دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر
پروژه درس مبانی بیوانفورماتیک

عنوان:

تصمیم‌گیری بیان یا عدم بیان ژن از طریق اندازه‌گیری میزان RNA موجود در نمونه‌ها

نگارش:
تینا خواجه
۹۳۲۱۰۷۶۱

استاد راهنما:
دکتر سید ابوالفضل مطهری

بهمن ۹۴

چکیده :

تصمیم گیری درباره فعالیت یا عدم فعالیت ژن با اندازه‌گیری mRNAهای موجود از آن، در نمونه ها در بسیاری از مسائل همچون شبکه های تنظیم ژنی کاربرد ویژه دارد. در این موارد احتیاج است تا میزان بیان ژن را در سطوح مشخص، معمولا دو سطح بیان شده و بیان نشده را تعیین کرد. از طرفی به دلیل اینکه این مسئله در حالت باینری که عبارت است از تعیین بیان یا عدم بیان ژن با استفاده از mRNA اندازه‌گیری شده، مسئله‌ای است که توصیف ریاضیاتی مستقلی ندارد و پاسخ آن همواره با عدم قطعیت مربوط به مدل همراه می‌باشد. در این گزارش در ابتدا به معرفی مختصری از چهار روش برای مشخص کردن بیان یا عدم بیان ژن می پردازیم و در انتها با در نظر داشتن این چهار روش روشی جدید را معرفی می نماییم که حاصل از تجمیع روش‌های قبل می باشد.

مقدمه:

فعالیت‌های ژنی از طریق اندازه‌گیری میزان RNAهای موجود تعیین می‌گردند. این اندازه گیری‌ها و یا بیان ژن‌ها در آرایه‌ای متشکل از بیان ژن‌های اندازه‌گیری شده ذخیره می شوند. این آرایه را می توان به صورت زیر در نظر گرفت :

$$G = [\, G_i(j)] \; 1 \le i \le M, 1 \le j \le N$$

در این حالت هر i تعیین کننده یک ژن و هر j تعیین کننده یک بار آزمایش و اندازه گیری میزان mRNA برای یک نمونه می‌باشد. به بیان دیگر ردیف شماره i که با G_i نشان داده می‌شود بیان ژنi می‌باشد. بسته به نوع آزمایش بردار G_i نشان دهنده اندازه‌گیری این ژن در نمونه‌های مختلف و یا اندازه گیری ژن در یک نمونه و در زمان‌های مختلف می باشد. با توجه به این نوع داده‌ها مسئله تصمیم‌گیری بیان ژن به این صورت تعریف می‌شود: با در اختیار داشتن آرایه G تعیین نماییم که آیا ژن بیان شده است یا نه. به این منظور و برای پاسخ گویی به این سوال الگوریتم‌های مختلفی بوجود آمده است. بعضی از این روش ها بر پایه روش‌های آماری و برخی نیز بر پایه روش‌های تصمیم‌گیری بیان ژن به صورت قطعی استوار هستند، که در آنها یک آستانه مشخص می شود و مقادیر زیر آن آستانه معادل با بیان نشدن ژن و مقادیر بالای آن بیان‌گر بیان ژن می باشند. به صورت اختصاری ژن‌های بیان شده و فعال را با عدد یک و ژن های بیان نشده را با عدد صفر نمایش می دهیم. در نهایت الگوریتم مورد استفاده رشته‌ای از اعداد باینری را بر می‌گرداند که در آن اعداد صفر و یک وجود دارند. برای هر آزمایش مشخص می‌شود آیا یک ژن بیان شده است یا نه. عدم وجود روابط ریاضی به منظور فرموله کردن مسئله منجر می شود تا امکان ارائه دقیق نتایج وجود نداشته باشد و نتایج حاصل همواره با عدم قطعیت همراه باشند.

در این گزارش تمرکز بر روی روش های قطعی برای تصمیم گیری میزان بیان ژن می باشد. به همین منظور ابتدا چهار روش از این موارد را ارائه می دهیم، سپس به بررسی نتایج حاصل از آنها می‌پردازیم. در انتها نیز به ارائه روشی دیگر حاصل از تجمیع روش‌های ابتدایی و با رویکرد احتمالاتی می‌پردازیم.

در زیر توصیف چهار مورد از الگوریتم‌های قطعی بررسی میزان بیان ژن را معرفی می کنیم که با A,B,C,D مشخص می شود. در هر کدام از این الگوریتم ها در نهایت آستانه‌ای را به عنوان خروجی به منظور تبدیل داده‌ها به دو دسته در اختیار خواهیم داشت. ورودی و خروجی الگوریتم ها به صورت زیر تعریف می شوند:

$$\tau_i = \mathrm{X}(\mathrm{G_i}), \mathrm{X}\epsilon\{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$$

آستانه که در بالا به وسیله τ_i نمایش داده شده است مستقل از زمان نمونه برداری می باشد. همچنین در الگوریتم‌هایی که در ادامه آمده‌اند ورودی π_i همان بردار $\mathrm{G_i}$ می باشد که به صورت صعودی مرتب شده است.

الگوریتم A:

انتخاب یک آستانه به صورت سراسری و تصمیم گیری برای ژن‌های مختلف انتخاب مناسبی نمی‌باشد، به همین منظور آستانه‌های مختلفی را برای ژن‌های مختلف اندازه گیری می‌نماییم. در این الگوریتم از این ایده استفاده می‌شود که آستانه مناسب محلی است که فاصله میان ژن‌هایی با بیان کم و زیاد بیشترین فاصله را داشته باشند. بیان دیگری از این ایده بدین صورت خواهد بود که اگر داده‌ها را مرتب نماییم آستانه محلی خواهد بود که اولین پرش بزرگ در دو داده پشت سر هم صورت می‌گیرد. در این الگوریتم به مقایسه نرخ تغییرات هر دو جفت داده پشت سر هم می پردازیم که به صورت زیر مشخص می‌گردد:

$$\pi_i(\mathrm{j}+1) - \pi_i(\mathrm{j}), \mathrm{j} = 1, \dots, \mathrm{N}-1$$

$$A = \frac{\pi_i(\mathrm{N}) - \pi_i(1)}{N-1}$$

عبارت مشخص شده در بالا با عنوان A را میانگین نرخ تغییرات می نامند. آستانه در این الگوریتم عبارت است از بزرگترین نقطه در اولین زوجی که نرخ تغییرات آن بزرگتر از A است. در نهایت این الگوریتم تعداد M آستانه را به عنوان آستانه‌های هر یک از ژن های موجود مشخص می نماید. شبه کد مربوط به این الگوریتم که دارای پیچیدگی زمانی $O(MN)$ و پیچیدگی فضای $O(N)$ می باشد در ادامه آمده است.[۱]

الگوریتم B :در این الگوریتم آستانه تصمیم‌گیری با بهره‌گیری از دنباله‌ای از توابع پله مشخص می‌گردد. اولین تابع پله از طریق مرتب سازی داده‌ها به صورت صعودی به دست می‌آید، سپس توابعی با تعدا گام کمتر از روی آن بدست می‌آیند. تعداد گام‌های

الگوریتم ۱: شبه کد الگوریتم A

$S_i \leftarrow \text{sort}(G_{i,1}, G_{i,2}, \ldots, G_{i,k})$
for $j = 1$ to $k - 1$ **do**
  $D_{i,j} \leftarrow (S_{i,j+1} - S_{i,j})$
end for
$t \leftarrow (S_{i,k} - S_{i,1})/(k - 1)$
$m = \min\{j : D_{i,j} > t\}$
for $j = 1$ to k **do**
  if $G_{i,j} \geqslant S_{i,m+1}$ **then**
    $B_{i,j} \leftarrow 1$
  else
    $B_{i,j} \leftarrow 0$
  end if
end for

هر یک از توابع پله در هر مرحله کم می‌شود. تابع پله مرحله بعد با تعداد کمتر گام‌ها از طریق کمینه کردن فاصله اقلیدسی با تابع پله اولیه بدست می‌آید. محاسبه فاصله اقلیدسی بین دو تابع از فرمول زیر محاسبه می‌شود :

$$||f_1 - f_2|| = \sqrt{\sum_{x=1}^{N} \left(f_1(x) - f_2(x)\right)^2}$$

به منظور کاهش تعداد گام توابع پله در هر مرحله گام‌هایی با دو مشخصه‌ی بلند بودن گام و نیز کم بودن خطای تخمین مشخص شده و در آرایه ذخیره می‌شوند. اگر γ را به عنوان میانه این آرایه تعریف کنیم آستانه بدست آمده از این الگوریتم از فرمول زیر محاسبه می شود :

$$\tau_{i,B} = \frac{\pi_i([\gamma]) + \pi_i([\gamma] + 1)}{2}$$

الگوریتم B برای یافتن توابع گام نزدیک به π_i از روش برنامه نویسی پویا استفاده می کند. در نهایت تعداد M آستانه برای آرایه $M \times N$ در زمان $O(MN^3)$ با پیچیدگی حافظه $O(N^2)$ در اختیار ما قرار می دهد[۲].

الگوریتم C:

روشی برای پیدا کردن تابع پله ای با یک یا دو گام که به بهترین نحو با داده ها تطابق دارد. در این حالت تخمین تابع گام با استفاده از رگرسیون خطی با درجه آزاد یک یا سه صورت می گیرد. صحت این تخمین با استفاده از محاسبه p-value اندازه گیری می شود. این الگوریتم از روش SetpMiner برای محاسبه زیر مجموعه تمام توابع تک پله ای استفاده می کند. اگر

مجموعه بازگردانده شده تهی باشد در این صورت خروجی الگوریتم عدم تصمیم گیری[1] می باشد. در غیر این صورت نقطه میانی بین گام های توابع را بر می گرداند که گام ها بیشترین فاصله را دارند. پیچیدگی زمانی و حافظه ای در این حالت عبارت است از : $O(MN^2)$ و $O(N)$ [۳].

الگوریتم D:

این الگوریتم همان الگوریتم k-means classification می باشد که درآن k=2 است. در این روش داده ها به دو گروه مختلف بر اساس نزدیکی به نقاطی به نام مرکز برای هر گروه تقسیم می شوند. سپس آستانه مور نظر به صورت نقطه میانی بین دو مرکز تعریف می شود. در نهایت مثل قبل تعداد M آستانه برای آرایه $M \times N$ معرفی می شود. میزان زمان مورد نیاز در این حالت عبارت است از $O(kMN)$ و فضای مصرفی نیز برابر است با $O((N+k)M)$علاوه بر این روش روش‌های گسترش یافته تری نیز برای این الگوریتم ارائه شده است که در آن دسته بندی ها به صورت تکراری صورت می گیرد که در آنها تلاش می گردد تا اثر تغییرات و نویز موجود در داده ها کمتر شود.

حال به بررسی نتایج بدست آمده از این چهار روش می پردازیم :

معیار pearson correlation که توسط این الگوریتم ها و بر روی تعداد ۱۰۰۰ داده ۱۰ بعدی اندازه گیری شده است در جدول زیر قابل مشاهده می باشد همانطور که مشخص است این نتایج شباهتی به همدیگر ندارند. عدم شباهت بدست آمده در این حالت حاکی از عدم قطعیت در تصمیم گیری اشاره شده در این روش‌ها می‌باشد.

جدول۱: معیار همبستگی خروجی الگوریتم‌ها

	A	B	C	D
A	1	0.1007	0.1907	0.2345
B	0.1007	1	0.1810	0.7030
C	0.1907	0.1810	1	0.3237
D	0.2345	0.7030	0.3237	1

حال با در نظر داشتن الگوریتم های ارائه شده و نتایج حاصل از آنها به سراغ معرفی روشی دیگر می‌رویم[۴] : این الگوریتم دارای دو ویژگی ۱) استفاده از الگوریتم‌های ارائه شده در مرحله قبل به صورت ترکیبی و استفاده از سیستم رای گیری ۲)اختصاص میزان احتمال برای هر یک از رشته‌ها می‌باشد.

در ابتدا به معرفی مفاهیم مورد استفاده در این روش خواهیم پرداخت:

برای داده های ثبت شده در طول زمان می توان فرض کرد که G_iیک داده N بعدی، پیوسته و در بازه اعداد بین ۰ و یک می‌باشد. که در زمان‌های t_j که $t_{j-1} < t_j$ می‌باشد ثبت شده است.

$$G_i(j) = g_i(t_j), \qquad \forall j = 1, \dots, N, i = 1, \dots, M$$

[1] undecidable

فرض می کنیم که از طریق نوعی از درون یابی با نام cubic spline بر اساس G_i می‌توانیم تابعی به نام f را تخمین زده که معادل با g_i باشد. بنابراین از طریق داشتن G_i با N نقطه برداری با $2(N-1)$ نقطه را با تعریف زیر می سازیم.

$$G_{\mathrm{i}}^{1}(2\mathrm{j}) = \begin{cases} \mathrm{f}\left(\frac{\mathrm{t_{j-1}} - \mathrm{t_j}}{2}\right) & , \quad \mathrm{f}\left(\frac{\mathrm{t_{j-1}} - \mathrm{t_j}}{2}\right) \geq 0 \\ 0 & , \quad \text{otherwise} \end{cases}$$

$$G_i^1(2j-1) = G_i(j)$$

به دنباله $G_{\mathrm{i}}^{0} \ldots G_{\mathrm{i}}^{\mathrm{L}}$. دنباله درون‌یابی Cubic Spline معادل با $G_i(j)$ می گویند که به صورت زیر تعریف می شود:

۱- $G_{\mathrm{i}}^{0} = \mathrm{G_i}$

۲- $\forall\, n \geq 1$ آنگاه $G_{\mathrm{i}}^{\mathrm{n}}$ با استفاده از $G_{\mathrm{i}}^{\mathrm{n-1}}$ و طبق روابط تعریف شده در قبل ساخته می‌شود.

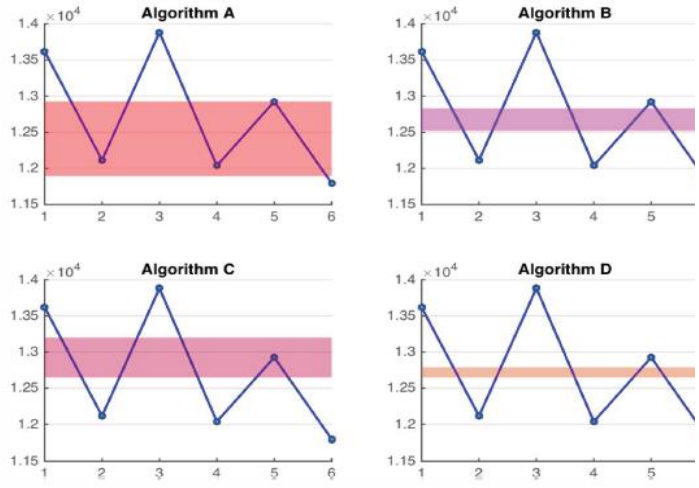
با داشتن $\mathrm{G_i}$ و همچنین انتخاب X به گونه ای که $X \in \{A, B, C, D\}$ باشد :

$$\psi_X(\mathrm{G_i}) = \{\mathrm{X}(G_{\mathrm{i}}^{\mathrm{n}}): G_{\mathrm{i}}^{0}, \ldots, G_{\mathrm{i}}^{\mathrm{L}} \mathrm{a\ SCSI}\}$$

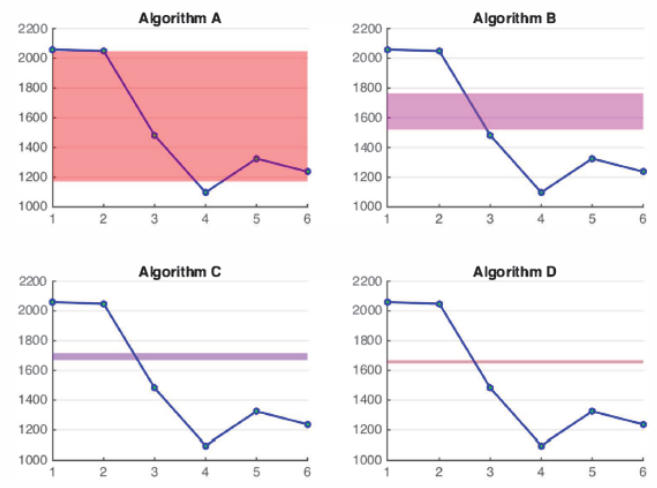
به این ترتیب میزان تغییرات در آستانه مربوط ژن $\mathrm{G_i}$ تحت الگوریتم X را می توان به صورت زیر تعریف نمود :

$$d_{\mathrm{x,i}} = |\max \psi_X(\mathrm{G_i}) - \min \psi_X(\mathrm{G_i})|$$

در تصویر زیر میزان تغییرات آستانه بدست آمده تحت الگوریتم‌های A, B, C, D ارائه شده در قبل بر روی نمودار قابل مشاهده می‌باشد .

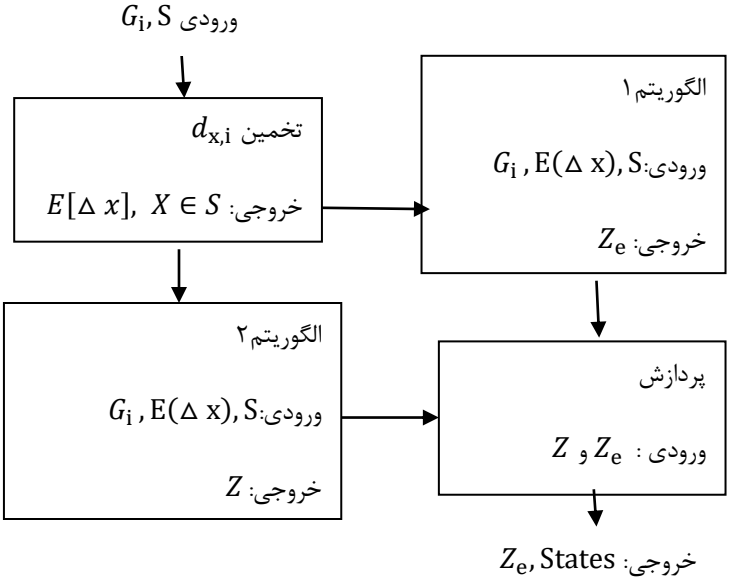


شکل۱ – میزان تغییرات آستانه در الگوریتم‌های A, B, C, D متعلق به ژن RKR1B1



شکل۲ – میزان تغییرات آستانه در الگوریتم‌های A, B, C, D متعلق به ژن ADM

روش پیش رو الگوریتمی برای انتخاب رشته وضعیت بیان ژن مانند رشته Z_e را ارائه می دهد. در ابتدا این روش مجموعه Z با تعداد اعضای N^3 بروی حروف الفبا $\{0, 1, \perp\}$ تعریف می‌شود. و در نهایت با محاسبه احتمال هر یک از اعضای مجموعه محتمل‌ترین دنباله از این مجموعه را با نام Z_e به عنوان خروجی انتخاب می‌گردد. مراحل این الگوریتم را به صورت تصویر زیر می‌توان خلاصه کرد:



۱)تخمین جابه جایی آستانه:

متغیر تصادفی Δx را جابه‌جایی صورت گرفته در آستانه تعریف می‌نماییم. حال از امید ریاضی $E(\Delta x)$ برای تخمین $d_{x,i}$ بهره می‌گیریم. از آنجایی که میزان جابه‌جایی در آستانه تقریبا محدود به بازه معرفی شده در زیر می باشد.

$$\rho_{x,i} = \max(G_i) - \min(G_i$$

بنابراین تخمین$E(\Delta x)$ بازه‌ها را، برای بازه‌های $\frac{k}{10}$ به ازای $k = 1, \dots, 10$ انجام می‌دهیم. میزان جابه جایی آستانه برای الگوریتم‌های مختلف در بازه های مختلف در جدول زیر مشخص است :

۲) انتخاب دنباله‌ای از حالات بیان ژن :

الگوریتم انتخاب دنباله متناظر با بیان ژن ها برتعریف ۴ تایی (S, Γ, R, Λ) استوار است. که در آن S بیانگر مجموعه‌ای محدود از الگوریتم ها، Γ تعیین کننده مجموعه‌ای محدود از عبارات منطقی، R قوانین تجمیع و Λ نیز عبارت است از اصول. این روش از ۴ الگوریتم معرفی شده در ابتدای گزارش استفاده می نماید. بنابراین $S \subset \{A, B, C, D\}$، $\Gamma = \{U, N\}$ که هر کدام از U و N

جدول۲: مقدار متوسط جابه‌جایی آستانه

Range	$\mathbb{E}[\Delta_A]$	$\mathbb{E}[\Delta_B]$	$\mathbb{E}[\Delta_C]$	$\mathbb{E}[\Delta_D]$
0.1	0.0297	0.0234	0.0165	0.0122
0.2	0.0490	0.0366	0.0292	0.0188
0.3	0.0699	0.0580	0.0508	0.0252
0.4	0.0845	0.0785	0.0480	0.0307
0.5	0.1107	0.0938	0.0660	0.0397
0.6	0.1356	0.0967	0.0823	0.0432
0.7	0.1435	0.1107	0.0796	0.0502
0.8	0.1795	0.1425	0.0975	0.0570
0.9	0.1949	0.1557	0.1389	0.0685
1.0	0.2244	0.1732	0.1487	0.0691

متغیر هایی دودویی هستند که به صورت زیر تعریف می‌شوند :

$$U = \left(G_i(j) + \ d_{x,i} < \tau_{x,i}\right) \wedge \left|\tau_{x,i} - G_i(j)\right| > d_{x,i}$$

$$N = |\tau_{x,i} - G_i(j)| \leq d_{x,i}$$

همانطور که مشخص است در صورتی که متغیر U مقدار یک را داشته باشد به معنای این است که ژن $G_i(j)$ بیان نشده است. همچنین در صورتی که $N = 1$ شود بدین معنا است که حالت این ژن از نظر بیان یا عدم بیان نامشخص است و نمی‌توان برای آن تصمیم گیری کرد. علاوه بر این اصول به منظور تکمیل این الگوریتم یک متغیر دودویی دیگر نیز به صورت زیر تعریف می‌نماییم :

$$E \Leftrightarrow \neg(U \vee N)$$

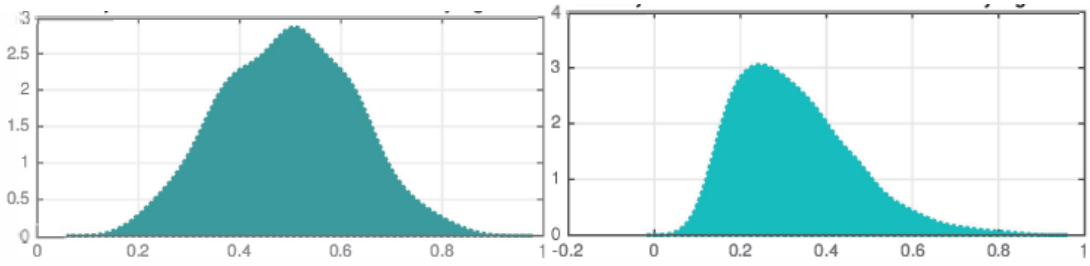
با توجه به تعریف بالا در صورتی که مقدار E برابر با یک شود معادل است با اینکه $G_i(j)$ را بیان شده در نظر بگیریم. در الگوریتم شماره دو که در زیر آمده است هر متد $x \in S$ یک آستانه $\tau_{x,i}$ را محاسبه می کند. سپس از آن برای محاسبه دقیق متغیرهای U و N بهره می گیرد. و سپس بعد از محاسبه این دو مقدار، مقدار معادل با E را از طریق فرمول زیر تعیین می نماید. در نهایت مجموعه تصمیم گیری ها از طریق انتساب E, N, U مشخص می شوند. و در ادامه صحت انتساب این تصمیم ها با استفاده از فرمول بالا و درستی آن در ارتباط تعریف شده تعیین می‌گردد. در صورتی که مقادیر معادل با یک فرمول valid نباشند مجموعه تصمیمات را inconsistent می گویند. در چنین حالتی بیان ژن را غیرقابل تصمیم گیری می نامند.

الگوریتم ۲: شبه کد انتخاب رشته بیان کننده حالت

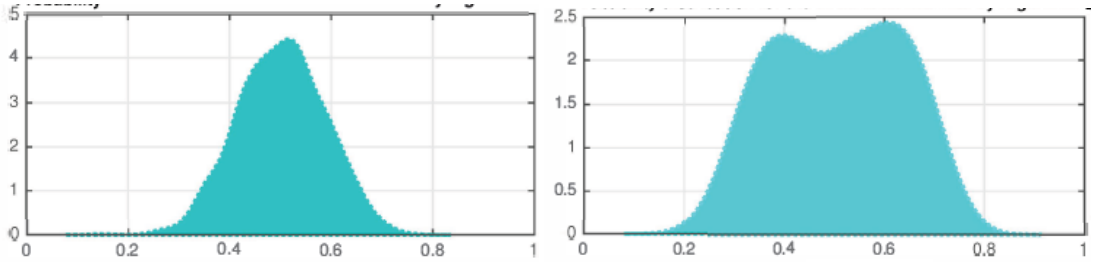
Input: $G_i(j), j = 1, \ldots, N$, S, and $d_{X,i}, \forall X \in S$
Initialize: $Z_e = \epsilon$, where ϵ is the null string
for $X \in S$ **do**
  Compute $\tau_{X,i}$
end for
for $j = 1$ to N **do**
  for $X \in S$ **do**
    Use (12) and (13) to evaluate N and U
    $E \leftarrow \neg(U \vee N)$
    Store N, U, and E
  end for
  Compute majority of N, U and E
  if $E \neq \neg(U \vee N)$ or $N = 1$ or the vote is a tie **then**
    $Z_e \leftarrow cat(Z_e, \perp)$, where cat is concatenation
  else if $E = 1$ **then**
    $Z_e \leftarrow cat(Z_e, 1)$
  else
    $Z_e \leftarrow cat(Z_e, 0)$
  end if
end for
return Z_e

۳) تعیین احتمال رشته‌های بیان کننده ژن :

در این مرحله با تعریف T_x به عنوان متغیر تصادفی آستانه بازگردانده شده توسط الگوریتم مورد استفاده، توزیع احتمال این متغیر تصادفی بر روی تمام حالات $x \in \{A, B, C, D\}$ محاسبه می‌شود. این توزیع احتمال برای الگوریتم‌ها بر روی ۱۰۰۰ داده ۱۰ نقطه ای تولید شده به صورت تصادفی محاسبه شده و در شکل های زیر برای الگوریتم‌های ارائه شده در ابتدا قابل ملاحظه می‌باشد.



ب الف



د ج

شکل۳ – توزیع احتمالات آستانه های تعیین شده به ترتیب توسط الگوریتم‌های A, B, C, D

با در اختیار داشتن مقادیر بیان ژن G_{i}، الگوریتم X و تخمینی از میزان جابه‌جایی آستانه $d_{\mathrm{x,i}}$ با استفاده از توزیع بدست آمده در این مرحله بر روی متغیر تصادفی T_{x} ، احتمال $P_{\mathrm{x,j}}$ برای وضعیت‌های متفاوت بیان ژن با توجه به فرمول‌های زیر تعریف می شود:

$$P_{\mathrm{x,j}}(0) = \mathrm{P}(\mathrm{T_x} > \mathrm{G_i(j)} + \mathrm{d_{x,i}})$$

$$P_{\mathrm{x,j}}(1) = \mathrm{P}(\mathrm{T_x} > \mathrm{G_i(j)} - \mathrm{d_{x,i}})$$

$$P_{\mathrm{x,j}}(\perp) = 1 - (P_{\mathrm{x,j}}(1) + P_{\mathrm{x,j}}(0))$$

احتمالات بالا که به ترتیب احتمال بیان نشدن ژن اندازه‌گیری شده، احتمال بیان شدن و احتمال غیر قابل تصمیم بودن به ازای نمونه j و ژن i می‌باشند به ازای تمام مقادیر $j = 1, \dots, N$ محاسبه می شود و در آرایه $3 \times N$ با نام Ω_x طبق تعریف زیر ذخیره می‌گردد.

$$\Omega_{x,i} = \left[\mathrm{P_{x,j}(Y)}\right], \mathrm{Y} \in \{0, 1, \perp\}$$

بعد از محاسبه این احتمالات و به منظور انتساب یک احتمال واحد به هریک از رشته های بیان کننده وضعیت بیان از میانگین گیری بر روی الگوریتم های مختلف و موجود در مجموعه S استفاده می‌شود.

$$\Omega_i = \left[\mathrm{P_j(Y)}\right] = \frac{1}{|\mathrm{S}|}\sum_{\mathrm{x}\epsilon\mathrm{s}} \Omega_{\mathrm{x,i}}$$

به این ترتیب به هر کدام از مقادیر $P_{\mathrm{j}}(\mathrm{Y})$ یک احتمال خاص نسبت داده می شود که در نهایت تصمیم گیری را آسان می‌کند. در نهایت با استفاده از احتمالات بدست آمده میزان احتمال هر یک اعضای مجموعه Z که با Z_{k} نمایش می‌دهیم محاسبه می شود.

$$P_{\mathrm{k}} = \prod_{\mathrm{j=1}}^{\mathrm{N}} \mathrm{P_j}\,(\mathrm{z_k(j)})$$

شبه کد الگوریتم دوم برای محاسبه احتمالات در زیر قابل مشاهده می‌باشد.

الگوریتم ۳: شبه کد رشته‌های احتمالاتی از حالت‌های بیان

Input: $G_i(j), j = 1, \ldots, N$, S, and $d_{X,i}, \forall X \in S$
for $X \in S$ **do**
  for $j = 1$ to N **do**
    Use (16), (17) and (18) to compute $\Omega_{X,i}$
  end for
end for
$\Omega_i = \frac{1}{|S|} \sum_{X \in S} \Omega_{X,i}$
for $k = 1$ to 3^N **do**
  Use the lexicographic order to produce Z_k
  $P_k \leftarrow 1$
  for $j = 1$ to N **do**
    Use Ω_i to update $P_k \leftarrow P_k P_j(Z_k(j))$
  end for
  Write Z_k, P_k in the k-th row in Z
end for
return Z

۴) پردازش :

در ابتدا بازه $[0,1]$ را به h قسمت با طول $\frac{1}{h}$ تقسیم می‌نماییم سپس با در اختیار داشتن مجموعه رشته‌های Z و احتمال هر یک از این رشته‌ها هیستوگرام مربوطه را با مشخص نمودن فرکانس تکرار رشته‌های Z با احتمال‌ها مشخص می‌نماییم. برای تعیین میزان اهمیت هر یک از اعضای مجموعه از میانگین μ و واریانس که عبارت است از فاصله احتمال هر یک از رشته ها تا میانگین، بهره می‌گیریم.

نتیجه گیری :

الگوریتم های مخلفی برای دسته بندی داده های بیان ژن در حالت باینری معرفی شده اند. در ابتدا به بررسی چهار مورد از این الگوریتم ها پرداخته شد که در تمام آنها با در نظر داشتن استراتژی خاص مرز یا آستانه ای برای دسته بندی مشخص می شد.
آخرین الگوریتم ارائه شده در این گزارش، به منظور تصمیم‌گیری درباره وضعیت بیان ژن در زمانی که مدلهایی به همراه عدم قطعیت در تصمیم‌گیری را در اختیار داریم کاربرد دارد. این الگوریتم قابلیت استفاده از مدل‌های مختلف با ویژگی‌های آماری گوناگون را دارا می‌باشد. به همین دلیل برای تصمیم گیری وضعیت بیان ژن قادر به استفاده از دیدگاه‌های متفاوت و موجود در این زمینه می‌باشد. این دیدگاه‌های متفاوت را می‌توان از طریق تفاوت در آستانه محاسبه شده برای آنها مشخص نمود.
برای استفاده بهتر از این روش می توان راه‌های دیگری نیز برای تطبیق روش و تجمیع الگوریتم‌های مختلف استفاده کرد به عنوان مثال می‌توان از استفاده از وزن برای محاسبه احتمال نهایی محاسبه شده حاصل از احتمالات بدست آمده از الگوریتم‌های مختلف، استفاده و ترکیب الگوریتم‌هایی که منجر به حصول نتیجه دقیق‌تر می‌شوند و با نگاهی دیگر استفاده از مجموعه‌ای از الگوریتم‌های

به مناسب برای هر گروه از ژن‌ها، استفاده از اطلاعات لازم برای پیش بینی مقادیر U و N استفاده نمود تا در نهایت دقت خوبی در پیش‌بینی وضعیت بیان ژن و نیز تغییرات آن با عوض شدن داده‌ها را شاهد باشیم.
الگوریتم ارائه شده در بخش آخر به منظور انتخاب توالی بیان کننده ژن و میزان عدم قطعیت آن ارائه شده است. این الگوریتم امکان استفاده از روش های مختلف و ترکیب آنها برای اندازه گیری میزان بیان ژن را در اختیار کاربران خود قرار می دهد. در نهایت برای تمام حالات ممکن برای نمونه های اندازه گیری شده یک ژن از طریق انتخاب محتمل ترین حالت ممکن بیان ژن را تعیین می‌کند.

منابع و مراجع

[۱] Shmulevich, Ilya, and Wei Zhang. "Binary analysis and optimization-based normalization of gene expression data." *Bioinformatics* 18.4 (2002): 555-565.

[۲] Hopfensitz, Martin, et al. "Multiscale binarization of gene expression data for reconstructing Boolean networks." *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 9.2 (2012): 487-498.

[۳] Sahoo, Debashis, et al. "Extracting binary signals from microarray time-course data." *Nucleic acids research* 35.11 (2007): 3705-3712.

[۴] Seguel, Jaime, and Marie Lluberes. "A unified approach to the computation and analysis of strings of gene expression states." *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE, 2015.