Article: A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures

Journal: Plos Genetics

Authors: Shiraishi, Y. et al.

Year: 2015

Citiations: 46

link: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005657#sec015

جهش های جسمی (غیر ارثی) در طول زندگی در سلول های بدن ما به دست می‌آید. برخی از این جهش‌ها می‌توانند منجر به سرطان شوند. سرطان‌های مختلف و حتی موارد مختلف از یک نوع سرطان می‌توانند الگوهای متفاوتی از جهش‌های جسمی را نشان دهند. این الگوهای خاص به عنوان "امضاهای جهش‌زا" شناخته شده‌اند. هر امضای جهش ممکن است با نوع خاصی از مواد سرطان زا ، مانند دود دخانیات یا نور پرتوی بنفش مرتبط باشد. بنابراین ، شناسایی امضاهای جهش‌زا این پتانسیل را دارد که عوامل سرطان‌زا جدید را شناسایی کرده و بینش جدیدی در مورد مکانیسم‌ها و دلایل سرطان ایجاد کند.

اگرچه بیشتر جهش‌ها بی‌ضرر هستند (بنام "جهش گذرگر") ، بخش کوچکی از جهش‌ها در بعضی از مکان‌های خاص در ژن‌های سرطانی ("جهش محرک") بر رشد سلولی تأثیر می‌گذارد ، و باعث تکثیر بی‌رویه ، تهاجم بافت‌ها و کمک به تبدیل سلول سالم به سلول سرطانی می شود.

مطالعات ژنوم سرطان به طور معمول بر شناسایی جهش‌های محرک تمرکز دارد ، تا به درک مکانیسم رشد سرطان کمک کند. با این حال ، جهش‌های گذرگر همچنین می‌تواند اطلاعات مهمی را به همراه آورد ، زیرا اغلب الگوهایی ("امضاهای جهش") را نشان می‌دهند که می‌توانند درک بهتری از نیروهایی که باعث ایجاد جهش‌های جسمی شدند، فراهم کند.

بررسی ۷۰۳۴ نمونه سرطانی با ۳۰ نوع سرطان مختلف

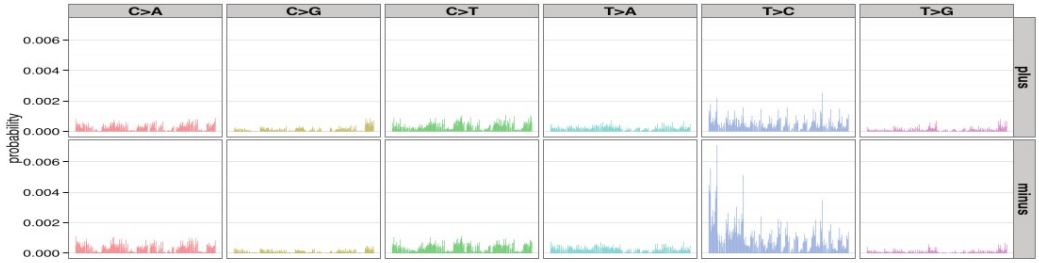
محدودیت‌های روش NMF و Emu:

۱. به علت استفاده از روش بوت‌استرپینگ در NMF و حذف انواع جهش‌هایی که تعداد کمی جهش برای آن‌ها وجود دارد، مجبور به محدود کردن دامنه امضاهای جهش‌زا هستند.

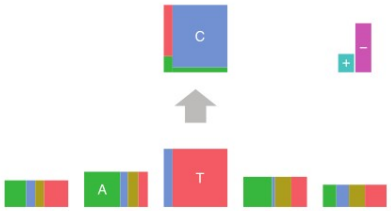
هرچه نوکلئوتیدهای مجاور یک جهش را زیاد کنیم، پارامتر های مدل نیز افـزایش پیـدا می‌کند. افزایش تعداد پارامترها در NMF باعث می‌شـود در به‌روزرسـانی دو مـاتریس مجهول ثبات و پایداری ایجاد نشود و یک جواب نخواهیم داشت.
۲. هر امضا به عنوان توزیع احتمالاتی در یک فضا با پارامتر بالا به سختی قابل تفسـیر خواهد بود.

روش جدید:
مرحلـه ۱: ساده‌سـازی مدل‌سـازی امضـاهای جهش بـا تجزیـه آن‌هـا بـه «ویژگی‌هـای جهش» مجزا
برای مثال، نوع جایگزینی جهش یک ویژگی و نوکلئوتیدهای مجـاور جهش یـک ویـژگی دیگر
مرحلـه ۲: بـا اسـتفاده از یـک مـدل احتمـالاتی بـرای امضـاها بـا فـرض استقلال بین ویژگی‌ها ، از این تجزیه استفاده می‌کنیم.
در نتیجه تعداد پارامترهای هر امضا کاهش پیدا می‌کند.
از آنجـا کـه هـدف از امضـای جهش ، بـه نـوعی ، بدسـت آوردن وابسـتگی در میـان ویژگی‌ها است ، فرض استقلال در میان ویژگی‌هـای یـک امضـا ممکن اسـت در ابتـدا غیرطبیعی به نظر برسـد. بـا این حـال ، اسـتفاده از آن در اینجـا شـبیه بـه "مـدل‌های ماتریس وزن موقعیت (PWM)" است که برای مدل‌سـازی موتیف‌هـای اتصـال فـاکتور رونویسی بسیار موفق بوده است. در واقـع ، سـهم مهمی در ایجـاد ایـده بـرای ارائـه نمایش قابل درک‌تر برای امضـاهای جهش ، مشـابه "sequence logos" اسـت کـه بـرای نمایش موتیف‌های اتصال اسـتفاده می‌شـود. در نهـایت ، همچـنین ارتبـاط نزدیـک بین مـدل‌های امضـای جهش‌زا و مـدل‌های "mixed-membership" ، همچـنین بـه عنـوان "مدل‌های admixture" یا مدل های "Latent Dirichlet Allocation" که در ژنتیک‌های جمعیت و کاردبردهای خوشه‌بندی document استفاده می‌شـود ، مـورد توجـه قـرار می‌دهیم. این ارتباطات باید برای جزییات بیشتر روشهای محاسباتی و آماری برای تشخیص امضـای جهش سرطان مفید باشد.
مدل جدید برای امضاهای جهش‌زا:
الگوی جایگزینی ممکن (C>A,C>G, C>T, T>A, T>C, T>G) ← ۶
نوکلئوتیدهای مجاور الگو جهش ← ۴ × ۴
رشته (مثبت یا منفی) ← ۲

در نهایت ۱۹۲ الگو به وجود می‌آید.
در مدل‌های قبلی امضاهای جهش‌زا با استفاده از توزیع غیرشرطی بـر روی الگوهـای جهش به تعداد M مشخص می‌شد. بنابراین هر امضا توسـط یـک بـردار احتمـالاتی بـه طور M مشخص می‌شد. مشکلی که در این روش وجود دارد ، تعداد زیاد پارامترها به ازای هر امضا است. با در نظـر گـرفتن تنهـا یـک نوکلئوتیـد مجـاور M=96 اسـت و بـا اضافه کـردن نوکلئوتیدهای مجاور بیشتر، M به صورت نمایی افـزایش پیدا می‌کند ($4^{2n} \times$). تعداد زیاد پارامترها دو مشکل ایجاد می‌کند. ۱) تخمین پارامترهای امضا از نظر آماری غیرپایدار است ۲) تفسیر امضاها به سختی صورت می‌گیرد.
با جدا کردن هر الگـوی جهش بـه عنـوان یـک «ویـژگی» و فـرض مسـتقل بـودن این ویژگی‌ها، می‌توانیم تعداد پارامترها را کاهش دهیم.
در مورد بالا ۴ ویژگی داریم (جایگزینی، نوکلئوتید مجاور چپ، نوکلئوتید مجاور راست، نوع رشته). و هر امضا را توسط یک توزیع احتمال از هر ویژگی توصیف می‌کـنیم (بـا فـرض اسـتقلال، توزیـع ویژگی‌هـا را بـرای بدسـت آوردن توزیـع امضـا در هم ضـرب می‌کنیم). در نتیجه تعداد پارامترهای مورد نیاز، ۵+۳+۳=۱۱ پارامتر به جـای ۹۶-۱= ۹۵ است. با توسعه مدل برای افزایش تعداد نوکلئوتید مجاور تنها ۶ n + ۵ پارامتر بـه جای ۱- $4^{2n} \times$ مورد نیاز است.
شکل زیر نمایش جدیـد از امضـایی کـه قبلاً مشـخص شـده اسـت را نشـان می‌دهـد. تصویری قابل تفسیر از امضا را ارائـه می دهـد کـه مشـابه sequence logos اسـت. عناصـر اصـلی این امضـای جهش نسـبت بـه مـدل قبلی (بـا نمـایش مـیزان احتمـال) راحت‌تر و سریع‌تر مشاهده می‌شود.
شکل زیر روشی را نشان می دهد که نمایش جدید امضاها می تواند بـه سـادگی یـک امضا را که قبلاً مشخص شده است ضبط کند و تصـویری بـه راحـتی قابـل تفسـیر از امضا را ارائه می دهد که مشابه sequencing logos است.



(B)



(C) substitution pattern

C>A	C>G	C>T	T>A	T>C	T>G
0.033	0.000	0.106	0.062	0.799	0.000

flanking bases

position	A	C	G	T
-2	0.340	0.143	0.139	0.378
-1	0.567	0.119	0.166	0.148
+1	0.451	0.048	0.255	0.246
+2	0.211	0.198	0.252	0.339

transcription strand

plus strand	minus strand
0.272	0.728

مدل ریاضی:

جهش با L ویژگی جهش: $m = (m_1, m_2, \ldots, m_L)$

مقادیر ممکن هر ویژگی: $M = (M_1, M_2, \ldots, M_L)$

مثال: ۶ الگو جایگزینی به همراه ۲ نوکلئوتید مجاور در هر سمت جهش

$M = (6,4,4,4,4,4)$

mutation pattern	full model	independent model
L	1	3
M	(96)	(6, 4, 4)
ApCpA → ApCpA	(1)	(1, 1, 1)
ApCpC → ApApC	(2)	(1, 1, 2)
ApCpG → ApApG	(3)	(1, 1, 3)
ApCpT → ApApT	(4)	(1, 1, 4)
CpCpA → CpApA	(5)	(1, 2, 1)
...	...	...
ApCpA → ApGpA	(17)	(2, 1, 1)
...	...	...
TpTpT → TpGpT	(96)	(6, 4, 4)

I نمونه ژنوم سرطانی

J_i تعداد جهش‌های مشاهده شده در ژنوم سرطانی i ام

$x_{i,j} = (x_{i,j,1}, \ldots, x_{i,j,L}), (i = 1, \ldots, I, j = 1, \ldots, J_i)$ بردار ویژگی جهش مشاهده شده برای j ام جهش از i امین ژنوم سرطان که در آن $x_{i,j,l} \in \{1, \ldots, M_l\}$

فرض کنیم هر جهش از یکی از K امضا جهش ممکن به وجود آید.

$q_{i,k}$ نسبت امضای k در نمونه i است.

$q_i = (q_{i,1}, q_{i,2}, \ldots, q_{i,K}) \in \Delta^K, (i = 1, ..., I)$ که در آن $\Delta^S = \{(t_1, \ldots, t_s) : ts > 0, \sum t_s = 1\}$

هر امضا جهش توسط بردارهای پارامتر $F_k := (f_{k,1}, \ldots, f_{k,L})$ که در آن $f_{k,l}$ یک بردار احتمالاتی از ویژگی l ام در امضا k ام است. $f_{k,l} = (f_{k,l,1}, \ldots, f_{k,l,Ml}) \in \Delta^{Ml}$

مدل مولد برای جهش‌های مشاهده شده $\{x_{i,j}\}$ در هر نمونه سرطانی می‌تواند در یک فرایند دو مرحله‌ای توصیف شود:

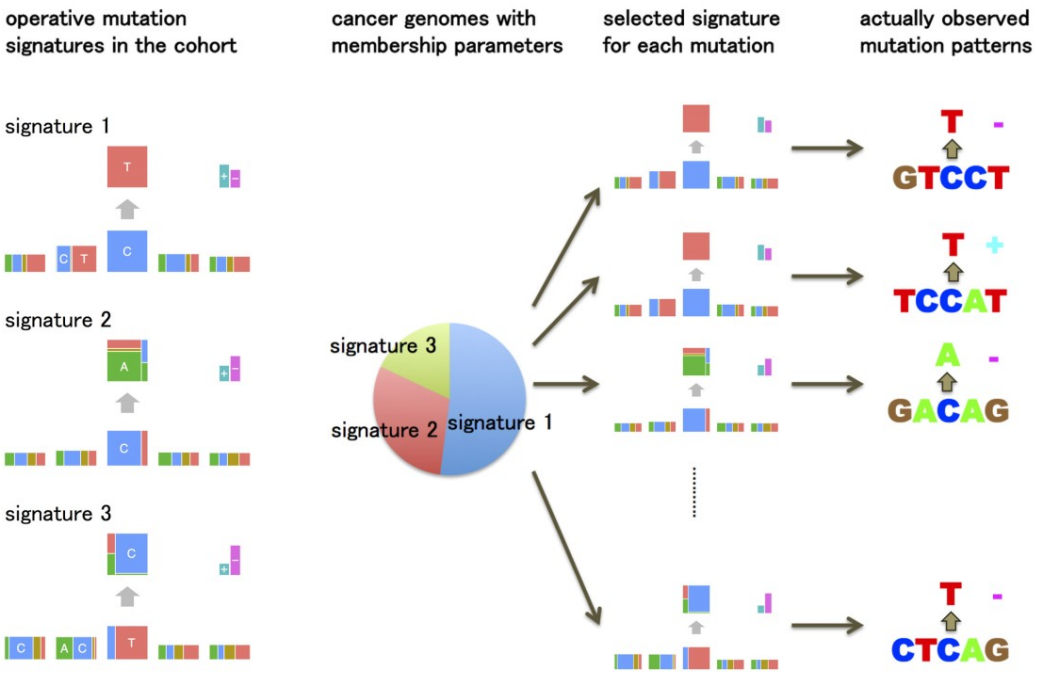
۱. تولید $z_{i,j} \sim \text{Multinomial}(q_m)$ که در آن $z_{i,j} \in \{1, \ldots, K\}$ نشان‌دهنده امضا جهش‌زا ناشی از j ام جهش در i ام نمونه سرطانی

۲. برای هر $l(= 1, \ldots, L)$ ، تولید $x_{i,j,l} \sim \text{Multinomial}(f_{z_{i,j},l})$ بنابراین

$$P(x_{i,j,l} = m \mid z_{i,j} = k, F_{k,lm}) = f_{k,lm}$$

پارامتر های کلیدی در این مدل، میزان مشارکت امضاها برای هر نمونه، q_i و پارامترهای امضا جهش‌زا، F_k است.

پیاده‌سازی روش بر روی داده‌های شبیه‌سازی شده نشان می‌دهد که اگر تعداد جهش‌ها و نمونه‌ها به اندازه کافی فراهم باشد، می‌توانیم امضاهای جهش‌زا با دقت بالا دوباره تولید کنیم.



داده‌های شبیه‌سازی شده:

الگو جهش: یک جایگزین و دو نوکلئوتید مجاور در هر طرف

تعداد ژنوم سرطانی: (I = 10, 25, 50, 100)

تعداد جهش در هر ژنوم سرطانی: (J = 10, 25, 50, 100, 250, 500, 1000)

تعداد امضاهای جهش‌زا: K = 5

پارامترهای ویژگی جهش و پارامترهای عضویت توسط توزیع دیریکله تولید می‌شوند:

$$\boldsymbol{f}_{k,l} \sim \mathrm{Dir}(\alpha \mathbf{1}),\ k = 1, \cdots, K,\ l = 1, \cdots, L.$$

$$\boldsymbol{q}_{i,k} \sim \mathrm{Dir}(\gamma \mathbf{1}),\ i = 1, \cdots, I,$$

که در آن‌ها α و γ میزان پراکندگی را برای پارامترهای امضا جهش و پارامترهای عضویت کنترل می‌کنند.

وقتی γ کوچک است، بیشتر نمونه‌ها، بیشتر جهش‌هایشان از یک امضا ایجاد شده‌اند (اما نه یک امضا برای هر نمونه)

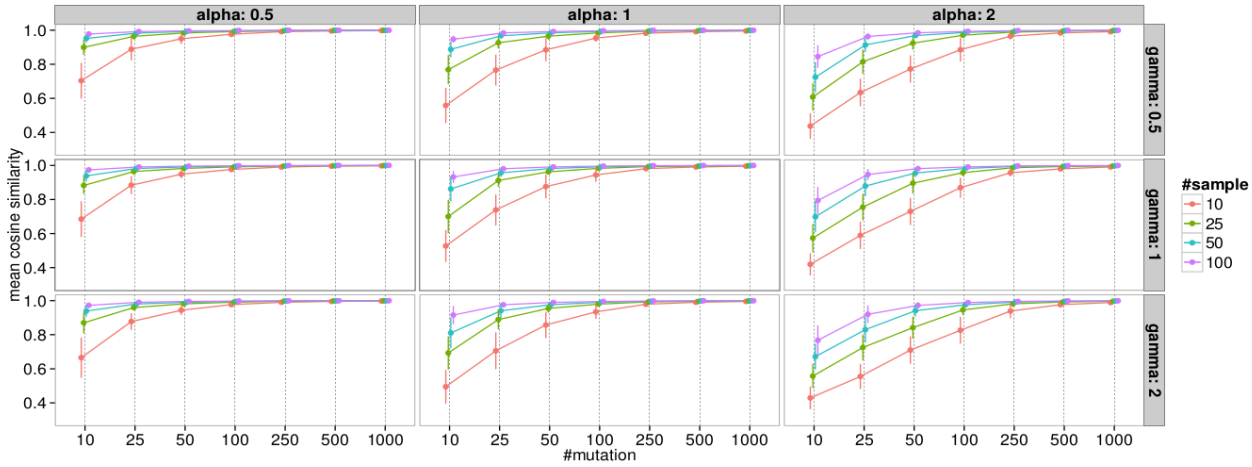
وقتی γ بزرگ است، نمونه‌ها، جهش‌هایشان تقریباً به یک اندازه از همه امضاها ایجاد می‌شوند.

با افزایش تعداد نمونه‌ها و جهش‌ها، دقت بالاتری داریم.

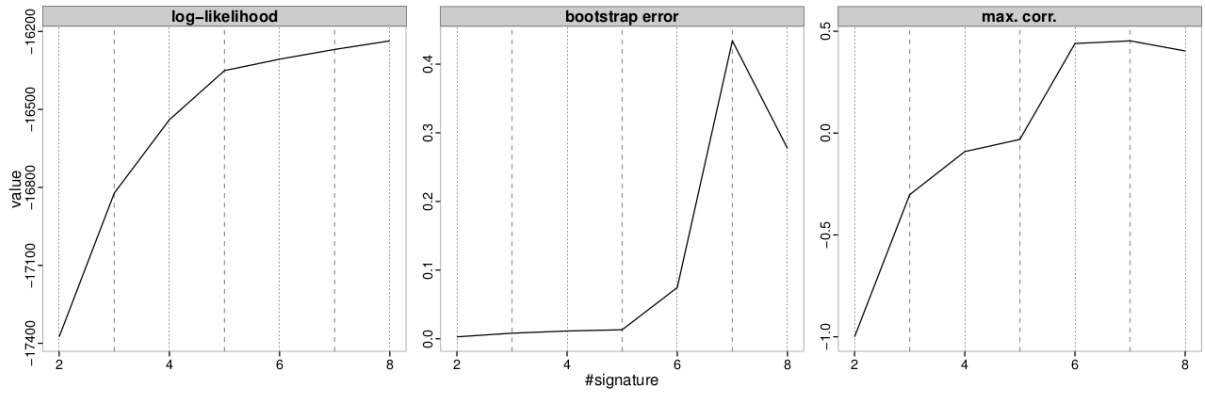
هر چه پراکندگی ویژگی جهش کاهش پیدا می‌کند (α افزایش)، دقت کم‌تر می‌شود.

با افزایش α ، امضاهای جهش بیشتر فازی می‌شوند، توده احتمالی در تعداد زیادی از الگوهای جهش گسترده می‌شود و نتیجه‌گیری سخت‌تر می‌شود (امضا جهش‌زا با اطلاعات مفید کمتر).

در مقایسته، دقت به پراکندگی عضویت فردی (γ) نسبتاً حساس نیست و حتی وقتی بیشتر افراد جهش هایشان از امضاهای متفاوت بیشتری ایجاد شده باشد، همچنان تخمین با دقت بالا به دست آمد.



در بیشتر موارد افزایش log-likelihood در k=5 امضا جهش‌زا متوقف می‌شود. هر چند که خطای استاندارد تخمین پارامترها از k=5 به بعد افزایش می‌یابد. نتیجه می‌گیریم که بررسی trade-off بین likelihood و خطای استاندارد برای انتخاب تعداد امضاهای جهش‌زای مناسب، می‌تواند مفید باشد.



تخمین این پارامترها با بیشینه کردن likelihood در الگوریتم EM انجام می‌شود.

روش‌ها:

تخمین پارامتر:

تخمین $\{ f_{k,l} \}$ و $\{ q_i \}$ با دانستن داده جهش موجود $\{ x_{i,j} \}$

استفاده از الگوریتم EM برای بیشینه کردن likelihood

$g_{i,m}$ تعداد جهش در نمونه سرطانی i ام با بردار ویژگی جهش m

مرحله E: محاسبه متغیرهای کمکی $\theta_{i,k,m}$

متغیری برای احتمال شرطی برای $z_{i,j}$ با دادن پارامترها و ویژگی‌های جهش $x_{i,j}$

$$\theta_{i,k,\boldsymbol{m}} = \Pr\left(z_{i,j} = k \middle| \boldsymbol{x}_{i,j} = \boldsymbol{m}, \{\boldsymbol{f}_{k,l}\}, \{\boldsymbol{q}_i\}\right)$$

این احتمال شرطی فقط به مقادیر ویژگی جهش $\mathrm{m} = (\mathrm{m}_1, \cdots, \mathrm{m}_L)$ وابسته است و نه شماره جهش j

$$\theta_{i,k,\mathbf{m}} = \frac{q_{i,k} \prod_{l=1}^{L} f_{k,l,m_l}}{\sum_{k'=1}^{K} q_{i,k'} \prod_{l=1}^{L} f_{k',l,m_l}}.$$

مرحله M: به روزرسانی پارامترهای $\{\mathrm{f}_{k,l}\}$ و $\{\mathrm{q}_{i,k}\}$

$$f_{k,l,p} = \frac{\sum_{\mathbf{m}:m_l=p} g_{i,\mathbf{m}} \theta_{i,k,\mathbf{m}}}{\sum_{p'} \sum_{\mathbf{m}:m_l=p'} g_{i,\mathbf{m}} \theta_{i,k,\mathbf{m}}},$$

$$q_{i,k} = \frac{\sum_{\mathbf{m}} g_{i,\mathbf{m}} \theta_{i,k,\mathbf{m}}}{\sum_{k'} \sum_{\mathbf{m}} g_{i,\mathbf{m}} \theta_{i,k',\mathbf{m}}}.$$

پیشنهادات:

ارتباط بین تشخیص امضاهای جهش‌زا و استفاده از مدل‌های mixed-membership در زمینه‌های دیگر به‌خصوص تجزیه و تحلیل admixture و خوشه‌بندی document می‌تواند در بهبود مدل ما کمک کند. برای مثال، استفاده از توزیع پیشین همبسته (correlated) بر روی امضاها، این اجازه را به بعضی امضاها می‌دهد - شاید در سرطان‌های مختلف - تا شاید به امضای دیگر شباهت داشته باشند (گرچه یکسان نیست). به طور کل، استفاده از توزیع‌های پیشین مشخص یا شرط‌های penalty، مانند sparsity-promoting penalties و احتمال پیشین determinantal point process (DPP) دقت و تفسیر می‌تواند بهبود یابد. علاوه بر این ، با بزرگتر شدن مقیاس داده های ژنوم سرطان ، ممکن است رویکردهای محاسباتی پیچیده تر برای تخمین پارامترها ضروری شود. ما می توانیم تعدادی از تکنیک های محاسباتی مانند آنهایی که از الگوریتم EM استفاده می‌کنند ، sequential quadratic programming، و Gibbs sampling، و روش‌های variational اقتباس کنیم. در نهایت برای حل مسأله تعداد امضاها می‌توانیم روش‌مان را به Hierarchical Dirichlet processes توسعه دهیم.

تمرکز این مقاله بر روی جهش‌های نقطه‌ای جایگزین است اما بسیاری از انواع جهش‌ها در ژنوم سرطانی رخ می‌دهد مانند حذف و اضافه ، جایگزین‌های دوتایی، تغییرات ساختاری و تعداد تکثیرها. حذف‌های بلند می‌تواند با طول حذف و نوکلئوتیدهای مجاور، حذف‌های کوتاه با نوکلئوتیدهای حذف شده به عنوان ویژگی جهش نمایش داده شود.

در یک تعداد از امضاهای جهش‌زا تفاوت در انتشار جهش‌ها بر رشته‌های رونویسی و غیر رونویسی (transcription strand biases) دیده شده است که با فعالیت‌های رونویسی در ارتباط است. در نتیجه، برای درک بیشتر تأثیرات فعالیت‌های رونویسی بر مکانیسم‌های جهش‌زا، می‌توانیم بیان ژن و محل اتصال RNA polymerase II به ویژگی‌های جهش اضافه کنیم تا بتوان ارتباط بین تفاوت انتشار و فعالیت‌های رونویسی را توضیح دهیم. همچنین ممکن است جالب باشد که یک الگوی احتمالی برای امضاهای جهش در جایی بین استقلال کامل و فرض عدم استقلال تدوین کنید ، برای مثال، استفاده از ایده های مشابه با ایده‌های [36]ـ که از یک ساختار Markovian برای محل‌های اتصال فاکتور رونویسی استفاده می‌کند. این ممکن است به بهبود انعطاف‌پذیری مدل‌سازی امضاهای جهش کمک کند در حالی که تعداد پارامترها را متوسط نگه دارید.

اگرچه ما معتقدیم که روشهای جدید ما قبلاً در مقایسه با رویکردهای موجود ، استفاده‌های مفیدی دارند ، اما این روش‌ها در آینده قابلیت‌های بیشتری دارند تا در تحلیل امضای جهش‌زا با سایر داده‌های زمینه‌ای، از جمله داده‌های اپی ژنتیکی آمیخته شوند. این مهم است ، زیرا تراکم جهش موضعی با تعدادی از عوامل ژنومی و اپی ژنتیکی مانند محتوای GC ، توالی های مکرر (repeat sequences) ، قابلیت دسترسی و تغییرات کروماتین (chromatin accessibility and modifications) و زمان بندی تکثیر (replication timing) مرتبط است [37–40].ـ یک مطالعه جدید نشان داد که اطلاعات اپی ژنتیکی در انواع سلول منشا تومورهای مربوطه [41] پیش بینی کننده تراکم جهش محلی است. طیف گسترده ای از داده های اپی ژنتیکی از بسیاری از انواع سلول در حال حاضر در دسترس است ، و جالب خواهد بود برای ادغام این عوامل اپی ژنتیک در تجزیه و تحلیل امضای جهش برای کمک به درک چگونگی این عوامل اپی ژنتیکی بر آسیب DNA و مکانیسم های ترمیم تأثیر می گذارد. کار ما در اینجا یک روش رو به جلو برای انجام این کار ارائه می دهد: داده های اپی ژنتیکی را می توان به سادگی به عنوان ویژگی هایی به امضای جهش اضافه کرد. این پتانسیل

برای بهبود دقت در تشخیص امضا و تولید بینش بیولوژیکی جدیـد است. مـا معتقـدیم که ارزش و تأثیر کار ما ، و به طـور خـاص رویکـرد پیشـنهادی مـا بـرای مـدل سـازی امضاهای جهش از طریق ویژگی‌های مستقل ، با ویژگی های بیشتر در تجزیه و تحلیل رشد می کند.

36. Zhao X, Huang H, Speed TP. Finding short DNA motifs using permuted Markov models. J Comput Biol.
2005; 12(6):894–906. doi: 10.1089/cmb.2005.12.894 PMID: 16108724

37. Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in
human cancer cells. Nature. 2012 Aug; 488(7412):504–507. doi: 10.1038/nature11273 PMID: 22820252

38. Hodgkinson A, Chen Y, Eyre-Walker A. The large-scale distribution of somatic mutations in cancer
genomes. Hum Mutat. 2012 Jan; 33(1):136–143. doi: 10.1002/humu.21616 PMID: 21953857

39. Liu L, De S, Michor F. DNA replication timing and higher-order nuclear organization determine single-
nucleotide substitution patterns in cancer genomes. Nat Commun. 2013; 4:1502. doi: 10.1038/ncomms2502 PMID: 23422670

40. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heteroge-
neity in cancer and the search for new cancer-associated genes. Nature. 2013 Jul; 499(7457):214–218. doi: 10.1038/nature12213 PMID: 23770567

41. Polak P, Karli R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin orga-
nization shapes the mutational landscape of cancer. Nature. 2015 Feb; 518(7539):360–364. doi: 10.1038/nature14221 PMID: 25693567