Motif attractiveness and stability with respect to mutational process

Ilya E. Vorontsov

March 2, 2019

Модель мутационного процесса

На геноме действует мутационный процесс. Мы предполагаем, что он случайным образом равновероятно выбирает одну из позиций генома, а затем вносит или не вносит мутацию с некоторой вероятностью, которая зависит лишь от геномного контекста позиции и направления мутации.

Направлением мутации мы называем тип замены, которая происходит. Следуя за Александровым и Стрэттоном мы рассматриваем лишь тринуклеотидные контексты и однонуклеотидные замены в них. Это объясняется в большей степени практическими соображениями, чем теоретическими сложностями. В частности, работать с контекстами большей длины можно совершенно аналогично, если статистика позволяет надежно оценить частоты подобных мутаций. В дальнейшем ненаправленными контекстами или просто контекстами мы называем окно из трех нуклеотидов с центром на заданной позиции. Направленный контекст – это пара из контекста и нуклеотида, который заменяет центральную позицию. Мы отождествляем замены на разных цепях ДНК и обозначаем контекст всегда так, чтобы центральный нуклеотид был пиримидином (т.е. C или T). В таком случае мы можем рассматривать $4 \times 2 \times 4 = 32$ конекста. Типов мутаций для заданного контекста – три (по числу возможных замен нуклеотида) – итого 96 направленных контекстов.

Чтобы вести дальнейшее повествование максимально аккуратно договоримся об обозначениях.

Обозначения

Две величины, которые играют особую роль в вычислениях: распределение числа мутаций по типам и встречаемость различных контекстов в геноме. Эти величины считаются известными нам заранее.

Обозначение $m^{wg}(ctx \to ctx')$ показывает число мутаций в полном геноме, произошедших в контексте ctx и заменивших его на контекст ctx'.

Обозначение $N^{wg}(ctx)$ характеризует число вхождений последовательности ctx в полный геном.

Верхний индекс $\cdot^{wg}$ следует понимать как “whole genome”. Индекс говорит о том, что величина посчитана для полного генома.

Отметим, что ключевое для нашей модели условие равновероятности случайного выбора мутагеном позиций генома едва ли реализуется. Однако

мы можем рассматривать сужение генома, на котором мы предполагаем это условие приближенно выполненным. В качестве такого сужения, можно взять, например, области открытого хроматина. Полным геномом $\cdot^{wg}$ мы будем в действительности называть именно такое сужение.

Изучать же мы будем действие мутационного процесса на подмножествах этого сужения полного генома. Для величин относящихся к этому подмножеству мы будем использовать индекс $\cdot^{rs}$, обозначающий "region specific". Например, $N^{rs}(ctx)$ – это число вхождений контекста ctx в области rs. Область rs может быть как непрерывной (например, один конкретный промотер), так и разрывной (например, все сайты связывания некоторого транскрипционного фактора).

Для тех формул, которые верны и для полного генома, и для его подмножества, мы будем писать вместо индекса плейсхолдер $\bullet$.

Число мутаций в области может быть и посчитана непосредственно, и оценено из других соображений. Мы будем использовать крышечку $\widehat{\dots}$, чтобы различать эти две величины. Так $\widehat{m}^{rs}(ctx \to ctx')$ – это реальное число мутаций типа $ctx \to ctx'$ в области rs, а $m^{rs}(ctx \to ctx')$ – это предсказанное число мутаций данного типа в области.
Величины $m^{wg}(ctx \to ctx')$ и $\widehat{m}^{wg}(ctx \to ctx')$ мы будем отождествлять, хотя из байесовских оценок можно было бы получить более точное значение m^{wg} на основе экспериментально измеренного $\widehat{m}^{wg}$. Но при высоком числе мутаций эти оценки не должны значительно отличаться.

По ходу статьи у нас будут постоянно возникать частично агрегированные распределения: например, число мутаций в данном мутационном контексте, но без различения направления мутации. Или вовсе общее число мутаций в области. Для подобных частично агрегированных величин мы будем придерживаться следующей схемы обозначений:

1. $m^{\bullet}(ctx \to ctx')$ – число мутаций вида $ctx \to ctx'$

2. $m^{\bullet}(ctx)$ – число мутаций, произошедших в контексте ctx с любым направлением мутации.
   Для числа мутаций верно следующее:
   $$m^{\bullet}(ctx) = \sum_{ctx'} m^{\bullet}(ctx \to ctx')$$

3. $m^{\bullet}$ – полное число произошедших мутаций любых типов.
   $$m^{\bullet} = \sum_{ctx} m^{\bullet}(ctx) = \sum_{ctx \to ctx'} m^{\bullet}(ctx \to ctx')$$
   Аналогично, $N^{\bullet}$ – полное число позиций в рассматриваемой области, независимо от контекста.

4. При изучении мутаций в сайтах связывания нам также встретится обозначение:
   $m_i^{\bullet}(ctx \to ctx')$ – число мутаций, случившихся в i-ой позиции изучаемого мотива. Ограничение на позицию мутации может комбинироваться с любыми описанными агрегациями по типам мутации.

Мутационный процесс в заданной области

Первой величиной, которая нас интересует является ожидаемое число мутаций, попадающих в изучаемую область генома. Если бы вероятность внести мутацию никак не зависела от контекста, то ожидаемое число мутаций было бы произведением общего числа мутаций в полном геноме и отношения длины области к длине полного генома. Однако в нашем случае можно ожидать, что некоторые области будут мутировать чаще или реже, чем эта величина – в зависимости от того, насколько состав контекстов в области похож на контексты, в которые мутационный процесс успешно вносит замены.

В случае, когда область задана явно мы можем просто посчитать количество мутаций, которые реально попали в область. Это позволяет нам оценить, насколько количество мутаций, внесенных в реальности деле, отличается от спрогнозированного. Отклонение реального распределения мутаций от ожидаемого может быть истрактовано как действие эффекта отбора в исследуемой области.

Это же дает нам способ валидировать модель: если модель верна, отклонение реального и спрогнозированного числа мутаций должно быть невелико для случайной (не обладающей выраженной функциональностью) области. Если же отклонение велико, это укажет на неучтенные моделью особенности мутационных процессов.

Количество мутаций, попавших в область, подчинено биномиальному распределению: $m^{rs} \sim Binom(m^{wg}, p^{rs})$, в котором число испытаний – это общее число мутаций в геноме m^{wg}. Если мы умеем вычислять ожидаемое число мутаций m^{rs}, то вероятность успеха единичого испытания p^{rs} (успехом мы называем попадание в область) мы можем оценить, исходя из оценки ожидаемого числа успехов биномиального распределения:

$$m^{rs} = m^{wg} \cdot p^{rs} \tag{1}$$

Знание параметров биномиального распределения даёт нам возможность оценить отклонение значение от ожидаемого – и измерить силу эффекта и значимость отличия при условии, что верна нулевая гипотеза о равновероятности внесения мутации в любую позицию с заданным контекстом.

Итак, нам заданы полногеномные количества мутаций в определенном (направленном) контексте: $m^{wg}(ctx \to ctx')$. Также нам известно количество позиций в геноме с определенным контекстом: $N^{wg}(ctx)$ и количество позиций различного контекста в изучаемой области $N^{rs}(ctx)$.

Важно помнить, что количество мутаций в полном геноме не является истинными частотами мутационного процесса, оно так же как и число мутаций в некоторой подобласти зависит от распределения контекстов в геноме. Пусть мутационный процесс, действует с определенной интенсивностью m_0 (число позиций, с которыми мутаген успел столкнуться), вероятность внести замену при единичном взаимодействии – $g(ctx \to ctx')$. Тогда ожидаемое число мутаций будет следующим:

$$m^{\bullet}(ctx \to ctx') = m_0 \cdot \frac{N^{\bullet}(ctx)}{N^{wg}} \cdot g(ctx \to ctx')$$

На практике мы не знаем ни m_0, ни g, но можем получить соотношение между числом мутаций в области и числом мутаций в полном геноме. Введём

обозначение $\mu(ctx \to ctx')$ - контексто-зависимую плотность мутаций:

$$\mu(ctx \to ctx') = \frac{m^{\bullet}(ctx \to ctx')}{N^{\bullet}(ctx)} \tag{2}$$

Нам будет удобно пользоваться разложить эту плотность на две компоненты:

$$\mu(ctx \to ctx') = \mu(ctx) \cdot \chi(ctx \to ctx') \tag{3}$$

Здесь $\mu(ctx)$ - плотность, которая не учитывает направление замены.

$$\mu(ctx) = \frac{m^{\bullet}(ctx)}{N^{\bullet}(ctx)} = \sum_{ctx'} \mu(ctx \to ctx')$$

Величина $\chi(ctx \to ctx')$ характеризует долю замен в контексте ctx, которые преобразуются в ctx'. Как положено распределению вероятностей, она суммируется к единице:

$$\sum_{ctx'} \chi(ctx \to ctx') = 1$$

Величина $\mu(ctx \to ctx')$ в нашей модели не зависит от региона, поэтому её возможно оценить на полном геноме, и использовать в вычислениях на области rs. Таким образом мы можем записать:

$$\begin{aligned} m^{rs}(ctx \to ctx') &= m^{wg}(ctx \to ctx') \cdot \frac{N^{rs}(ctx)}{N^{wg}(ctx)} \\ &= \mu(ctx \to ctx') \cdot N^{rs}(ctx) \\ &= \mu(ctx) \cdot \chi(ctx \to ctx') \cdot N^{rs}(ctx) \end{aligned} \tag{4}$$

Полное число мутаций в области вычислим как сумму чисел мутаций различных типов:

$$m^{rs} = \sum_{ctx} \mu(ctx) \cdot N^{rs}(ctx) \tag{5}$$

В случае, когда число мутаций мало, чтобы надежно оценить частоты мутаций в различных контекстах, нам следовало бы воспользоваться формулой Байеса и в дальнейших вычислениях работать не с константным значением $\mu(ctx)$, а вероятностным распределением этой величины, обусловленным наблюдаемыми данными.

Стоит отметить, что зависимость числа мутаций в регионе линейна по числу мутаций в геноме; этот довольно тривиальный факт дает нам возможность работать со смесями мутационных процессов, не проводя предварительную декомпозицию на мутационные подписи независимых процессов.

Вероятность единичного попадания мутации в область rs теперь можно оценить как

$$p^{rs} = \frac{m^{rs}}{m^{wg}} = \sum_{ctx} \frac{m^{wg}(ctx)}{m^{wg}} \cdot \frac{N^{rs}(ctx)}{N^{wg}(ctx)} \tag{6}$$

Также мы можем ввести характеристику притягательность Q (attractiveness) региона для мутационного процесса как отношение вероятности мутации попасть в регион к доле генома, которую составляет регион:

$$Q = \frac{p^{rs}}{N^{rs}/N^{wg}} = \sum_{ctx} \frac{m^{wg}(ctx)}{m^{wg}} \cdot \frac{N^{wg}}{N^{wg}(ctx)} \cdot \frac{N^{rs}(ctx)}{N^{rs}} \tag{7}$$

Фактически, она оценивает, во сколько раз больше мутаций попадает в этот регион, благодаря мутационной специфичности, в сравнении со случаем, когда вероятность внесения мутации от контекста бы не зависела (в этом случае $m^\bullet(ctx)/m^\bullet = N^\bullet(ctx)/N^\bullet$ и $Q = 1$).

Мутационный процесс на ансамбле сайтов

Притягательность

Теперь мы можем рассмотреть частный случай области – ансамбль сайтов связывания некоторого транскрипционного фактора. Мы не хотим искать в геноме все сайты связывания, поскольку сайты связывания тканеспецифичны, а набор сайтов сильно зависит от выбранного порога. Чтобы избавиться от этих деталей, вспомним, что ансамбль сайтов связывания можно описать вероятностной моделью, которая позволяет оценить доли контекстов в этом ансамбле.

Пусть ансамбль сайтов описывается вероятностной моделью S. Вероятность встретить контекст ctx на позиции i мы обозначим как $S(ctx)$.

Для ансамбля из K сайтов длины l мы можем посчитать распределение контекстов[1]:

$$N^{rs}(ctx) = K \cdot \sum_{i=1}^{l} S_i(ctx)$$

$$N^{rs} = K \cdot l$$

Даже не зная числа сайтов мы можем таким образом вычислить притягательность ансамбля сайтов:

$$Q = \sum_{ctx} \frac{m^{wg}(ctx)}{m^{wg}} \cdot \frac{N^{wg}}{N^{wg}(ctx)} \cdot \frac{\sum_{i=1}^{l} S_i(ctx)}{l} \tag{8}$$

Позиционные предпочтения

Для изменения аффинности сайта связывания имеет существенное значение позиция, в которую попадает мутация, и направление мутации. Оценим наиболее подверженные мутированию позиции и изменение аффинности сайтов под воздействием мутационного процесса. Здесь и далее мы считаем, что вероятностью того, что один сайт будет мутирован дважды, можно пренебречь.

Пусть ${m_i}^{rs}(ctx \to ctx')$ – число мутаций заданного типа в i-ой позиции сайта. Оно вычисляется при помощи подстановки в формулу (4) области, состоящей из i-ых позиций во всех сайтах ансамбля:

$$\begin{aligned} {m_i}^{rs}(ctx \to ctx') &= K \cdot S_i(ctx) \cdot \frac{m^{wg}(ctx \to ctx')}{N^{wg}(ctx)} \\ &= K \cdot S_i(ctx) \cdot \mu(ctx) \cdot \chi(ctx \to ctx') \end{aligned} \tag{9}$$

Позиционным профилем ${\pi_i}^{rs}$ назовем долю мутаций, попадающих в i-ую позицию мотива, среди всех мутаций, попавших в сайты мотива. Соответственно

[1] Фланки в длину сайта не входят.

$\pi_i(ctx \to ctx')$ – доля мутаций определенного типа в i-ой позиции мотива среди всех мутаций, попавших в сайты. Можно рассматривать π_i, как условную вероятность попасть в определенную позицию сайта при условии, что мутация уже попала в сайт.

$$\begin{aligned}{\pi_i}^{rs}(ctx \to ctx') &= \frac{{m_i}^{rs}(ctx \to ctx')}{m^{rs}} \\ &= \frac{K \cdot S_i(ctx) \cdot \mu(ctx)}{\sum_i \sum_{ctx \to ctx'} K \cdot S_i(ctx) \cdot \mu(ctx) \cdot \chi(ctx \to ctx')} \cdot \chi(ctx \to ctx')\end{aligned}$$

Итого:

$${\pi_i}^{rs}(ctx \to ctx') = \frac{S_i(ctx) \cdot \mu(ctx)}{\sum_i \sum_{ctx} S_i(ctx) \cdot \mu(ctx)} \cdot \chi(ctx \to ctx') \tag{10}$$

Для компактности введём нормировочный множитель

$$\varkappa = \frac{1}{\sum_i \sum_{ctx} S_i(ctx) \cdot \mu(ctx)} \tag{11}$$

и перепишем выражение для позиционного профиля:

$${\pi_i}^{rs}(ctx \to ctx') = \varkappa \cdot S_i(ctx) \cdot \mu(ctx) \cdot \chi(ctx \to ctx') \tag{12}$$

Тогда позиционный профиль без учёта контекста может быть вычислен как:

$${\pi_i}^{rs} = \varkappa \cdot \sum_{ctx} S_i(ctx) \cdot \mu(ctx) \tag{13}$$

Изменение аффинности

Далее мы хотели бы узнать, как меняется аффинность связывания в результате мутирования сайтов ансамбля. Для оценки аффинности мы будем использовать весовую матрицу W (она должна быть согласована по позициям с частотной матрицей S, хотя и не обязана быть получена непосредственно из неё). Нас интересует, как W меняется под действием мутационного процесса. Оценим среднее изменение веса $\mathbb{E}\Delta W$ и средний квадрат изменения веса $\mathbb{E}(\Delta W)^2$.

Мы будем рассматривать только мутации, попавшие в один из сайтов ансамбля. Мутации, в сайт не попавшие, афинность изменить не могут (т.е. для них $\Delta W = 0$), поэтому связь между матожиданием на множестве всех мутаций $\mathbb{E}^{wg}$ и матожиданием на множестве мутаций в сайтах $\mathbb{E}^{rs}$ выражается простым умножением на вероятность $p^{rs} = Q \cdot \frac{N^{rs}}{N^{wg}}$ мутации попасть в сайт.

Множитель $\frac{N^{rs}}{N^{wg}}$ не зависит от нуклеотидного состава сайтов, а лишь указывает, насколько сайтов в геноме много, тогда как притягательность Q определяет, насколько много мутаций сайт набирает в сравнении со случайным участком генома, и таким образом является важной характеристикой региона.

$$\mathbb{E}^{wg}\Delta W = \frac{N^{rs}}{N^{wg}} \cdot Q \cdot \mathbb{E}^{rs}\Delta W$$

$$\mathbb{E}^{wg}(\Delta W)^2 = \frac{N^{rs}}{N^{wg}} \cdot Q \cdot \mathbb{E}^{rs}(\Delta W)^2$$

Зная условную вероятность ${\pi_i}^{rs}(ctx \to ctx')$ мутации $ctx \to ctx'$ попасть в позицию i (при условии, что мутация уже попала в сайт), мы можем вычислять моменты распределения изменений аффинности. Обозначим изменение веса ΔW при замене нуклеотида $ctx \to ctx'$ в позиции i как $\Delta W_i(ctx \to ctx')$, оно вычисляется непосредственно из весовой матрицы. Тогда

$$\mathbb{E}^{rs}\Delta W = \sum_i \sum_{ctx \to ctx'} {\pi_i}^{rs}(ctx \to ctx') \cdot \Delta W_i(ctx \to ctx') \tag{14}$$

$$\mathbb{E}^{rs}(\Delta W)^2 = \sum_i \sum_{ctx \to ctx'} {\pi_i}^{rs}(ctx \to ctx') \cdot (\Delta W_i(ctx \to ctx'))^2 \tag{15}$$

Мы можем считать, что у мотива есть две важные характеристики: притягательность Q, описывающая как часто он подвергается мутации. И распределение ΔW, которое говорит, как сильно изменяется аффинность сайта, если мутация в него попала. Проблема в том, что ΔW не сравнима между мотивами. Чтобы преодолеть это ограничение, мы можем смоделировать частоты нуклеотидов в сайтах, подвергшихся действию мутационного процесса, и затем сравнить частотные матрицы одним из множества методов. Например, мы можем сравнить пересечение множества слов, преодолевающих некоторый порог для исходного и итогового ансамблей сайтов [MacroAPE].

Смещение ансамбля сайтов под действием мутационного процесса

Рассмотрим как меняется ансамбль сайтов под действием мутационного процесса. Пусть мутационный процесс вносит в среднем ρ мутаций на сайт. Вычислим вероятность встретить определенный контекст в ансамбле сайтов после мутирования, т.е. опишем вероятностную модель $\widetilde{S}_i(ctx)$ ансамбля сайтов после мутирования.

Число позиций с определенным контекстом в ансамбле сайтов размера K выразится как:

$$K \cdot \widetilde{S}_i(ctx) = K \cdot S_i(ctx) - \sum_{ctx'} K \cdot \rho \cdot {\pi_i}^{rs}(ctx \to ctx') + \sum_{ctx'} K \cdot \rho \cdot {\pi_i}^{rs}(ctx' \to ctx) \tag{16}$$

Таким образом для изменения частот

$$\Delta S_i(ctx) := \widetilde{S}_i(ctx) - S_i(ctx)$$

можно записать:

$$\Delta S_i(ctx)/\rho = \sum_{ctx'} {\pi_i}^{rs}(ctx' \to ctx) - \sum_{ctx'} {\pi_i}^{rs}(ctx \to ctx') \tag{17}$$

Давайте теперь рассмотрим в качестве частного случая вероятностной модели позиционную матрицу частот[2]. $S_i(\alpha)$ обозначает вероятность встретить

[2]Мы дополняем частотную матрицу S фланкирующими колонками, которые подчинены равномерному (или в случае смещенности геномного GC-состава от 0.5 – среднему по геному) распределению нуклеотидов. Это позволяет учитывать контексты, чья центральная буква лежит на краю мотива. В то же время операция не требует дополнительных допущений: если бы во фланках частоты нуклеотидов отличались бы от среднегеномного, то они вошли бы дополнительной позицией в мотив, как добавляющие ненулевую информацию о последовательности сайта связывания.

нуклеотид α на позиции i. Нас будут интересовать тринуклеотидные контексты. Вероятности различных позиций в нашей модели независимы друг от друга, поэтому мы можем вычислить вероятность $S_i(\alpha\beta\gamma)$ встретить контекст $\alpha\beta\gamma$, центрированный на позиции i:

$$S_i(\alpha\beta\gamma) = S_{i-1}(\alpha)S_i(\beta)S_{i+1}(\gamma)$$

Вычислим изменение частотной матрицы.

$$\Delta S_i(\alpha\beta\gamma)/\rho = \sum_{\delta} {\pi_i}^{rs}(\alpha\delta\gamma \to \alpha\beta\gamma) - \sum_{\delta} {\pi_i}^{rs}(\alpha\beta\gamma \to \alpha\delta\gamma) \tag{18}$$

Снова используя предположение о независимости позиций, мы вычислим изменение распределения вероятностей центрального нуклеотида:

$$\Delta S_i(\beta) = \sum_{\alpha,\gamma} \Delta S_i(\alpha\beta\gamma) \tag{19}$$

$$\Delta S_i(\beta)/\rho = \sum_{\alpha,\delta,\gamma} ({\pi_i}^{rs}(\alpha\delta\gamma \to \alpha\beta\gamma) - {\pi_i}^{rs}(\alpha\beta\gamma \to \alpha\delta\gamma)) \tag{20}$$

$$\boxed{\Delta S_i(\beta) = \rho \cdot \varkappa \cdot \sum_{\alpha,\gamma} \left(\sum_{\delta} S_i(\alpha\delta\gamma) \cdot \mu(\alpha\delta\gamma \to \alpha\beta\gamma) - S_i(\alpha\beta\gamma) \cdot \sum_{\delta} \mu(\alpha\beta\gamma \to \alpha\delta\gamma) \right)} \tag{21}$$