Хрупкость мотива по отношению к мутационному процессу

Воронцов И.Е.

December 21, 2017

Мы рассмотрим мутационный процесс, вносящий однонуклеотидные замены, действующий на (ансамбле) всех сайтов, задаваемых мотивом. В качестве модели мотива мы используем матрицу частот $M^i_{\alpha_i}$. Строго говоря, мы не ограничиваемся мотивами сайтов связывания; это могут быть и сайты другой природы.

1 Мотив

По построению матрицы частот (и при условии независимости отдельных позиций)

$$P(w|w \text{ is a site}) = \prod_{i=1}^{n} M^i_{w_i}$$

Фактически, матрица частот описывает всё множество сайтов и показывает, какой сайт там встречается с какой частотой.

Введём также вероятность того, что тринуклеотидный контекст позиции i в сайте будет $\alpha\beta\gamma$:

$$M^i_{\alpha\beta\gamma} = M^{i-1}_{\alpha} M^i_{\beta} M^{i+1}_{\gamma}$$

2 Мутационный процесс

В нашей модели мутационный процесс $f^{\delta}_{\alpha\beta\gamma}$ задаётся мутационной подписью, т.е. распределением частот тринуклеотидных мутаций $\alpha\beta\gamma \to \alpha\delta\gamma$ (или кратко $\alpha\beta\gamma \to \delta$).

Мы полагаем, что мутационный процесс не может сохранить нуклеотид (этот случай мы рассматриваем отдельно как “мутация не произошла”):

$$f^{\beta}_{\alpha\beta\gamma} = 0$$

Также мы полагаем, что мутации не имеют предпочтений по нити ДНК, так что

$$f^{\delta}_{\alpha\beta\gamma} = f^{\delta'}_{\gamma'\beta'\alpha'},$$

где α' - это нуклеотид, комплиментарный α.

Мутационный процесс задаёт частоты, следовательно суммируется к единице:

$$\sum_{\alpha,\beta,\gamma,\delta} f^{\delta}_{\alpha\beta\gamma} = 1$$

3 Действие мутационного процесса на ансамбль сайтов, характеризуемый мотивом

Чтобы оценить влияние мутационного процесса на ансабль сайтов, мы построим (тоже в форме частотной матрицы) множество сайтов, каким оно будет при условии, что на исходное множество сайтов подействовал мутационный процесс. При этом мутации, внесенные в ансамбль, будут так распределены между различными позициями различных вариаций сайта, чтобы частоты тринуклеотидов были такими, как задаёт их мутационная подпись.

Дополнительное условие, которое мы считаем выполняющимся: в одном и том же сайте не могло случиться две замены одновременно.

Изучим вероятность нуклеотида на фиксированной позиции после действия мутационного процесса быть δ. Благодаря тому, что частоты нуклеотидов на различных позициях, независимы, нам достаточно рассматривать действие мутационного процесса локально - на определенной позиции с её тринуклеотидным контекстом. Необходимости перебирать все сайты для этого нет.

Рассмотрим следующее дерево исходов с условными вероятностями (при условии, что родительское условие выполнено):

- с вероятностью ρ сайт мутировал
  - с вероятностью $f^{\delta}_{\alpha\beta\gamma}$ произошла мутация типа $\alpha\beta\gamma \to \delta$
    - с вероятностью $\frac{M^i_{\alpha\beta\gamma}}{M_{\alpha\beta\gamma}} \equiv \frac{M^i_{\alpha\beta\gamma}}{\sum_i M^i_{\alpha\beta\gamma}}$ мутация в тринуклеотиде $\alpha\beta\gamma$ произошла именно на позиции i
- с вероятностью $(1-\rho)$ сайт вообще не мутировал

Для удобства мы обозначили сумму вероятностей встретить тринуклеотид $\alpha\beta\gamma$ на разных позициях сайта:

$$M_{\alpha\beta\gamma} = \sum_i M^i_{\alpha\beta\gamma}$$

Стоит отметить, что эта величина не является суммарной вероятностью встретить этот тринуклеотид на любой из позиций (более того, это вообще не вероятность; например, она может принимать значение больше единицы). Эта величина играет роль нормировочного множителя в ситуации, когда мутация "выбирает", какую позицию сайта она заденет.

Вычислим теперь вероятность получить нуклеотид δ в модели нового множества сайтов на i-ой позиции. Во-первых, выразим её как сумму вероятностей получить определенный контекст:

$$\widetilde{M}^i_{\delta} = \sum_{\alpha,\gamma} \widetilde{M}^i_{\alpha\delta\gamma}$$

Далее, вероятность получить нуклеотид δ складывается из вероятности того, что нуклеотид исходно был δ и не мутировал, а также вероятности того, что нуклеотид был неким β, но стал δ. При этом мутация меняет позицию, если она попала в сайт (с вероятностью ρ), притом контекст мутации совпал с контекстом нуклеотида в сайте и эта мутация из всех

мест с таким контекстом попала именно в нужную позицию. Не меняет позицию, соответственно, во всех других случаях.

Изменения фланкирующих нуклеотидов не играют роли при вычислении частоты встречаемости центрального нуклеотида, поскольку это лишь перераспределит вероятности получить различные контексты нуклеотида δ. Сумма частот контекстов с фиксированным центром не изменится.

$$\widetilde{M}^i_{\alpha\delta\gamma} = \rho \sum_\beta M^i_{\alpha\beta\gamma} f^\delta_{\alpha\beta\gamma} \frac{M^i_{\alpha\beta\gamma}}{M_{\alpha\beta\gamma}} + M^i_{\alpha\delta\gamma} \left(1 - \rho \sum_\xi f^\xi_{\alpha\delta\gamma} \frac{M^i_{\alpha\delta\gamma}}{M_{\alpha\delta\gamma}}\right)$$

$$\Delta M^i_{\alpha\delta\gamma} = \widetilde{M}^i_{\alpha\delta\gamma} - M^i_{\alpha\delta\gamma}$$

$$\Delta M^i_{\alpha\delta\gamma}/\rho = \sum_\beta M^i_{\alpha\beta\gamma} f^\delta_{\alpha\beta\gamma} \frac{M^i_{\alpha\beta\gamma}}{M_{\alpha\beta\gamma}} - M^i_{\alpha\delta\gamma} \sum_\xi f^\xi_{\alpha\delta\gamma} \frac{M^i_{\alpha\delta\gamma}}{M_{\alpha\delta\gamma}}$$

$$\Delta M^i_\delta = \widetilde{M}^i_\delta - M^i_\delta = \sum_{\alpha,\gamma} \Delta M^i_{\alpha\delta\gamma}$$

$$\Delta M^i_\delta/\rho = \sum_{\alpha,\gamma} \left(\sum_\beta M^i_{\alpha\beta\gamma} f^\delta_{\alpha\beta\gamma} \frac{M^i_{\alpha\beta\gamma}}{M_{\alpha\beta\gamma}} - M^i_{\alpha\delta\gamma} \sum_\xi f^\xi_{\alpha\delta\gamma} \frac{M^i_{\alpha\delta\gamma}}{M_{\alpha\delta\gamma}} \right)$$

$$\Delta M^i_\delta/\rho = \sum_{\alpha,\gamma} \left(\sum_\beta \frac{\left(M^i_{\alpha\beta\gamma}\right)^2}{M_{\alpha\beta\gamma}} f^\delta_{\alpha\beta\gamma} - \frac{\left(M^i_{\alpha\delta\gamma}\right)^2}{M_{\alpha\delta\gamma}} \sum_\xi f^\xi_{\alpha\delta\gamma} \right)$$

Мы вычислили, в какую сторону и насколько сильно мутационный процесс будет перетягивать ансамбль сайтов. Если мы зафиксируем интенсивность мутационного процесса, положив $\rho = 1$, мы можем применить метрики схожести мотивов для оценки того, насколько новое множество сайтов отличается от старого. Например, при помощи macro-ape мы можем вычислить Джаккарову похожесть множеств топовых слов. Это будет некая характеристика прочности (хрупкости) мотива против данной мутационной подписи.

4 Коррекция мутационной подписи

В методе явно предполагается, что мутационная подпись сайта нам задана. Но на практике мы знаем скорее полногеномную мутационную подпись. Нередко на ансамбле сайтов (особенно для консервативных мотивов) достигнуть заданной мутационной подписи не представляется возможным: в сайтах просто нет необходимых контекстов. Это отражается в том, что для некоторого набора α, β, γ, стоящая в знаменателе сумма вероятностей встретить соответствующий контекст $M_{\alpha\beta\gamma} = 0$, тогда как частота мутаций $f^\delta_{\alpha\beta\gamma} \neq 0$.

В качестве простейшего решения можно добавить фланки, содержащие равнораспределенные нуклеотиды - это позволит мутациям с нетипичными для сайта контекстами попадать во фланки. Остаётся однако не вполне ясным, как корректно сгенерировать фланки так, чтобы они моделировали полногеномное распределение тринуклеотидов.

Кроме того, стоит понять, не размывает ли добавление к мотиву фланок понятие интенсивности мутационного процесса ρ. Кажется, что нет: $\rho = 1$ отвечает фактически ситуации, когда на каждый из 4^n сайтов приходится в среднем по одной мутации (но размазывается это общее количество по всему ансамблю, так что могут быть сайты, не имеющие мутации).

В качестве альтернативного варианта, можно конвертировать полногеномное распределение мутаций ${}^{\text{wg}}f^{\delta}_{\alpha\beta\gamma}$ в сайт-специфическое распределение мутаций ${}^{\text{ss}}f^{\delta}_{\alpha\beta\gamma}$. Обозначим число тринуклеотидов в сайтах как ${}^{\text{ss}}N$, в полном геноме как ${}^{\text{wg}}N$. Для частот тринуклеотидов воспользуемся ${}^{\text{ss}}\widetilde{M}_{\alpha\beta\gamma}$ и ${}^{\text{wg}}\widetilde{M}_{\alpha\beta\gamma}$.

$$\widetilde{M}_{\alpha\beta\gamma} = \frac{M_{\alpha\beta\gamma}}{\sum_{\alpha,\beta,\gamma} M_{\alpha\beta\gamma}}$$

Мы предполагаем, что вероятность позиции мутировать зависит только от контекста, т.е. все позиции с одинаковым контекстом мутируют равновероятно. В геноме, однако, существуют позиции с различной доступностью для мутационного процесса. В данной модели можно учесть это лишь грубо: считая частоты (и количества тринуклеотидов) не по полному геному, а по той его части, что доступна для мутаций, например, для открытого хроматина.

Тогда можно выразить частоту попаданий мутации определенного типа в сайт следующим образом:

$$P_0 \cdot {}^{\text{ss}}f^{\delta}_{\alpha\beta\gamma} = {}^{\text{wg}}f^{\delta}_{\alpha\beta\gamma} \cdot \frac{{}^{\text{ss}}N \cdot {}^{\text{ss}}\widetilde{M}_{\alpha\beta\gamma}}{{}^{\text{wg}}N \cdot {}^{\text{wg}}\widetilde{M}_{\alpha\beta\gamma}} = J \cdot {}^{\text{wg}}f^{\delta}_{\alpha\beta\gamma} \cdot \frac{{}^{\text{ss}}\widetilde{M}_{\alpha\beta\gamma}}{{}^{\text{wg}}\widetilde{M}_{\alpha\beta\gamma}},$$

$${}^{\text{ss}}f^{\delta}_{\alpha\beta\gamma} = \frac{J}{P_0} \cdot {}^{\text{wg}}f^{\delta}_{\alpha\beta\gamma} \cdot \frac{{}^{\text{ss}}\widetilde{M}_{\alpha\beta\gamma}}{{}^{\text{wg}}\widetilde{M}_{\alpha\beta\gamma}},$$

где P_0 – вероятность мутации попасть в сайт (при условии, что мутация случилась), а $J = {}^{ss}N/{}^{wg}N$ – доля генома, которую занимают сайты.

Из условия нормировки

$$\sum_{\alpha,\beta,\gamma,\delta} {}^{\text{ss}}f^{\delta}_{\alpha\beta\gamma} = 1$$

можно вычислить значение нормировочного множителя

$$\varkappa = \left(\frac{J}{P_0}\right)^{-1} = \sum_{\alpha,\beta,\gamma,\delta} {}^{\text{wg}}f^{\delta}_{\alpha\beta\gamma} \frac{{}^{\text{ss}}\widetilde{M}_{\alpha\beta\gamma}}{{}^{\text{wg}}\widetilde{M}_{\alpha\beta\gamma}} = \sum_{\alpha,\beta,\gamma} \left(\frac{{}^{\text{ss}}\widetilde{M}_{\alpha\beta\gamma}}{{}^{\text{wg}}\widetilde{M}_{\alpha\beta\gamma}} \sum_{\delta} {}^{\text{wg}}f^{\delta}_{\alpha\beta\gamma} \right)$$

Отсюда, зная долю генома, покрытую сайтами, J можно вычислить также вероятность мутации попасть в сайт P_0. Этой оценкой вероятности можно пользоваться при оценке правдоподобия доли попавших в сайт мутаций (которое моделируется биномиальным распределением).

Можно трактовать J как вероятность мутации попасть в сайт, если бы мутации не имели бы контекстной специфичности – тогда вероятность зависела бы только от суммарной доли генома, покрытой сайтами. $\varkappa$ же представляет из себя, во сколько раз чаще или реже мутация должна попадать в сайт, чем при таком равновероятном процессе. Фактически, это “подверженность” сайта мутационному процесса (но попавшие в сайт мутации вовсе не обязаны сайт разрушать).

Чтобы оценить хрупкость конкретного сайта можно смоделировать его как индикаторную частотную матрицу с единичными вероятностями для нуклеотидов в этом сайте.

Также, возможно, имеет смысл оценивать перекошенность мутационного процесса, измеряя подверженность $\varkappa$ мутационному процессу "мотива", представляющего собой равновероятное распределение нуклеотидов. Эксперимент показывает, что реальные мутационные процессы имеют $\varkappa > 1$, т.е. такие нейтральные области более часто поражаются мутациями. Как я понимаю, это можно объяснить тем, что часть неоднородных мотивов избегается мутациями.

Ruby-скрипт, применяющий мутационный процесс к частотной матрице вместе с тестовыми примерами можно найти в репозитории https://github.com/VorontsovIE/motif_fragility