**Предсказание пре-миРНК**

Для предсказания последовательности пре-миРНК человека мы строим контекстно-структурную марковскую модель. Пре-миРНК человека представляют собой нуклеотидные последовательности, образующие вторичную структуру в форме шпильки (рисунок 1а).

С учетом вторичной структуры шпильку пре-миРНК можно интерпретировать как парные последовательности, где в каждой позиции две характеристики: символ из алфавита {A, C, G, U, -} и структурное свойство (позиция находится в петле или в спирали). Последовательность начинается с 5′ края пре-миРНК.



Рисунок 1. Представление шпилечной структуры пре-миРНК в виде последовательности пар, где каждая пара содержит информацию о структуре и последовательности предшественника. а) Шпилечная структура пре-миРНК *hsa-let-7a-1* с двумя зрелыми микроРНК *hsa-let-7a-5p* и *hsa-let-7a-3p*. b) Парные последовательности, соответствующие шпильке пре-миРНК. c) Схема контекстно-структурной марковской модели.

Структура модели включает пять состояний: A, C, G, U и X и одно начальное состояние S. Конечное состояние в модели не задается. Состояния A, C, G, U соответствуют символам нуклеотидов. Состояние Х соответствует символам “-”, которые уравнивают число символов между ветвями шпильки посредством вставки напротив нуклеотидов боковых или ассиметричных внутренних петель (рисунок 1a, b).

Полученная модель может порождать два типа символов для каждого состояния: T - для символов, имеющих комплементарную пару во вторичной структуре в рассматриваемом состоянии, F - для символов, не имеющих комплементарную пару (рисунок 1b). В состоянии Х порождается только символ “-”. Все допустимые переходы между состояниями модели показаны на рисунке 1c.

Определим путь $π=π\_{1}π\_{2}…π\_{L}$ - последовательность наблюдаемых состояний модели, в которых порождается последовательность наблюдений $x=x\_{1}x\_{2}…x\_{L}$.

Пусть $T\_{km}=P(π\_{l+1}=m|π\_{l}=k)$ – вероятность перехода из состояния *k* в состояние *m*, где $π\_{l}$ - *l*-е состояние в пути π. В частном случае $T\_{0m}$ - вероятность перехода из начального состояния S в состояние $π\_{1}=m$.

Пусть $E\_{l}(b)=P(x\_{l}=b|π\_{l}=k)$ – вероятность порождения символа *b* в состоянии *l*.

Вероятностная модель λ определяется набором параметров {*Т, Е*}, где $T=\{T\_{km}\}$ - матрица, состоящая из вероятностей переходов между состояниями, $E=\{E\_{l}(b)\}$ – матрица, состоящая из вероятностей порождения символов. Параметры модели {*T, E*} оцениваются с помощью метода, предложенного Баумом с коллегами [1], на основе информации об известных пре-миРНК человека – их последовательностях и структурах.

Пусть модель λ = {*Т, Е*} порождает последовательность $x=x\_{1}x\_{2}…x\_{L}$ в состояниях модели $π=π\_{1}π\_{2}…π\_{L}$. Используя вероятности переходов T и вероятности порождения символов E можно вычислить полную вероятность порождения моделью последовательности $x$ в состояниях модели $π$:

$$P(x)=\prod\_{n=1}^{L}T\_{π\_{n-1}π\_{n}}E\_{π\_{n}}(x\_{n})$$

, где L – длина последовательности, а $π\_{0}$= S - стартовое состояние модели.

При увеличении длины последовательности величина вероятности *P(x)* экспоненциально уменьшается. Так как длины пре-миРНК достигают 150 нуклеотидов, в вычислениях используется нормированная на длину L величина:

$K= -\frac{log\_{10}P\left(x\right)}{L}= -\frac{\sum\_{n=1}^{L}\left(log\_{10}\left(E\_{π\_{n}}\left(x\_{n}\right)\right)+log\_{10}\left(T\_{π\_{n-1}π\_{n}}\right)\right)}{L}$ (1)

Поиск пре-миРНК в геномной последовательности начинается с выделения подпоследовательностей, способных формировать вторичную структуру в форме шпильки и удовлетворяющих условиям на длину, G+C контент, величины е-score и энергии. Так как программа для расчёта структур GArna [2] основана на стохастическом алгоритме, для каждого скользящего окна длиной *L*=120 нт и шагом 10 нт строится три варианта шпилечных структур с разными начальными рандомизационными параметрами. Для получения шпилечных структур в алгоритм программы GArna добавлены штрафы за мульти-петли и принудительное связывание концов последовательности в спирали. Из трех структур выбирается оптимальная структура с наименьшей энергией - кандидат в пре-миРНК.

По первичной и вторичной структурам кандидатов однозначно восстанавливается последовательность наблюдений $x=x\_{1}x\_{2}…x\_{120}$ и состояния модели $π=π\_{1}π\_{2}…π\_{120}$, в которых последовательность порождается. Кандидат классифицируется как пре-миРНК, если значение *K* (формула 1) для него меньше заданной величины порога. Величина порога задается пользователем и определяется из анализа графика оценки качества предсказания метода.

Среди предсказанных пре-миРНК с совпадающими шпилечными петлями (расстояния между шпилечными петлями меньше 5нт) в результат оставляется кандидат с наименьшей энергией.

**Предсказание микроРНК**

Для поиска гомологов известных микроРНК применяются фильтры по нуклеотидному контенту и алгоритм выравнивания Нидлмана-Вунша [3] со штрафом -1 за вставку/удаление/замену. Выбираются только те гомологи, расстояние Хэмминга между которыми и микроРНК не больше 1.

Метод предсказания *ab initio* микроРНК основан на парной скрытой марковской модели, предложенной авторами программы ProMiR [4]. Метод ProMiR предсказывает границы дуплекса микроРНК-микроРНК\* и определяет из двух последовательностей одну зрелую микроРНК.

Шпилька пре-миРНК интерпретируется в виде последовательности пар символов от основания шпильки до ее шпилечной петли (рисунок 1а, 2а). Соответствующая последовательности парная скрытая марковская модель (СММ) содержит четыре структурных состояния - M (спираль), N (двухстороння петля), I (вставка) и D (удаление), и два скрытых состояния – m (позиция предшественника в микроРНК) и n (позиция вне микроРНК).



Рисунок 2. Представление шпилечной структуры пре-миРНК в виде парной последовательности структурных и скрытых состояний, каждое из которых содержит информации о последовательности пре-миРНК и о расположении микроРНК в предшественнике. а) Шпилечная структура пре-миРНК hsa-let-7a-1 с двумя зрелыми микроРНК hsa-let-7a-5p и hsa-let-7a-3p. b) Последовательность состояний СММ, соответствующая 5р микроРНК в предшественнике. c) Структура скрытой марковской модели.

Обозначим СММ как λ = {*Т, Е*}. $T^{'}=T^{⊥}$ - транспонированная матрица вероятностей перехода между состояниями модели.

Для каждой позиции парной последовательности вычисляются величины:

$S\left(i\right)=\frac{P\_{m}\left(i-1\right)T\_{mn}}{P\_{m}\left(i-1\right)T\_{mn}+P\_{n}\left(i-1\right)T\_{nn}}$ и $S'\left(i\right)=\frac{P\_{m}\left(i+1\right)T\_{mn}^{'}}{P\_{m}\left(i-1\right)T\_{mn}^{'}+P\_{n}\left(i-1\right)T\_{nn}^{'}}$

,где *Pm(i)* и *Pn(i)* – вероятности позиции *i* находиться в области микроРНК или вне микроРНК, соответственно, вычисляются итерационно через параметры модели [4]; *Tnm* – вероятность перехода из скрытого состояния *n* в состояние *m*, *Tnn* – вероятность перехода из скрытого состояния *n* в состояние *n*; величины *S(i)* и *S′(i)* - численно характеризуют позиции границ микроРНК у шпилечной петли и у основания шпильки: чем больше значения величин – тем вероятнее расположение границ в позициях.

Основные отличия нашего подхода от работы [4] заключаются в следующем.

При обучении скрытых марковских моделей и при предсказании микроРНК человека мы предполагаем, что длина последовательностей составляет 24 нуклеотида против наблюдаемых 22 нуклеотидов: граница микроРНК, ближайшая к шпилечной петле предшественника, смещается нами на два нуклеотида. После предсказания микроРНК граница корректируется.

Для предсказания зрелых микроРНК мы используем не одну, а две СММ, по одной для 3′ и 5′ ветвей предшественника, что обусловлено частичной (неполной, со смещением) комплементарностью микроРНК в дуплексе микроРНК-микроРНК\*. Так как многие пре-миРНК порождают две зрелые микроРНК, в процессе предсказания нами определяется дуплекс микроРНК-микроРНК\*, а не одна последовательность.

Процедура предсказания состоит из поиска границы, ближайшей к шпилечной петле предшественника, и с определения по локальному максимуму второй границы, на расстоянии 24±2 нт от первой. Это обусловлено тем, что граница у петли шпильки более ярко выражена с точки зрения вторичной структуры (увеличение числа петлевых нуклеотидов) и лучше предсказывается.

Наконец, результатом предсказания являются оптимальная и субоптимальная пары микроРНК, которые соответствуют глобальному и следующему за ним максимумам функции *S(i)* для границы у шпилечной петли.

**Предсказания сайтов связывания микроРНК-мРНК**

Реализованные в представленной работе алгоритмы предсказания сайтов связывания микроРНК-мРНК основываются на двух известных принципах - полной комплементарности позиций № 2-8 в 5′-крае микроРНК к 3′ нетранслируемой области мРНК [5] и термодинамической стабильности дуплекса микроРНК-мРНК [6].

Метод «5’ context search» ищет в исходной последовательности мРНК 7 нт подпоследовательности, которые полностью комплементарны инвертированному участку в позициях № 2-8 5′-края микроРНК.

Метод поиска «Free energy calculation» вычисляет дуплекс микроРНК-мРНК с наименьшей энергией связывания с помощью метода полного перебора с ограничениями в виде размера петель (50нт) и порогом свободной энергии. Энергия связывания получается путем суммирования энергии участков полной комплементарности последовательностей, вычисленной по четырем нуклеотидам, участков внутренних и боковых петель. В вычислениях предполагается, что участки непрерывного взаимодействия взаимно не пересекаются, один нуклеотид не может вступать во взаимодействие с несколькими нуклеотидами.

Метод поиска «Partition function» вычисляет среднюю населенность последовательностями микроРНК каждой позиции большой РНК последовательности. Метод основан на вычислении статистических сумм дуплексов микроРНК-мРНК. Результатами расчёта являются вероятности образования дуплексов в зависимости от позиции мРНК и с учётом конкуренции микроРНК за сайт связывания. Последнее обстоятельство может оказаться существенным при предсказании множественных сайтов связывания, которые обеспечивают ингибирование трансляции мРНК при значительных концентрациях микроРНК.

Каждая последовательность микроРНК имеет свой статистический вес в каждой координате *k* последовательности мРНК *S*. Статистический вес *Wk(S)* определяется через энергию микроРНК-мРНК дуплекса *Ek*:

, где T – температура (константа).

Допустимой конфигурацией *K* для последовательности мРНК называется один из возможных способов корректного (без взаимных пересечений) размещения микроРНК на последовательности мРНК. Допустимую конфигурацию можно определить, как множество позиций *K={k}* в последовательности мРНК, соответствующих началам микроРНК из конфигурации. Понятие допустимой конфигурации исключает варианты одновременного связывания нескольких микроРНК с одним сайтом.

Статистический вес каждой допустимой конфигурации *K* для последовательности мРНК *S* рассчитывается как сумма статистических весов всех микроРНК конфигурации:

.

Пусть *N* – длина последовательности мРНК *S*, *m* – длина последовательности микроРНК.

τ – величина концентрации микроРНК в каждой позиции мРНК. При малых величинах параметра концентрации результаты расчетов близки к результатам поиска термодинамически наиболее стабильных микроРНК-мРНК дуплексов.

Вероятность расположения микроРНК, начиная с координаты *i*, равна сумме статистических весов всех допустимых конфигураций, в которых микроРНК располагаются с координаты i, деленое на сумму статистических весов всех допустимых конфигураций.

Эти суммы вычисляются в три этапа с применением алгоритма динамического программирования.

«Прямой проход». Вычисление F1, …, Fi,…, FN - суммы статистических весов всех допустимых конфигураций для подпоследовательностей мРНК S1,…,Si,…, SN, где Si – последовательность длиной *i* начиная с 5′ края мРНК, SN – вся мРНК.

F0 = 1;

Fi = Fi-1 для 1 ≤ *i* ≤ *m*;

Fi = Fi-1 + Fi-m-1×τ×P(Si-m), для *i* ≥ *m+1*; где , а *q* = 10-6 – нормировочный коэффициент для исключения арифметического переполнения при расчётах.

“Обратный проход”. Вычисление R1,…, Ri,…, RN – суммы статистических весов всех допустимых конфигураций подпоследовательностей мРНК S1′,…, Si′,…, SN′, где Si′ – последовательность длиной *i* начиная с 3′ края мРНК, S1′ – вся мРНК.

RN = 1;

Ri = Ri+1 для *i* ≥ *N-m+1*;

Ri = Ri+1 + Ri+m+1×τ×P(Si′), для *i* ≤ *N-m*;

“Финальный проход”. Вычисление *P(i)* – вероятность расположения последовательности микроРНК начиная с координаты i в мРНК:

 для *i* ≤ *N-m-1*.

Числитель представляет собой сумму статистических весов всех допустимых конфигураций, в которых микроРНК располагаются с координаты i в мРНК, знаменатель – это сумма статистических весов всех допустимых конфигураций для мРНК.

Тогда средняя населенность микроРНК позиции i в мРНК вычисляется как:

 для *i* ≤ *N-m-1*.

Программа демонстрирует график вероятности связывания микроРНК с мРНК для каждой позиции мРНК.

Ссылки:

1. Rabiner L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.
2. Титов И.И., Воробьев Д.Г., Иванисенко В.А., Колчанов Н.А. (2002). Быстрый генетический алгоритм для анализа вторичной структуры РНК. *Известия Академии Наук, Серия химическая,* **7**, 1047-1056.
3. Needleman S.B., Wunsch C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443-453.
4. Nam J.W., Shin K.R., Han J., Lee Y., Kim V.N., Zhang B.T. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Research*. **33**, 3570–3581.
5. Stark A., Brennecke J., Russell R.B., Cohen S.M. 2003. Identification of Drosophila microRNA targets. *PLoS Biol*, **1**, E60.
6. Lewis B.P., Burge C.B., Bartel D.P. (2004). Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*. **120**, 15–20.