Motif attractiveness and stability with respect to mutational process

Ilya E. Vorontsov

January 29, 2019

Statistical analysis of somatic mutations requires a proper background model of mutations distribution. Mutational processes (MPs) acting in different cancers have different sequence specificities. Thus regions with certain local genomic context can accumulate much more mutations than others. Preferences of various MPs to hit distinct regions is important for understanding differences between cancer types.

Transcription factor binding sites has distinct sequence motif which allows us to theoretically study action of MP onto an ensemble of binding sites. Similarity between motif and mutational signature can lead to enhanced accumulation of mutations in motif occurences. Effect of introduced mutation can range from very strong to negligible depending on position of mutated nucleotide and substitution type. Here we introduce notions of motif attractiveness and motif stability (fragility) with respect to a certain mutational process. Motif attractiveness describes is it more or less expected for a motif to be hit by a mutation. Motif fragility tells whether motif occurences tend to be disrupted or strengthen by introduced mutations.

(нам стоит оценить валидность этой модели на случайно сэмплированных участках генома)

(gene ontology анализ надо сделать)

Background

Cancer progression is driven by somatic mutations. We can separate mutations into “driver” and “passenger” ones. The first class of mutations give cells an advantage during clonal expansion and are a subject of positive selection. Though driver mutations is just a small fraction of all somatic mutations and we need powerful statistical methods to indicate which mutations are beneficial for cancer cells. Study of selection pressure requires a proper background model for a mutational process (MP). The majority of somatic mutations in cancer are single-nucleotide variations (SNVs), thus we focused on statistical properties of SNVs.

Here we study a simple model of MP based on the concept of mutational signatures introduced in [Alexandrov]. In our model probability to introduce a substitution into a certain position of genome depends only on genomic context of this position. One can think of this process in the following way: certain mutagen randomly encounter different positions of genome. At each encounter mutagen has some probability to mutate it, this probability depends on sequence context of that position. This process is repeated lots of times and finally mutagen introduces some number of mutations.

Each mutagen has its own mutational preferences and acts through cancer cell evolution with different cumulative intensivity (number of mutations caused by certain mutational process is called as its exposure). Each mutational process can be decomposed into a weighted sum of independent mutational processes. But mutational process

Number of mutations of certain type $m(ctx)$ will be proportional to an exposure time m_0 (measured as a number of collision events), to a probability $f(ctx)$ to introduce a mutation during a single encounter and to a frequency of a given genomic context among all possible mutations $\chi^{wg}(ctx)$. Index $\cdot^{wg}$ here and below means "whole genome", i.e. we count a fraction of positions with given context in a whole genome. Later notation $\cdot^{ss}$ will also be used to denote site-specific estimation of characteristic (i.e. related to a reduced subset of genome). For example m^{ss} means number of mutations that should occur in a region given a mutational process (not a number of mutations which actually occured).

$$m^{wg}(ctx) = m_0 \cdot f(ctx) \cdot \chi^{wg}(ctx) \tag{1}$$

$$\chi^{wg}(ctx) = N^{wg}(ctx)/N^{wg} \tag{2}$$

From sequencing cancer genomes we know $m^{wg}(ctx)$. $N^{wg}(ctx)$ and N^{wg} are also known given we have a genome assembly.

Let's now calculate number of mutations in some region. This number of mutations depend on a total length of that region and on context distribution in it. As mutagen doesn't distinguish between positions in and out of that region we can write:

$$m^{ss}(ctx) = m_0 \cdot f(ctx) \cdot \frac{N^{ss}(ctx)}{N^{wg}} = m^{wg}(ctx) \cdot \frac{N^{ss}(ctx)}{N^{wg}(ctx)} \tag{3}$$

We can define attractiveness Q of a region to a mutational process. It shows whether it's more or less likely that mutation will hit the region under specific mutational process related to a uniform distribution of mutations along genomic positions.

$$Q = \frac{m^{ss}}{m^{ss}_{uniform}} = \frac{\sum_{ctx} m^{ss}(ctx)}{m^{wg} \frac{N^{ss}}{N^{wg}}} = \frac{\sum_{ctx} m^{wg}(ctx) \cdot \frac{N^{ss}(ctx)}{N^{wg}(ctx)}}{m^{wg} \frac{N^{ss}}{N^{wg}}} \tag{4}$$

$$Q = \sum_{ctx} \frac{m^{wg}(ctx)}{m^{wg}} \cdot \frac{N^{ss}(ctx)}{N^{ss}} \cdot \frac{N^{wg}}{N^{wg}(ctx)} \tag{5}$$

Background probability that random mutation occur inside a region is

$$p = \frac{m^{ss}}{m^{wg}} = Q \cdot \frac{N^{ss}}{N^{wg}} \tag{6}$$

Number of mutations in a region is distributed binomially with probability p, that allows us to estimate p-value of getting a certain number of mutations using cumulative binomial distribution.

Given that region is defined as an ensemble of transcription factor binding sites we can estimate $\frac{N^{ss}(ctx)}{N^{ss}}$ using positional frequency matrix S.

$$\frac{N^{ss}(ctx)}{N^{ss}} = \frac{1}{l}\sum_{i=1}^{l} S_i(ctx) \tag{7}$$

Not all mutations are equally detrimental, so just a number of mutations doesn't tell us about an effect of mutational process on affinities of sites. We assess an effect of mutational process on distribution of motif sites in a way similar to [Impact of cancer mutational signatures on transcription factor motifs in the human genome]. In that work authors consider each mutation type in each genomic k-mer. Our approach differs: we don't try to assess weight of each k-mer in the genome, instead we model an ensemble of all sites using the definition of positional frequency matrix. It tells us that given w is a site, w has probability $S_i(\alpha_i)$ to have nucleotide α_i at position i and that probability doesn't depend on other nucelotides. We consider only sites under this model thus we can't consider events of novel site creation but can consider affinity gain or loss for existing sites.

Also we introduce a notation $S_i(ctx)$ that describes a probability to have context ctx centered at position i. For trinucleotide contexts $ctx = \alpha\beta\gamma$ we can write:

$$S_i(\alpha\beta\gamma) = S_{i-1}(\alpha)S_i(\beta)S_{i+1}(\gamma) \tag{8}$$

Mutation of a certain context has three different directions (α_i can be converted in any different nucleotide). We assume that mutation direction probability doesn't depend on anything but original context.

$$f(ctx \to ctx') = f(ctx) \cdot \frac{m(ctx \to ctx')}{m(ctx)} \tag{9}$$

Now let's consider an ensemble of K^{ss} sites distributed according to a frequency matrix S and let mutational process make $K^{ss} \cdot \rho$ mutations in these sites. ρ can be treated as a measure of (local) intensivity of a mutational process on these sites. ρ is small enough to neglect probability that some position of an ensemble will be mutated twice.

We denote the probability that a mutation will hit certain position with certain context as $p_i(ctx \to ctx')$:

$$p_i(ctx \to ctx') = f(ctx \to ctx') \cdot S_i(ctx) \tag{10}$$

Our goal is to assess frequency matrix $\widetilde{S}$ of an ensemble under action of a mutational process. We count number of positions of certain context after action of mutational process:

$$K^{ss} \cdot \widetilde{S}_i(ctx) = K^{ss} \cdot S_i(ctx) - \sum_{ctx'} K^{ss} \cdot \rho \cdot p_i(ctx \to ctx') + \sum_{ctx'} K^{ss} \cdot \rho \cdot p_i(ctx' \to ctx) \tag{11}$$

$$\Delta S_i(ctx) := \widetilde{S}_i(ctx) - S_i(ctx) \tag{12}$$

$$\boxed{\Delta S_i(ctx)/\rho = \sum_{ctx'} p_i(ctx' \to ctx) - \sum_{ctx'} p_i(ctx \to ctx')} \tag{13}$$

In particular case of trinucleotide contexts with substitutions altering central nucleotide we can rewrite it as follows:

$$\Delta S_i(\alpha\beta\gamma)/\rho = \sum_\delta p_i(\alpha\delta\gamma \rightarrow \alpha\beta\gamma) - \sum_\delta p_i(\alpha\beta\gamma \rightarrow \alpha\delta\gamma) \tag{14}$$

Using assumption of independence of site positions we can calculate shift of central nucleotide distribution:

$$\Delta S_i(\beta) = \sum_{\alpha,\gamma} \Delta S_i(\alpha\beta\gamma) \tag{15}$$

$$\Delta S_i(\beta)/\rho = \sum_{\alpha,\delta,\gamma} \left(p_i(\alpha\delta\gamma \rightarrow \alpha\beta\gamma) - p_i(\alpha\beta\gamma \rightarrow \alpha\delta\gamma)\right) \tag{16}$$

$$\boxed{\Delta S_i(\beta) = \rho \cdot \sum_{\alpha,\gamma} \left(\sum_\delta S_i(\alpha\delta\gamma) \cdot f(\alpha\delta\gamma \rightarrow \alpha\beta\gamma) - S_i(\alpha\beta\gamma) \cdot \sum_\delta f(\alpha\beta\gamma\ \rightarrow \alpha\delta\gamma) \right)} \tag{17}$$

We derived direction in which mutational process will shift an ensemble of sites. For a fixed exposure (e.g. $\rho = 1$) we can get a frequency matrix describing altered ensemble of sites. With initial and altered PFMs we can estimate similarity between site ensembles. For instance we can calculate Jaccard similarity between sets of top-scoring words in each ensemble.

Also we can calculate mean weight of sites in ensemble given some weight matrix W. Weight matrix obtained by initial frequency matrix looks like a legit choise.

$$\mathbb{E}_{m \sim M} W(m) = \sum_i \sum_\alpha W_i(\alpha) \cdot S_i(\alpha) \tag{18}$$

Change in mean weight characterizes whether site affinity to a TF was gained or lost under action of mutational process:

$$\Delta\mathbb{E}_{m \sim M} W(m) = \sum_{i,\alpha} W_i(\alpha) \cdot \Delta S_i(\alpha) \tag{19}$$

ToDo: - написать про профиль позиционных предпочтений. - про нулевые фланки - убить кусок про изменение среднего веса и рассказать про среднее изменение веса и среднеквадратичное изменение веса.

Где у нас проблема: можно придумать процесс, который одни сайты усиливает, а другие с той же вероятностью - ослабляет. Средний вес не меняется. Даже частотная матрица может не поменяться. Однако отсюда не следует, что мутационный процесс не ломает сайты.

Action of mutational process on motif

Transcription factors (TF) play crucial role in transcription regulation. They bind promoters and enhancers at specific positions named transcription factor binding sites (TFBS) and facilitate or restrict ablility of RNA-Polymerases to initiate transcription. Binding sites recognized by a transcription factor share a similar sequence - binding motif. Typically binding motif is represented by a nucleotide frequency matrix built from multiple gapless alignment of a set of binding sites.

As was shown in [doi:10.1038/nature12477] different mutation types are prevalent in different cancer types. There was coined a notion of mutational signature. Mutational signature refer to a fraction of substitutions which occur in a specific context. Trinucleotide contexts were considered, such as C → T in context ACA (ACA → ATA), but we can take longer contexts in the same way.

Progress in next generation sequencing allowed to track mutations in the whole genome. In the last years lots of efforts were directed to tackle mutations in regulatory regions such as promoters and enhancers. Most of somatic mutations hit DNA outide of coding regions but not that much cases with recurrent mutations in regulatory regions were found and even less were experimentally verified [TERT promoter, ...]. Some studies indicate that somatic mutations in regulatory regions may experience both positive and negative selection pressure [...] but evidences of such selection are still scarce. One of the major obstacles to reveal selection pressure is a difficulty in defining background level of mutational burden.

In papers [Negative selection, Snyder] approach was to generate random set of mutations in binding sites. But different ways to generate a set of random mutations lead to different results. We try to overcome restrictions of sampling random mutations by modeling alteration of an ensemble of all possible sites under action of mutational process.