информативностью.

Дальнейший анализ проводился с использованием такого понятия, как сила воздействия, которая определялась тем, на сколько изменилась активность того или иного гена в результате того или иного нокаута по сравнению с его работой в нормальном состоянии сети. Для каждого гена выделялась группа генов, на которые его нокаут оказал наиболее сильное воздействие. Затем проверялось, нет ли в этой группе гена, через который воздействие передавалось остальным. Другими словами, проверялось насколько правдоподобна гипотеза о прямом воздействии нокаутированного гена на выделенную группу генов. В случае обнаружения кандидата на роль посредника, эксперимент с тем же нокаутом повторялся, но этот посредник изолировался от всех внешних воздействий: его МРНК нарабатывалась с помощью конститутивного синтеза. В результате из первоначально сформированной группы убирались те гены, сила воздействия на которые в последнем эксперименте резко ослабевала. Эти гены связаны с нокаутированным геном не на прямую, а через посредника. Из первоначального полного графа возможных связей этих генов с нокаутированным геном удалялись ребра прямой связи. Структура генной сети становилась более определенной.

Как оказалось, результаты такого рода обработки позволяют определить транскрипционные факторы в рассматриваемой сети с высокой точностью. Кроме того, исследовалось возможность решения сходной задачи, наблюдая за поведением генной сети без нокаутов, а лишь с использованием изоляции выделенного гена от внешних воздействий. Выяснилось, что при таком подходе данные о поведении динамических траекторий не несут достаточной информации для восстановления связей в данной генной сети.

Дальнейшее развитие методов восстановления структуры генных сетей возрастающей сложности ведется с использованием разработанных в ИЦиГ СО РАН машинных моделей, имитирующих работу гипотетических генных сетей.

Выводы

В исходном виде каждое состояние сети описывается большим количеством характеристик (3400 чисел) и смысл такого описания практически недоступен для непосредственного восприятия человеком. Описание на языке результатов таксономии становится более прозрачным. Теперь о каждом веществе при заданной мутации можно сказать, что оно принадлежит такому-то таксону, следовательно, по форме траектории похоже на типичного представителя этого таксона (см. рис.1 и 2), а среднее значение концентрации этого вещества равно такой-то величине. По поводу конкретной мутации можно сказать, что при этой мутации такие-то вещества (попавшие в один большой таксон) не меняют своего поведения, а такие-то (оказавшиеся в "единичных" таксонах) реагируют на данную мутацию сильно, что указывает на то, что данное вещество либо само подверглось мутации, либо имеет прямую связь с веществом-мутантом.

Результаты таксономии могут оказаться полезными также и при изучении разнообразия поведения отдельных элементов генной сети в норме и при той или иной мутации. В частности, модельные эксперименты показали, что ее можно с успехом использовать при решении задачи построения структурной схемы генной сети по реакциям ее отдельных элементов на определенные мутации.

Литература

1. RATUSHNY A.V., PODKOLODNAYA O.A.,ANANKO E.A., LIKHOSHVAI V.A. Mathematical model of critroid cell differentiation regulation// Proc. of the 2nd Int. Conf. on Bioinformatics of Genome Regulation and Structure. – Novosibirsk, 2000. – Vol.1. – P.203–206.

2. ЗАГОРУЙКО Н.Г. Прикладные методы анализа и знаний. – Новосибирск: Изд.ИМ СО РАН, 1999. – 270 с.

3. BORISOVA I.A., ZAGORUIKO N.G., LIKHOSHVAI V.A., RATUSHNY A.V., KOLCHANOV N.A. Diagnostics of mutations based on analysis of gene networks// Proc. of the Third Int. Conf. on Bioinformatics of Genome Regulation and Structure. – Novosibirsk, 2002. – Vol.2. – P.163–165.

Поступила в редакцию
6 января 2004 года