Хрупкость мотива по отношению к мутационному процессу

Воронцов И.Е.

August 19, 2018

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

Также мы рассмотрим уязвимость произвольного участка генома к мутационному процессу.

1 Обозначения

В наших вычислениях множество раз необходимо будет отнормировать некоторый набор величин по всем вариантам значений набора переменных var. Чтобы не расписывать всякий раз нормировочный множитель, мы введем специальное обозначение:

$$\widehat{\eta}_{var}\left[f(var)\right] := \frac{f(var)}{\sum_{var} f(var)} \tag{1}$$

2 Мотив

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

$$P(w|w \text{ is a site}) = \prod_{i=1}^{n} M_i(w_i) \tag{2}$$

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

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

$$M_i(\alpha\beta\gamma) = M_{i-1}(\alpha)M_i(\beta)M_{i+1}(\gamma) \tag{3}$$

Мы будем работать с ансамблем слов одинаковой длины, представляющих сайты связывания. В ансамбле будут встречаться все возможные слова заданной длины: мы не задаем порог на минимальный вес распознаваемого слова. Но слова отвечающие сильным сайтам будут встречаться чаще, чем слова, которые не распознаются как сайт - в полном соответствии с формулой (2).

3 Мутационный процесс

В нашей модели мутационный процесс задаётся мутационной подписью, т.е. долей замен последовательности ctx на ctx' среди всех произошедших замен. Эти частоты обозначим $f(ctx \to ctx')$.

Для случая однонуклеотидных замен в тринуклеотидном контексте: $\alpha\beta\gamma \to \alpha\delta\gamma$ (или кратко $\alpha\beta\gamma \to \delta$) мы введём специальное обозначение $f^{\delta}_{\alpha\beta\gamma}$.

Мутационный процесс задаёт частоты, следовательно суммируется к единице:

$$\sum_{ctx,ctx'} f(ctx \to ctx') = 1 \tag{4}$$

Так как он описывает лишь произошедшие мутации (незатронутые мутационным процессом позиции мы рассматриваем отдельно), поэтому положим

$$f_{ctx \to ctx} = 0 \tag{5}$$

Также мы полагаем, что мутации не имеют предпочтений по нити ДНК. Мы симметризуем частоты мутаций так, чтобы выполнялось условие:

$$f(ctx \to ctx') = f(revcomp(ctx) \to revcomp(ctx')) \tag{6}$$

4 Действие мутационного процесса на ансамбль сайтов, характеризуемый мотивом

Рассмотрим ансамбль из N слов одинаковой длины, который подвергается действию мутационного процесса. Мы будем рассматривать внесение мутации как два последовательных случайных события:

- выбор типа мутации $ctx \to ctx'$ с частотой $f(ctx \to ctx')$
- равновероятный выбор одной позиции, в которую вносится мутация, среди всех, имеющих контекст ctx, позиций всех сайтов ансамбля

Мутационный процесс интенсивностью ρ вносит $N \cdot \rho$ мутаций. Мы полагаем, что ρ достаточно мало, чтобы можно было пренебречь вероятностью того, что одна и та же позиция ансамбля быть мутирована дважды. В то же время $N \cdot \rho$ достаточно велико, чтобы частоты мутаций стабилизировались.

В этой главе будем считать, что $f(ctx \to ctx')$ - это частоты мутаций среди мутаций, попавших в сайты ансамбля. Эта частота отличается от частоты мутаций среди всех мутаций генома. Соответствующие поправки мы рассмотрим в следующей главе.

Мы хотим узнать, как поменяется ансамбль под воздействием заданного мутационного процесса. Мы будем предполагать независимость нуклеотидов на различных позициях сайтов и описывать ансамбль частотной матрицей M. Мутировавший ансамбль также опишем частотной матрицей $\widetilde{M}$.

$M_i(ctx)$ - это вероятность того, что в случайном сайте из ансамбля на i-ой позиции находится контекст ctx.

Вычислим, как изменяется число сайтов с контекстом ctx' на позиции i под действием мутационного процесса. Исходная число таких сайтов в ансамбле равно $N \cdot M_i(ctx')$.

Далее случается $N \cdot \rho \cdot f(ctx' \to ctx)$ мутаций меняющих ctx' на ctx.

Вероятность этой мутации попасть именно в i-ю позицию учитывает равновероятность всех позиций ансамбля. Введём вероятность выбора одной из позиций, при условии, что контекст известен:

$$\mu_i(ctx') = \widehat{\eta}_i\left[M_i(ctx')\right] \tag{7}$$

Обозначим вероятность того что мутация внесёт определенную замену на конкретной позиции как $p_i(ctx' \to ctx)$:

$$p_i(ctx' \to ctx) = f(ctx' \to ctx) \cdot \mu_i(ctx') \tag{8}$$

Теперь мы готовы написать, как изменится количество позиций ансамбля, имеющих контекст ctx' и пришедших с i-й позиции сайта.

$$N \cdot \widetilde{M}_i(ctx') = N \cdot M_i(ctx') - \sum_{ctx} N \cdot \rho \cdot p_i(ctx' \to ctx) + \sum_{ctx} N \cdot \rho \cdot p_i(ctx \to ctx') \tag{9}$$

$$\Delta M_i(ctx') := \widetilde{M}_i(ctx') - M_i(ctx') \tag{10}$$

$$\boxed{\Delta M_i(ctx')/\rho = \sum_{ctx} p_i(ctx \to ctx') - \sum_{ctx} p_i(ctx' \to ctx)} \tag{11}$$

Рассмотрим теперь частный случай, когда контекст представляет из себя тринуклеотид, а замены меняют центральный нуклеотид. Тогда можно переписать формулу как

$$\Delta M_i(\alpha\delta\gamma)/\rho = \sum_{\beta} p_i(\alpha\beta\gamma \to \alpha\delta\gamma) - \sum_{\beta} p_i(\alpha\delta\gamma \to \alpha\beta\gamma) \tag{12}$$

Воспользуемся теперь независимостью нуклеотидов и выразим изменение частоты центрального нуклеотида:

$$\Delta M_i(\delta) = \sum_{\alpha,\gamma} \Delta M_i(\alpha\delta\gamma) \tag{13}$$

$$\Delta M_i(\delta)/\rho = \sum_{\alpha,\beta,\gamma} \left(p_i(\alpha\beta\gamma \to \alpha\delta\gamma) - p_i(\alpha\delta\gamma \to \alpha\beta\gamma)\right) \tag{14}$$

$$\boxed{\Delta M_i(\delta) = \rho \cdot \sum_{\alpha,\gamma} \left(\sum_{\beta} \mu_i(\alpha\beta\gamma) \cdot f(\alpha\beta\gamma \to \alpha\delta\gamma) - \mu_i(\alpha\delta\gamma) \cdot \sum_{\beta} f(\alpha\delta\gamma \ \to \alpha\beta\gamma) \right)} \tag{15}$$

Мы вычислили, в какую сторону мутационный процесс будет перетягивать ансамбль сайтов. Если мы зафиксируем интенсивность мутационного процесса(например, положив $\rho = 1$), то по этой формуле мы легко можем вычислить частотную матрицу, описывающую мутировавший ансамбль сайтов. Зная исходную и полученную частотные матрицы, 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.

Также мы можем посчитать как изменится средний по ансамблю вес сайта $\mathbb{E}_{m\sim M}W(m)$, используя произвольную весовую матрицу $W_i(\delta)$.

$$\mathbb{E}_{m\sim M}W(m) = \sum_i \sum_\delta W_i(\delta)\cdot M_i(\delta) \tag{16}$$

Логичным выбором будет взять весовую матрицу, построенную по частотной матрице исходного ансамбля. Изменение её среднего веса будет характеризовать, уменьшилась или увеличилась в среднем аффинность ТФ к сайтам связывания под действием мутационного процесса:

$$\Delta\mathbb{E}_{m\sim M}W(m) = \sum_{i,\delta} W_i(\delta)\cdot \Delta M_i(\delta) \tag{17}$$

5 Коррекция мутационной подписи

До сих пор в наших вычислениях мы предполагали, что мутационный процесс действует только на позиции ансамбля сайтов. Однако мутационный процесс работает на всём геноме, и на практике мы знаем только полногеномную мутационную подпись. Нам же необходимо вычислить мутационную подпись на множестве сайтов связывания, а также перенормировать интенсивность мутационного процесса.

На этот раз мутация случайным образом выбирает одну из $K^{wg}(ctx)$ позиций полного генома (whole genome) с соответствующим контекстом ctx. Среди этих позиций есть $K^{ss}(ctx)$ позиций, принадлежащих ансамблю (site specific). Число позиций с некоторым контекстом в полном геноме мы можем посчитать непосредственно.

Для ансамбля сайтов длины L, имеющего K^{ss} позиций, задаваемого частотной матрицей можно написать:

$$K^{ss}(ctx) = K^{ss}\cdot\widehat{\eta}_{ctx}\left[\sum_j M_j(ctx)\right] = \frac{K^{ss}}{L}\cdot\sum_j M_j(ctx), \tag{18}$$

(кстати) Мы предполагаем, что сайты расширены однородными фланками, позволяющими к любой позиции сайта “приложить” контекст.

Пусть мутационный процесс внёс $N^{wg}(ctx)$ мутаций контекста ctx, из них $N^{ss}(ctx)$ мутаций попало в позиции ансамбля. Число мутаций пропорционально числу соответствующих позиций:

$$\frac{N^{ss}(ctx\to ctx')}{K^{ss}(ctx)} = \frac{N^{wg}(ctx\to ctx')}{K^{wg}(ctx)} \tag{19}$$

Выразим теперь частоты полногеномного $f^{wg}(ctx\to ctx')$ и сайт-специфичного $f^{ss}(ctx\to ctx')$ мутационных процессов через число мутаций разных типов, а затем перепишем их через соотношение частот встречаемости различных контекстов.

$$f^{wg}(ctx\to ctx') = N^{wg}(ctx\to ctx')/N^{wg} \tag{20}$$

$$f^{ss}(ctx \to ctx') = N^{ss}(ctx \to ctx')/N^{ss} \tag{21}$$

$$N^{ss} = \sum_{ctx,ctx'} N^{ss}(ctx \to ctx') = \sum_{ctx,ctx'} N^{wg}(ctx \to ctx') \frac{K^{ss}(ctx)}{K^{wg}(ctx)} \tag{22}$$

Осталось написать частоты мутаций:

$$f^{ss}(ctx \to ctx') = \widehat{\eta}_{ctx,ctx'} \left[N^{wg}(ctx \to ctx') \frac{K^{ss}(ctx)}{K^{wg}(ctx)} \right] \tag{23}$$

Вероятность мутации попасть в ансамбль P_{ss}:

$$P_{ss} = \frac{N^{ss}}{N^{wg}} = \frac{\sum_{ctx,ctx'} N^{wg}(ctx \to ctx') \frac{K^{ss}(ctx)}{K^{wg}(ctx)}}{\sum_{ctx,ctx'} N^{wg}(ctx \to ctx')} \tag{24}$$

Мы можем выразить интенсивность мутационного процесса на ансамбле ρ как

$$\rho = \frac{N^{ss}}{K^{ss}/L} \tag{25}$$

Если бы мутации выбирали позицию независимо от контекста, то такая вероятность P_0 попасть в ансамбль зависела бы только от общей доли позиций ансамбля в геноме:

$$P_0 = \frac{K^{ss}}{K^{wg}} \tag{26}$$

Введём характеристику Q "притягательности" мотива для мутационного процесса:

$$Q = \frac{P_{ss}}{P_0} \tag{27}$$

Узнаем необусловленные геномом частоты мутационного процесса $f^0(ctx \to ctx')$.

Рассмотрим процесс, в котором мутации разной контекстной специфичности пытаются внести мутации в случайные позиции число раз, пропорциональное их частотам. Если контекст случайной позиции совпал с контекстом мутации, то она вносится. В таком случае число мутаций внесённых мутационным процессом $N^{wg}(ctx \to ctx')$ будет пропорционально встречаемости контекстов в геноме $K^{wg}(ctx)$:

$$N^{wg}(ctx \to ctx') = const \cdot f^0(ctx \to ctx') \cdot K^{wg}(ctx) \tag{28}$$

$$f^0(ctx \to ctx') = \widehat{\eta}_{ctx,ctx'} \left[\frac{N^{wg}(ctx \to ctx')}{K^{wg}(ctx)} \right] \tag{29}$$

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