نشریه علمی- ترویجی محاسبات نرم

شمارهٔ ششم/ پاییز و زمستان ۹۳/ صفحهٔ ۴۲-۵۳

ارسال مقاله: ۹۳/۱۰/۱

پذیرش مقاله: ۹٤/٤/۱۳

روشی جدید برای تفکیک و طبقه‌بندی توالی‌های سرطانی و غیرسرطانی نواحی کد شده پروتئینی DNA با استفاده از الگوریتم‌های مبتنی بر LPC و SVD

امین خدائی[۱]، بهزاد مظفری تازه‌کند[۲]

[۱] دانشجوی کارشناسی ارشد، دانشکده فنی و مهندسی ـ موسسه آموزش عالی نبی‌اکرم ـ تبریز – ایران
amin.khodaei.13@gmail.com

[۲] دانشیار، دانشکده برق و کامپیوتر ـ دانشگاه تبریز ـ تبریز – ایران
mozaffary@tabrizu.ac.ir

چکیده: سرطان یکی از بیماری‌هایی است که روند رو به افزایش ابتلا به آن، محققان را به مطالعه ابعاد مختلف آن ترغیب می‌کند. منشا ژنتیکی سرطان، لزوم بررسی اجزای ژنی درونی سلول را نشان می‌دهد. در این مقاله سعی شده است تا با بهره‌گیری از تکنیک‌های تحلیل توالی‌های DNA موجود، برخی مؤلفه‌ها و ویژگی‌های خاص ژنتیکی منحصر به فرد از توالی‌های بزرگ DNA استخراج و آشکار شوند. شبیه‌سازی الگوریتم پیشنهادی روی توالی‌های عضو خاصی از بدن انسان که از یک بانک اطلاعاتی معتبر تهیه شده است، انجام گرفته است. در الگوریتم ارائه شده از روش نگاشت منحنی Z برای تبدیل رشته‌های DNA به سیگنال بهره گرفته شده است. روش پیشنهادی ارائه شده برای تحلیل سیگنال‌های توالی‌های DNA به منظور استخراج ویژگی، مبتنی بر الگوریتم پیشگوی خطی (LPC) است که از تکنیک‌های محاسباتی ماتریس کواریانس و تجزیه مقدار منفرد (SVD) به منظور انتخاب ویژگی و کاهش ابعاد بهره می‌گیرد. با مقایسه برخی پارامترهای آماری، تفکیک و تمایز خوبی بین نمونه‌های سرطانی و غیرسرطانی مشاهده می‌شود که قابل طبقه‌بندی است. این سطح تمایز بیانگر مفهوم جهش بیولوژیکی و تغییرات ژنتیکی بیماری سرطان است.

واژه‌های کلیدی: سرطان، نواحی کد شده پروتئینی DNA، مدل پیشگوی خطی، تجزیه مقدار منفرد، ماشین بردار پشتیبان

۱. مقدمه

امروزه با پیشرفت علم در زمینه‌های مختلف، تعداد و تنوع بیماری‌ها هم رو به افزایش است. یکی از این بیماری‌ها، سرطان می‌باشد که تاکنون بیش از ۵۰ نوع از این بیماری در نقاط مختلف بدن انسان شناسایی شده است. سرطان یکی از دلایل عمده مرگ و میر انسان در کل دنیا محسوب می‌شود، به طوری که طبق آمار رسمی سازمان بهداشت جهانی (WHO) سالانه بیش از ۱۰ میلیون نفر به این بیماری مبتلا می‌شوند و بیش از ۵ میلیون نفر هم بر اثر ابتلا به این بیماری جان خود را از دست می‌دهند [۱و۲]. یکی از دلایل اصلی خطرناک بودن این بیماری به عدم تشخیص سرطان در مراحل ابتدایی بر می‌گردد، چرا که در صورت تشخیص به موقع و سریع آن، قابل درمان است؛ به عنوان مثال طبق تحقیقات صورت گرفته، احتمال اینکه یک بیمار مبتلا به سرطان ریه، بیش از ۵ سال مقاومت کند، چیزی حدود ۱٤٪ است، در حالی که در صورت تشخیص این سرطان در مراحل اولیه، این احتمال تا ۷۰٪ افزایش می‌یابد [۳].

در این بین، به دلیل ماهیت ژنتیکی سرطان، مطالعه و پژوهش عناصر ژنتیکی بدن انسان می‌تواند در تشخیص این بیماری بسیار مفید باشد. مولکول‌های DNA که مؤلفه ژنتیکی اصلی سلول است، توجه محققان را در چند دهه اخیر بیشتر جلب کرده است. از بین تحقیقاتی که در این حوزه صورت گرفته است، می‌توان از مقالاتی که در حوزه نگاشت DNA ارائه شده است، شروع کرد. یکی از این موارد، تحقیق اختر با همکارانش بود که ضمن معرفی روش‌های نگاشت معمول، به مقایسه آن‌ها پرداخته‌اند [٤]. از دیگر کارهای این حوزه، پژوهش ابو-زهاد و همکارانش در سال ۲۰۱۱ و در مقاله [۵] بود که به مقایسه روش‌های مختلف نمایش DNA پرداختند و آن‌ها را از حیث معیارهای مختلف مورد ارزیابی قرار دادند. در زمینه تحلیل توالی‌های ژنومیکی خاره و همکارانش در [٦] ضمن مرور روش‌های مبتنی بر تبدیل فوریه تحلیل و پیش‌بینی نواحی کد شده پروتئینی توالی‌های DNA، به بررسی تأثیر طول و نوع پنجره و لغزش آن در طول کل توالی ژنومیکی پرداختند. صابرکاری و همکارانش در مقاله [۷]، الگوریتم جدیدی که مبتنی بر الگوریتم پیشگوی خطی (LPC) است، برای پیش‌بینی نواحی کد شده پروتئینی ارائه کردند. همچنین اختر و همکارش [۸] در سال ۲۰۰۵ با استفاده از الگوریتم تجزیه مقادیر منفرد به تحلیل توالی‌های ژنومیکی پرداخته است. مهم‌تر از همه، ستاپاتی به کمک همکارانش، با طراحی یک فیلتر خاص در [۹]، توالی‌های سرطانی و غیرسرطانی را به خوبی از هم تفکیک کردند. با مطالعه این مقالات می‌توان پی برد که می‌توان با تحلیل توالی‌های DNA سرطانی و غیرسرطانی با هر یک از روش‌های مذکور، ویژگی‌هایی متمایز از هر کدام استخراج کرد.

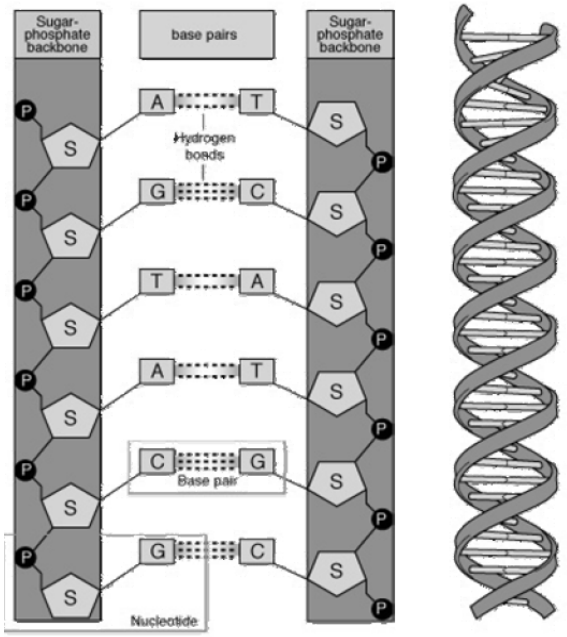
در این مقاله، ابتدا در بخش دوم برخی مفاهیم پایه و اساسی درون سلولی بیان می‌شود و در بخش سوم الگوریتم پیشنهادی بیان می‌شود و در بخش چهارم نتایج پیاده‌سازی الگوریتم پیشنهادی و نتیجه‌گیری تحقیق بیان خواهد شد.

۲. مفاهیم پایه

همانطور که اشاره شد، سرطان یک بیماری درون سلولی و ژنی محسوب می‌شود. از آنجا که بدن انسان دارای بیش از صد تریلیون سلول است، تحلیل و بررسی تمامی سلول‌ها غیرممکن به نظر می‌رسد. اما می‌توان بر اساس مولکول‌های DNA درون سلولی، آن‌ها را مورد مطالعه قرار داد. از این رو در نخستین گام برخی مفاهیم پایه زیستی برای ورود به بحث پردازش توالی‌های DNA مرور می‌شوند.

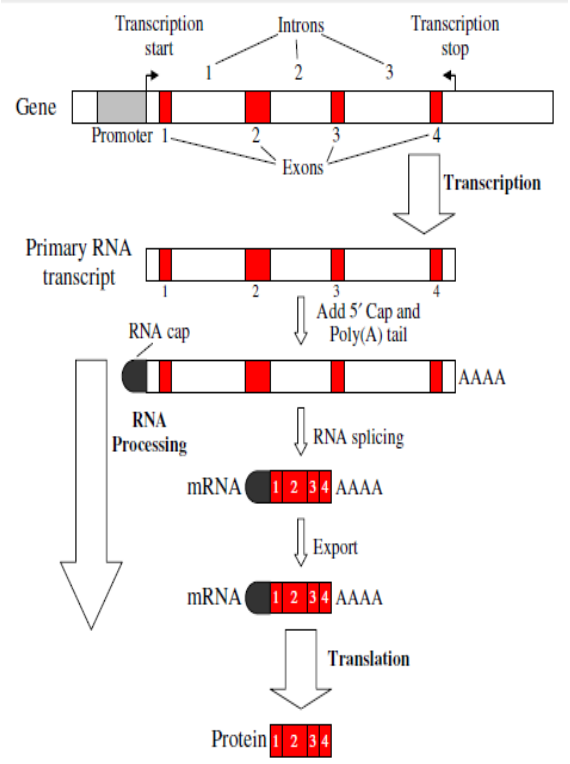
سلول به عنوان کوچکترین واحد سازنده موجودات زنده، از اجزای ریزتری تشکیل یافته است که عامل اصلی بیماری‌های ژنتیکی محسوب می‌شوند. هسته مهم‌ترین بخش تشکیل‌دهنده سلول است که درون آن کروموزوم‌هایی قرار دارند که از درهم پیچیدن رشته‌های اسید نوکلئیکی (مولکول DNA) تشکیل یافته‌اند [۱۰]. همانطور که در شکل (۱) مشاهده می‌شود هر DNA خود از به هم پیوستن دو زنجیره طولانی از نوکلئوتیدها تشکیل شده است که ساختار درونی آن‌ها، چهار نوع است که نشان‌گر تفاوت بازهای تشکیل دهنده آن‌هاست که به چهار صورت آدنین، سیتوزین، گوانین و تیمین

می‌باشند. همچنین در این دو زنجیره، علاوه بر پیوندهای بین دو نوکلئوتید مجاور، نوکلئوتیدهای روبرو هم بر اساس این قاعده که نوکلئوتیدهای آدنین با تیمین و سیتوزین با گوانین مکمل هم هستند، به هم وصل شده‌اند. شایان ذکر است که توالی‌های DNA را با حرف اول نوکلئوتید آن‌ها (A,C,G,T)، نشان می‌دهند [۱۱].



شکل (۱): ساختار درونی DNA [۹]

در هر لحظه و درون هر یک از کروموزوم‌های سلول‌های مختلف بدن انسان فعل و انفعالات شیمیایی انجام می‌گیرد که تحت عنوان «بیان ژن» شناخته می‌شود که طی این فرایندهای شیمیایی، اطلاعات ذخیره شده در DNA به پروتئین تبدیل می‌شوند که این تبدیل‌ها مطابق آنچه در شکل (۲) مشاهده می‌شود، طی مراحل نسخه‌برداری، ترجمه و سنتز پروتئین، اطلاعات پروتئین‌ها را در قالب ژن و با در کنار هم قرار دادن کدهای ژنتیکی که در ساختارهایی سه‌تایی از نوکلئوتیدها و تحت عنوان خاصیت «تناوب-۳» یا مفهوم «کودون» هستند، به صورت آمینواسید ذخیره می‌کنند [۱۲و۱۳].



شکل (۲): مراحل بیان ژن [۸]

نوکلئوتیدهای DNA یکی از سلول‌های بدن به دلایل نامشخصی می‌توانند جهش پیدا کنند که این جهش می‌تواند تغییر از یک نوع باز به باز دیگر و یا اضافه یا کم شدن یک باز باشد. این تغییر ناگهانی منجر به تغییر ساختار بیان ژن و تولید ژن‌هایی متفاوت با ژن‌های قبلی می‌شود. پروتئین‌های حاصل در صورتی که از نوع سرطانی باشند، آن سلول را تحت تأثیر قرار می‌دهند و طی تقسیمات سلولی که در هر لحظه انجام می‌گیرد، سلول‌های اطرافش را هم ممکن است از کار بیندازند. در صورتی که از این تغییرات آگاهی به عمل نیاید، این روند ادامه یافته و سلول‌های سرطانی به وسیله رگ‌های خونی، در بخش‌های دیگر بدن هم پخش می‌شوند [۱۱].

۳. مواد و روش‌ها

پس از آن که توالی‌های DNA در آزمایشگاه‌های پزشکی، با روش‌هایی خاص از بدن انسان استخراج شدند، به صورت یک مجموعه اطلاعاتی در قالب توالی‌های رشته‌ای ذخیره می‌شوند. این رشته‌های کاراکتری از چهار باز نوکلئوتیدی اشاره شده تشکیل شده‌اند.

۱،۳. نمونه‌های DNA مورد مطالعه

در بسیاری از مقالات حوزه پردازش توالی‌های DNA به طور مستقیم یا با استفاده از بانک‌های اطلاعاتی خاص یا وبسایت‌های اینترنتی از پایگاه‌داده GenBank استفاده می‌شود. یکی از پایگاه‌های اطلاعاتی که اجازه دسترسی به این داده‌ها را فراهم می‌سازد، پایگاه اطلاعاتی NCBI است. در این تحقیق هم با جستجو در بانک اطلاعاتی نوکلئوتیدی NCBI، چند توالی سرطانی و غیرسرطانی از یک نقطه خاص بدن انسان، با شماره دسترسی AC جمع‌آوری شده است [۱٤]. با توجه به این که تعداد محدودی نمونه سرطانی و غیرسرطانی از یک فرد خاص موجود است که قابلیت ارزیابی و اعتبارسنجی در انتهای کار را میسر نمی‌سازد، در این پژوهش ۱۰۰ نمونه غیرسرطانی و ۱۰۰ نمونه سرطانی مربوط به ژن‌های عامل بیماری سرطان سینه از این بانک اطلاعاتی به عنوان نمونه آزمایشی جمع‌آوری شده است تا در ادامه تحقیق، الگوریتم پیشنهادی روی این نمونه‌ها بررسی شود. در جدول (۱) چند نمونه از داده‌های مورد بررسی، نمایش داده شده است.

جدول (۱): توالی‌های DNA استفاده شده

سرطانی AC	غیر سرطانی AC
DI170721	NM_000518
DI170729	EU761578
NM_000059	AY356351
NM_006768	AF540397
NM_007294	AY027509

۲،۳. نگاشت رشته به سیگنال توالی‌های DNA

اولین چالش در تحلیل توالی‌های DNA، طرز نمایش توالی‌ها می‌باشد تا امکان اعمال پردازش‌های بعدی را فراهم سازد. در مقالات مختلفی که برای تحلیل توالی‌های DNA مطرح شده است، از روش‌های مختلفی بدین منظور استفاده شده است [۵].

در این مقاله از روش منحنی Z به منظور تبدیل توالی‌های رشته‌ای به عددی استفاده شده است. در این روش توالی‌های کاراکتری DNA بر اساس خاصیت تقارن چهاروجهی، به یک توالی سه بعدی متناظر تبدیل می‌شوند [۱۵]. به منظور تبدیل منحنی Z بر اساس روش نمایش دودویی، چهار بردار دودویی $x_A[n]$، $x_C[n]$، $x_G[n]$ و $x_T[n]$ نمایش داده شده و بر اساس رابطه یک باز پایه با دیگر بازها به صورت آنچه در رابطه (۱) نشان داده شده، محاسبه می‌شود:

$$\begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} = 2 \times \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} x_A[n] \\ x_C[n] \\ x_G[n] \\ x_T[n] \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad (۱)$$

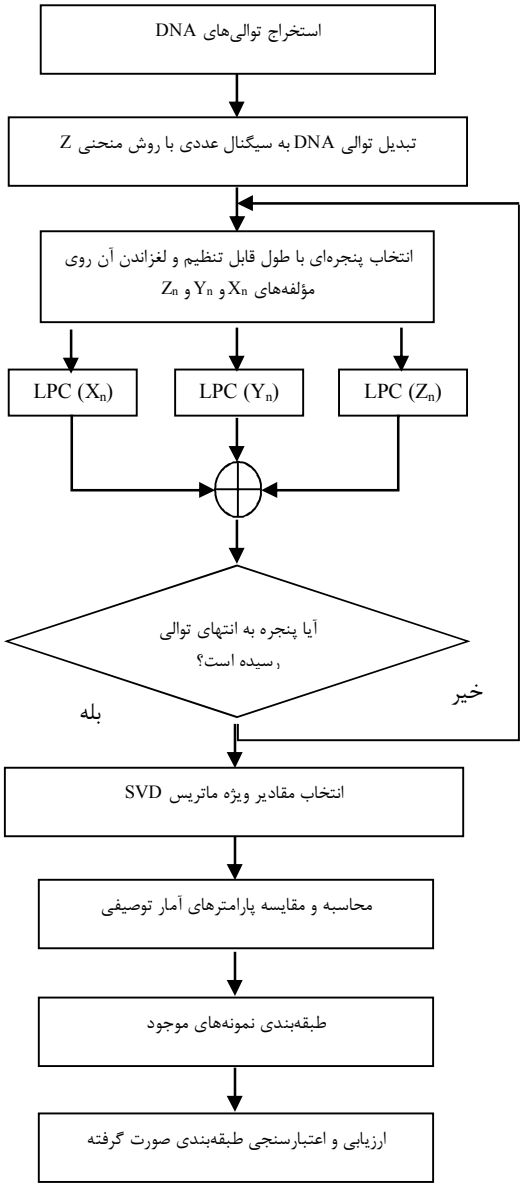
این نوع نگاشت عددی علاوه بر نمایش عددی رشته‌های مذکور، برخی ویژگی‌های بیولوژیکی آن‌ها را هم در قالب سه بردار X_n، Y_n و Z_n بیان می‌کنند. به طوری که X_n نشان‌گر حضور یکی از نوکلئوتیدهای A یا G، Y_n نشان‌گر حضور یکی از نوکلئوتیدهای A یا C و Z_n نشان‌گر حضور یکی از دو نوکلئوتید A یا T است. این ویژگی‌ها را می‌توان به روابط (۲) هم بیان کرد.

$$\begin{aligned} X_n &= (A_n + G_n) - (C_n + T_n) \\ Y_n &= (A_n + C_n) - (G_n + T_n) \\ Z_n &= (A_n + T_n) - (C_n + G_n) \end{aligned} \quad (۲)$$

٤. روش پیشنهادی

پس از تبدیل توالی‌های کاراکتری به دنباله‌های عددی، با استفاده از فیلترهای دیجیتال مخصوص و تبدیلات و الگوریتم‌های پردازش سیگنال، به تحلیل توالی‌های DNA و آشکارسازی اجزای درونی مخصوصی همچون خاصیت تناوب-۳ و مناطق کدشده پروتئینی می‌پردازند. با اجرای این تکنیک‌ها روی توالی‌های سرطانی و غیرسرطانی، ویژگی‌هایی متمایزی برای تفکیک سلول‌های سرطانی از غیرسرطانی استخراج می‌شود. در این بخش، برای آشنایی مختصر، برخی روش‌های مشهور و پر کاربرد تحلیل توالی‌ها، معرفی می‌شوند. پس از آن الگوریتم ارائه شده که مراحل کلی آن در شکل (۳) آمده است، در ادامه شرح داده خواهد شد. مشابه الگوریتم‌های مذکور، این روش هم مراحل یک سیستم شناسایی الگو را که

شامل بخش‌های توالی‌یابی DNA، استخراج و انتخاب ویژگی، طبقه‌بندی، تست و ارزیابی می‌باشد، را در بر می‌گیرد. لازم به ذکر است که شبیه‌سازی الگوریتم مذکور در نرم‌افزار متلب انجام گرفته است.



شکل (۳): فلوچارت کلی الگوریتم پیشنهادی

۱،٤. استخراج ویژگی

همانطور که اشاره شد روش‌های متعدد و متنوعی برای پیش‌بینی و تعیین نواحی کد شده پروتئینی ارائه شده است. تبدیل فوریه گسسته (DFT) و فیلتر Antinotch از جمله این روش‌ها هستند. تبدیل DFT یکی از ابزارهای حوزه پردازش سیگنال دیجیتال است که سیگنال‌های اولیه با طول محدود را به مجموعه‌ای از سیگنال‌های سینوسی با فرکانس‌های مختلف تبدیل می‌کند تا خواص سیگنال را در یک دوره زمانی مشخص بیان کند. از تبدیل فوریه گسسته برای تحلیل طیفی توالی‌های عددی DNA با طول محدود استفاده می‌شود که طبق تعریف برای توالی x[n] با طول محدود N، مقدار DFT بر اساس رابطه (۳) محاسبه می‌شود.

$$X[k]=\sum_{n=0}^{N-1}x[n]w[n-m]e^{-j\frac{2\Pi nk}{N}},0\le k\le N-1 \quad (۳)$$

در این رابطه $w[n-m]$ پنجره‌ای به طول n است که در برخی مقالات با لغزش آن به طول m در طول کل توالی، آن را بهبود بخشیده‌اند. با توجه به طول زیاد توالی‌های DNA و با رعایت دو پارامتر سرعت و دقت، انتخاب مقداری معقول برای طول پنجره اهمیت زیادی دارد. در ادامه از برخی معیارهای طیفی برای استخراج ویژگی استفاده می‌شود. یکی از این معیارها، مقدار طیف توان S[k] بر اساس رابطه (٤) است. مطابق این رابطه اگر رشته DNA ورودی متعلق به یک ناحیه کدشده پروتئینی باشد، S[k] در نقاط مضرب $\frac{N}{3}$ دارای مقدار بیشینه است [۱۳].

$$S[k] = |X[k]|^2 \quad (٤)$$

مقالات زیادی در این حوزه و مبتنی بر تبدیل فوریه ارائه شده است ولی یکی از مشکلات اساسی این روش انتخاب طول پنجره و محاسبات بالای تبدیلات فوریه به منظور ارائه نتایج دقیق‌تر است [۱٦-۱۹].

راهکار دیگری که برای یافتن مناطق کدشده پروتئینی استفاده می‌شود، طراحی فیلترهای خاصی برای اجرا روی توالی‌های DNA است. مشهورترین موردی که مقالات زیادی در این مورد ارائه شده است، فیلتر میان‌گذر Antinotch با رابطه (۵) است. خروجی این فیلتر برای نواحی ژنی بسیار بیشتر از نواحی غیرژنی می‌باشد و این موضوع راهکار مناسبی برای یافتن نواحی ژنی می‌باشد. پاسخ دامنه این فیلترها در

فرکانس $\omega_0 = \frac{2\pi}{3}$ مقدار بیشینه‌ای می‌دهد که به عنوان آشکارساز مؤلفه تناوب-۳ می‌تواند مورد استفاده قرار گیرد.

$$A(z) = \frac{R^2 - 2RCos\omega_0 z^{-1} + z^{-2}}{1 - 2RCos\omega_0 z^{-1} + R^2 z^{-2}} \qquad (۵)$$

در این رابطه ω_0 فرکانس مرکزی و R شعاع قطب است که با تغییر R پهنای باند فیلتر تعیین می‌شود. همچنین شرط R<1 همواره برقرار است و افزایش بیش از حد مقدار R تا مقدار ۱ منجر به تولید نویز بیشتر و کاهش دقت موقعیت مناطق کد شده پروتئینی می‌شود [۱۷].

در این تحقیق برای تحلیل توالی‌های DNA و استخراج ویژگی از این توالی‌ها از مدل پیشگوی خطی یا LPC استفاده شده است. مدل LPC یک الگوریتم فشرده‌سازی محسوب می‌شود که در زمینه تحلیل سیگنال صوت و گفتار کاربردهای زیادی دارد. یکی از ویژگی‌های حایز اهمیت این مدل، کمینه بودن مقدار خطای پیش‌بینی است. این مدل از دو بخش تشکیل می‌شود که یک بخش فقط شامل صفر و بخش دیگر تمام قطب است که در رابطه (٦) مشخص شده است.

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \qquad (٦)$$

ضرایب a_i در این رابطه را ضرایب پیشگوی خطی یا ضرایب LPC می‌نامند. عدد P نشان دهنده مقدار مرتبه پیشگویی الگوریتم LPC است. اساس این تبدیل فرض پیش‌گویی نمونه زمانی لحظه n سیگنال s یعنی s(n) با استفاده از ترکیب خطی P نمونه قبلی s است. این نمونه پیش‌گویی شده با $\hat{s}_{(n)}$ نشان داده می‌شود که رابطه اصلی این مدل به صورت رابطه (۷) خواهد بود:

$$\hat{s}_{(n)} = \sum_{i=1}^{p} a_i s(n-i) \qquad (۷)$$

در این رابطه $\hat{s}_{(n)}$ مقدار تخمینی خروجی است و هر یک از *s(n-i)*ها نشان دهنده مقادیر قبلی سیگنال هستند که از روی آن‌ها تخمین صورت خواهد گرفت، تا در نهایت مقادیر a_i به دست آید.

پس از اینکه رشته‌های کاراکتری به شکل عددی و در یک فضای سه بعدی نمایش داده شدند، از این مدل برای فشرده‌سازی و کاهش همبستگی نمونه‌های توالی‌های عددی DNA استفاده می‌شود. البته در سیگنال‌های ایستا، برای کسب نتایج بهتر و دقیق‌تر، از روش «پنجره‌بندی» استفاده می‌شود. بدین معنا که در این روش سیگنال به بازه‌های زمانی کوچکی تقسیم می‌شود و تحلیل سیگنال در این بازه‌های کوچک انجام می‌گیرد، که این بخش‌های کوچکتر را، پنجره می‌نامند. این الگوریتم در هر یک از پنجره‌های موجود اعمال می‌شود. از سوی دیگر برای در نظر گرفتن اطلاعات مرزهای این بخش‌ها، که ممکن است در تحلیل و پردازش تأثیرشان نادیده گرفته شود، پنجره‌های متوالی به شکل همپوشان انتخاب می‌شوند. در مجموع، این مراحل را با نام «پنجره‌بندی لغزان» شناخته می‌شود.

با این اوصاف، پس از دریافت تعداد زیادی رشته عددی، این رشته‌ها به چندین قسمت برابر تقسیم شده و عمل پنجره‌سازی انجام گرفته و با در نظر گرفتن مقداری هم‌پوشانی بین رشته‌ها و بر اساس یک مقدار صحیح مثبت که مرتبه p است، تخمینی از رشته‌های هر بخش در قالب بردار ماتریس ضرایب تخمین a_k ارائه می‌شود. البته تعیین مقدار مرتبه و طول پنجره قابل بحث است. این مقادیر روی هر سه بردار محاسبه شده و مجموع آن‌ها به صورت آنچه در رابطه (۸) مشاهده می‌شود، نمایش داده خواهد شد که مقدار ε نشان دهنده مقدار ناچیز خطای پیش‌بینی است.

$$y = Y * a + \varepsilon \qquad (۸)$$

۲,٤. انتخاب ویژگی

برای محاسبه ضرایب پیشگوی خطی، از روش‌های متعددی استفاده می‌شود که یکی از روش‌های معمول، استفاده از ماتریس‌های کواریانس است. ماتریس کواریانس، یکی از بهترین روش‌های مبتنی بر استخراج ویژگی کاهش ابعاد داده است که با حذف ضرایب کم اهمیت، ابعاد جدیدی برای داده تعریف می‌کند. ماتریس مربعی کواریانس، همبستگی دو به دو تمام ویژگی‌های داده‌ها را محاسبه می‌کند. پس از میان انبوه

ویژگی‌های به دست آمده، بر اساس بردارهای ویژه مذکور که یک فضای برداری m بُعدی را تشکیل می‌دهند، زیرفضای دومی را با به دست آوردن مقادیر ویژه ماتریس کواریانس می‌توان انتخاب کرد.

با این اوصاف، بعد از آن که الگوریتم LPC به تعداد طول پنجره m تعریف شده روی سیگنال‌های سه بعدی موجود اعمال شد، ماتریس X با ابعاد m در n تشکیل می‌شود. ماتریس کوواریانس متناظر آن به صورت رابطه (۹) و با استفاده از ترانهاده ماتریس محاسبه می‌شود که بر اساس بردارهای ویژه و مقادیر ویژه متناظر تجزیه می‌شود.

$$X = [x_1, x_2, \dots, x_m]^T,$$
$$C_{xx} = E\{XX^T\}, \qquad (۹)$$
$$D = diag[\lambda_1, \lambda_2, \dots, \lambda_m], V = [v_1, v_2, \dots, v_m]$$

در این معادله ماتریس D یک ماتریس قطری بوده و $\boldsymbol{\lambda}_i$ مقادیر ویژه ماتریس کوواریانس C_{xx} است و v_i بردار ویژه متناظر مقدار ویژه $\boldsymbol{\lambda}_i$ است. ستون‌های ماتریس V را بردارهای ویژه تشکیل می‌دهند.

تجزیه مقدار منفرد (SVD) یکی دیگر از روش‌های مفید در محاسبات ماتریسی است که به منظور تسهیل حل دستگاه‌های بزرگ خطی به کار می‌رود. در این روش برای هر ماتریس، یک ماتریس از چپ و یک ماتریس از راست، یک ماتریس قطری تشکیل می‌دهد. با فرض ماتریس X حاصل از اعمال الگوریتم LPC، تبدیل SVD آن با رابطه (۱۰) نمایش داده می‌شود.

$$X = U S V^T \qquad (۱۰)$$

در این رابطه ماتریس U به ابعاد M × M و V به ابعاد N× N هستند که به درایه‌های آن مقادیر منحصر به فرد تجزیه ماتریس A گویند. با توجه به مطالب بیان شده، تبدیل SVD روی مقادیر ویژه ماتریس کواریانس حاصل از مجموع سه بردار LPC اعمال می‌شود.

پس از کاهش ابعاد داده‌ای ماتریس‌های حاصل، می‌توان با محاسبه برخی پارامترهای آماری روی بردار به دست آمده، ویژگی‌هایی آماری برای تفکیک داده‌های سرطانی و غیرسرطانی استخراج کرد. بدین منظور میانگین یکی از شاخص‌های مهم و معمول در خلاصه‌سازی و متمرکز ساختن داده‌های حاصل در یک عدد است. علاوه بر شاخص‌های تمرکز داده، معیار تأثیرگذار دیگر در تحلیل یک مجموعه داده، پراکندگی داده‌هاست. شاخص پراکندگی، معیار سنجش میزان تغییرات داده‌هاست. پارامتری که به منظور بررسی پراکندگی بردار انتخاب شده است، انحراف‌معیار حاصل از میانگین انحراف‌های هر داده است. در روابط (۱۱) فرمول میانگین و انحراف معیار بیان شده است.

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n} \qquad (۱۱)$$

در این روابط میانگین و انحراف معیار به ترتیب با نماد μ و σ^2 مشخص شده‌اند که روی n داده X_i این مقادیر محاسبه می‌شوند.

۳,٤. طبقه‌بندی با SVM

ماشین‌های بردار پشتیبان یا SVM یک روش آماری غیرپارامتریک نظارتی طبقه‌بندی دودویی محسوب می‌شوند که دو کلاس را با استفاده از یک مرز خطی یا غیرخطی از هم جدا می‌سازند. یکی از مهمترین ویژگی‌های این الگوریتم توانایی بالای آن در رسیدن به یک طبقه‌بندی دقیق از داده‌ها با وجود نمونه آموزشی کمتر است. اگر داده‌های موجود به صورت مجموعه $((x_1,y_1), (x_2,y_2), \dots, (x_n,y_n))$ در نظر گرفته شوند، که x_i مجموعه داده آموزشی ورودی و y_i یکی از مقادیر ۱+ یا ۱- برای تفکیک دو دسته است، هدف یافتن خط یا صفحه‌ای است که نقاط با $y_i = 1$ را از نقاط $y_i = -1$ جدا سازد. در این صورت این صفحه را می‌توان با رابطه زیر تعریف کرد:

$$W.x + b = 0 \qquad (۱۲)$$

که منظور از W بردار وزن‌ها و n تعداد صفحه و b یک عدد برای تنظیم وزن در شرایط مختلف است و $\phi(x)$ نشان‌دهنده تابع انتقالی است که بردارهای ورودی را درون فضایی بزرگ نشان می‌دهد که با نام تابع هسته شناخته می‌شود.

به منظور طبقه‌بندی داده‌های غیرخطی، می‌توان آن‌ها را برای داده‌های غیرخطی تعمیم داد. الگوریتم SVM غیرخطی از دو مرحله اصلی تشکیل شده است. در مرحله اول داده‌های

ورودی به فضایی با ابعاد بالاتر نگاشت می‌شوند. پس از تبدیل داده‌ها، یک ابرصفحه جداکننده خطی در فضا اعمال می‌شود. از تابعی به نام تابع هسته بدین منظور استفاده می‌شود و روی داده‌های اولیه اجرا می‌شود. پس از آن مشابه حالت خطی، ابرصفحه‌ای با حداکثر حاشیه انتخاب می‌شود. توابع هسته متنوعی وجود دارند که هر یک ویژگی‌های خاصی دارند که بر اساس نوع داده و فضای قرارگیری داده‌ها یکی از آنها را می‌توان برگزید. یکی از توابع معروف هسته، تابع گاوسین است که به صورت زیر تعریف می‌شود:

$$k\left(x_i, x_j\right) = \frac{e^{-|x_i - x_j|^2}}{2\sigma^2} \qquad (۱۳)$$

۵. نتایج

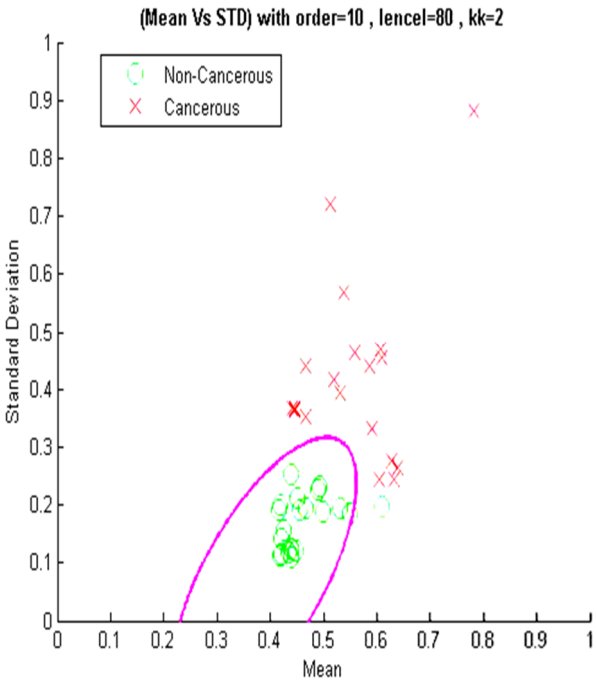
پس از اجرای مراحل مذکور، برای هر نمونه موجود، دو ویژگی میانگین و انحراف‌معیار حاصل می‌شود. در جدول (۲) نتایج ویژگی‌های پیاده‌سازی این روش روی داده‌های جدول (۱) و مقادیر آماری به دست آمده آورده شده است. البته در این جدول مقدار ۱۰ برای مرتبه مدل پیشگوی خطی و مقدار ۸۰ برای طول پنجره لغزان و مقدار هم پوشانی یک دوم طول پنجره در نظر گرفته شده است. بدیهی است که با تغییر متغیرهای مذکور، نتایج دیگری به همراه خواهد داشت.

جدول (۲): نتایج پیاده‌سازی الگوریتم پیشنهادی روی داده‌ها

نوع	شماره AC	میانگین	انحراف معیار
غیرسرطانی	NM_000518	۷/۲٤	٦/۱٥
	EU761578	۷/۱۱	٦/۳۷
	AY356351	۷/۱۲	٦/۲۹
	AF540397	۱۰/٦۱	۸/۹٦
	AY027509	۲۷/۹٥	۱٦/۷۸
سرطانی	DI170721	۲۸/۲۳	۱۷/۰۲
	DI170729	۲۸/۰۷	۱۷/۱۳
	NM_000059	۳۱/۷٥	۲۱/۱٥
	NM_006768	۳۱/۸٦	۲۱/٥۷
	NM_007294	۳۲/۰٤	۲۲/۰۱

اما با اعمال این شرایط، مقادیری که برای میانگین داده‌های سرطانی محاسبه شده است، مقداری بیشتر از ۲۷ را نشان می‌دهد، در حالی که برای نمونه‌های غیرسرطانی، مقدار میانگین کمتر از ۱۱ است و به همین ترتیب انحراف معیار داده‌های سرطانی هم بیشتر از ۱٦ بوده در حالی که داده‌های غیرسرطانی مقادیر کمتر از ۹ در این پارامتر آماری ارائه داده است. با اجرای این الگوریتم روی سایر داده‌ها هم، اعدادی مشابه به دست می‌آید که قابلیت محدود شدن و طبقه‌بندی دارند.

نتایج به دست آمده در جدول (۲) نشان داد که می‌توان مقادیری را به عنوان آستانه بالا یا پایین برای هر یک از دسته‌های سرطانی یا معمولی در نظر گرفت. همانطور که در جدول ۲ هم مشخص بود، می‌توان از یک خط یا منحنی برای جداسازی نمونه‌های سرطانی و غیرسرطانی بهره برد. در ادامه از ۲۰۰ نمونه سرطانی و غیرسرطانی مشابه دیگر استفاده شده است تا ارزیابی و اعتبارسنجی بهتری صورت گیرد. در شکل (٤) و در یک فضای ویژگی نرمال شده، با استفاده از یک طبقه‌بندی‌کننده غیرخطی نمونه‌های سرطانی و غیر سرطانی سابق به همراه نمونه‌های آزمایشی به کار گرفته شده به تصویر کشیده شده است. برای وضوح بیشتر، تنها ۲۰ نمونه سرطانی و ۲۰ نمونه غیرسرطانی از کل داده‌های موجود نمایش داده شده است. محور افقی بیان‌گر مؤلفه میانگین و محور عمودی بیان‌گر مؤلفه انحراف‌معیار است. در این شکل نمونه‌های سرطانی با نماد × و نمونه‌های غیرسرطانی با O مشخص شده‌اند. طبقه‌بندی‌کننده SVM هم به صورت یک منحنی این نمونه‌ها را از هم جدا ساخته است.



شکل (٤): تفکیک توالی‌های سرطانی از غیرسرطانی با SVM

یکی از مؤلفه‌های مهم در پیاده‌سازی مدل LPC، مرتبه آن است که تأثیر به سزایی در کارایی و زمان اجرای الگوریتم دارد. از سوی دیگر، بر حسب مقدار مرتبه انتخاب شده، مقداری مناسب برای طول پنجره لغزان الگوریتم پنجره‌بندی مدل LPC باید معین شود. از این رو در جدول (۳) تأثیر اعمال مقادیر مختلف مرتبه LPC و طول پنجره روی تعداد نمونه بیشتری نمایش داده شده است. باز هم در این جدول، علاوه بر ارزیابی هر یک از روش‌ها بر حسب شاخص‌های مبتنی بر درستی یا نادرستی پیش‌بینی‌های صورت گرفته، به اعتبارسنجی طبقه‌بندی صورت گرفته در هر مورد بر حسب شاخص‌های مذکور پرداخته شده است.

جدول (۳): نتایج ارزیابی و اعتبارسنجی الگوریتم پیشنهادی با مرتبه و طول پنجره متفاوت

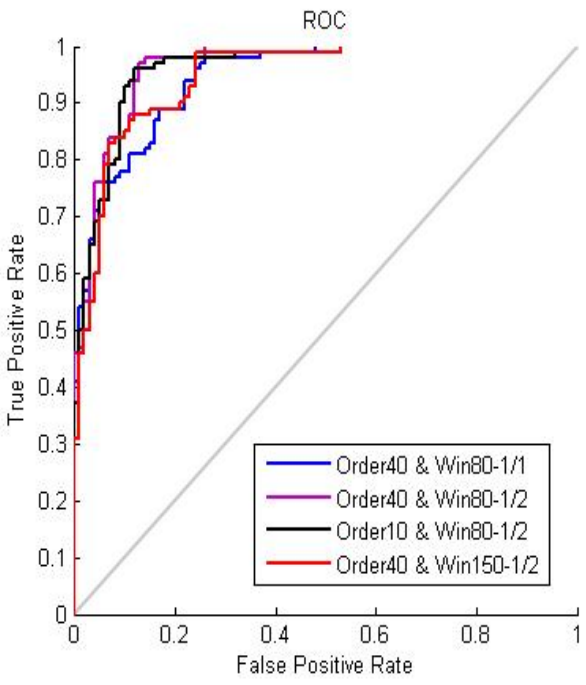
اجرا	مرتبه	پنجره	هم.پ	K-Fold	Resubs	AUC	CC
۱	٤۰	٦۰	۱	۰/۸۲۳	۰/۸۲۳	۰/۸۲٥	۰/٦٥۲
۲	٤۰	۸۰	۱	۰/۸۲٦	۰/۸۱٤	۰/۸۲٥	۰/٦٥۲
۳	٤۰	۱۲۰	۱	۰/۸٦۷	۰/۸۸	۰/۸۷	۰/۷٤۷
٤	٤۰	٦۰	۲	۰/۸٤۲	۰/۸٦۱	۰/۸٤٥	۰/٦۹۱
٥	٤۰	۸۰	۲	۰/۸۸۹	۰/۸۹۳	۰/۹	۰/۸۰۱
٦	٤۰	۱۲۰	۲	۰/۹۰۲	۰/۸۹۷	۰/۹	۰/۸۰۳
۷	٤۰	٦۰	۳	۰/۸۷۸	۰/۸٥۹	۰/۸۸	۰/۷٦۱
۸	٤۰	۸۰	۳	۰/۹۰٤	۰/۸۹۹	۰/۹	۰/۸۱
۹	٤۰	۱۲۰	۳	۰/۸۹٦	۰/۸۸۸	۰/۹۱	۰/۸۲٤
۱۰	۱۰	٤۰	۲	۰/۸٤۸	۰/۸٤۳	۰/۸٥٥	۰/۷۱۰
۱۱	۱۰	٦۰	۲	۰/۹۱	۰/۹۰۳	۰/۹۱٥	۰/۸۳۱
۱۲	۱۰	۸۰	۲	۰/۹	۰/۹۰۷	۰/۹۰٥	۰/۸۱

در این جدول ستون اجرا، شماره اجرای مورد نظر، مرتبه، مرتبه مدل LPC، پنجره نشان‌گر طول پنجره انتخابی برای لغزاندن و ستون هم.پ هم مقدار هم‌پوشانی دو پنجره متوالی را نشان می‌دهد. منظور از اعداد ۱، ۲ و ۳ در ستون هم.پ مقادیر یک یکم، یک دوم و یک سوم طول پنجره به عنوان مقدار هم‌پوشانی است. همچنین ستون K-Fold و Resubs بیان‌گر دو معیار اعتبارسنجی K-دسته و جایگزینی مجدد پس از طبقه‌بندی است. مقادیر AUC و CC هم به ترتیب مقادیر سطح زیر نمودار ROC و همبستگی حاصل از پارامترهای مبتنی بر دقت پس از طبقه‌بندی را مطابق رابطه (١٤) نشان می‌دهد. منظور از پارامترهای دقت مقادیر TP، TN، FP و FN است.

$$CC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (١٤)$$

نتایج حاصل، نشان می‌دهند که میانگین و انحراف‌معیار الگوریتم پیشنهادی روی داده‌های سرطانی و غیرسرطانی

موجود، می‌تواند معیار مناسبی برای تفکیک و طبقه‌بندی باشد. افزایش مقدار هم‌پوشانی، تأثیر مستقیمی در افزایش کارایی دارد. با فرض ثابت بودن مرتبه و هم‌پوشانی پنجره‌ها، مشخص است که در ردیف ۱ تا ۳، ٤ تا ٦ و ۷ تا ۹ مقادیر CC و AUC روند رو به رشدی دارند. همچنین مقادیر اعتبارسنجی این ردیف‌ها هم تا حدودی این موضوع را تایید می‌کنند. البته با افزایش مقدار هم‌پوشانی، دقت در موارد با پنجره به طول ۸۰ بیشتر از ۱۲۰ مشاهده می‌شود که به دلیل افزایش دقت آن‌ها با افزایش دامنه نمونه‌برداری است. همچنین با فرض ثابت بودن مرتبه و طول پنجره، افزایش مقدار هم‌پوشانی در سه اجرای ۱-٤-۷ یا ۲-۵-۸ یا ۳-٦-۹ افزایش مقدار CC و AUC را به وضوح نشان می‌دهد. معیارهای اعتبارسنجی طبقه‌بندی هم روند رو به رشدی نشان می‌دهند.



شکل (۵): نمودار ROC الگوریتم LPC با مقادیر مرتبه و طول پنجره مختلف

در شکل (۵) هم منحنی ROC در چهار حالت با مرتبه، طول پنجره و مقدار هم‌پوشانی مختلف رسم شده است و نشان می‌دهد که در هر FP، مقدار TP چقدر است. همان‌طور که انتظار می‌رود بر اساس این نمودار می‌توان گفت که دقت اجرای با مرتبه ۱۰، پنجره ۸۰ و همپوشانی ۲، بیشتر از موارد با مرتبه، پنجره و هم‌پوشانی <٤۰-۸۰-۲>، <٤۰-۸۰-۱۵۰> و <٤۰-۸۰-۱> است، چرا که در اکثر نقاط به نقطه (۰,۱) نزدیک‌تر است.

برای بررسی روش پیشنهادی، این راهکار روی نمونه‌های به کار رفته در مقاله [۹] هم آزمایش شد. در جدول (٤) ، نتایج اعمال روش پیشنهادی روی نمونه‌های استفاده شده در مقاله [۹] به تصویر کشیده شده است. بدین منظور الگوریتم پیشنهادی با مرتبه ۱۰ و طول پنجره ٦۰ روی نمونه‌های مقاله مذکور اعمال شده است.

جدول (٤): نتایج ویژگی‌های حاصل روی نمونه‌های مقاله [۹]

نمونه	Mean	STD	Mean/STD
AF083883	۱/۶۸۸	۲/۰۷۷	۰/۸۱۲
AF186607	۱/۶۵۷	۱/۸۸۲	۰/۸۸۰
AF186613	۱/۶۳۹	۱/۹۱۶	۰/۸۵۵
AF007546	۱/۵۷۳	۲/۲۸۲	۰/۷۹۳
AF008216	۱/۹٤۱	۳/۳۴۳	۰/۵۸۰
AF348525	۱/۸٤٤	۱/۶۳۱	۰/۹۵٤
NM_016346	۱/۹۷۳	۲/۷۵۸	۰/۷۱۵
NM_005732	۱/۹۷٤	۳/۱۲۴	۰/٦۳۱
AF348515	۱/۸٤۸	۱/۷۳٦	۰/۸٦۹
NM_012403	۱/۹۱۳	۱/۴۸۵	۱/۲۸۸

در این جدول چهار نمونه اول از بالا غیرسرطانی بوده و ٦ نمونه بعدی سرطانی هستند که با نام دسترسی منحصر به فرد به کار رفته خود در بانک اطلاعاتی genbank سایت NCBI مشخص شده‌اند. در ستون‌های بعدی نتایج حاصل برای ویژگی‌های میانگین و انحراف‌معیار سیگنال‌های خروجی، به ترتیب از راست به چپ با لغات Mean و STD به ازای هر نمونه، مشخص شده‌اند. همچنین در ستون بعدی نسبت این دو ویژگی آورده شده است که به خوبی نشان می‌دهد که در نمونه‌های غیرسرطانی انتخابی که از یک نوع ژن هستند، مقادیر تقریبا مشابهی برای هر دو ویژگی به دست آمده است.

همچنین این الگوریتم روی نمونه‌های محدود آزمایشگاهی با طول بسیار بیشتری هم آزمایش شده و یافته مهم دیگری که

در بر داشته است، توانایی شناسایی ژن‌های مختلف یک کروموزوم خاص با محدود ساختن هر یک از ویژگی‌های مذکور در بازه‌هایی خاص است. نکته حائز اهمیت در این بین، مقادیر به نسبت قابل توجه ویژگی‌های میانگین و انحراف‌معیار در نمونه‌های سرطانی به نسبت موارد غیرسرطانی است که به دلیل تغییر توزیع نوکلئوتیدی ناشی از ماهیت جهش ژنتیکی شکل‌گیری سرطان است. البته با توجه به این که درصد تشخیص و صحت ۱۰۰٪ آن مشخص است و تعداد به نسبت کمی داده استفاده شده است، قابلیت مقایسه مقادیر مختلف پارامترهای متغیر الگوریتم، مشابه معیارهای ارزیابی جدول (۳) یا شکل (۵) میسر نیست.

٦. نتیجه

الگوریتم پیشنهادی مبتنی بر مدل LPC که روش‌های تجزیه مقدار منفرد و ماتریس کواریانس را برای انتخاب ویژگی استفاده می‌کند، نشان داد که با فشرده‌سازی مبتنی بر پیش‌گویی سیگنال‌ها می‌تواند نمونه‌های سرطانی را از نمونه‌های غیرسرطانی جدا سازد. این الگوریتم به دلیل دقت بالای آن در پیش‌بینی نواحی کد شده پروتئینی در مرتبه‌های پایین‌تر، دقت بیشتری هم در طبقه‌بندی دو دسته موجود به همراه داشت. از جمله مزایای دیگر این روش، ماهیت پنجره‌بندی شده آن است؛ چرا که داده‌های این حوزه به صورتی هستند که توانایی پردازش آن در یک اجرا میسر نیست. از سوی دیگر، پنجره‌بندی سبب می‌شود تا تمام سیگنال‌های نمونه‌ها پردازش شوند. در نظر گرفتن مقادیر هم‌پوشانی بیشتر پنجره‌ها، علاوه‌بر بررسی مرز پنجره‌ها، به کسب اطمینان از بررسی همبستگی‌های سیگنال با سیگنال‌های مختلف و اجرای صحیح الگوریتم کمک می‌کند.

همانطور که مشاهده شد الگوریتم پیشگوی خطی که کاربردهای وسیعی در زمینه فشرده‌سازی و تخمین فرکانس در حوزه پردازش صوت دارد، به خوبی در تحلیل توالی‌های DNA هم توانست تفاوت‌های ساختاری مولکول‌های DNA سازنده سرطانی و غیرسرطانی را به خوبی نشان دهد. با درنظر گرفتن این مطلب که الگوریتم پیشگوی خطی پنجره‌بندی شده است و بر حسب مقداری که برای تعداد پنجره و طول آن انتخاب می‌شود، یکی از موارد مورد بحث در این مقاله انتخاب مقداری مناسب برای دو پارامتر مرتبه اجرایی و طول پنجره می‌باشد که به نمونه‌های انتخابی و تعداد و طول آن‌ها بستگی دارد.

نتایج حاصل، نشان می‌دهند که میانگین و انحراف‌معیار الگوریتم پیشنهادی روی داده‌های سرطانی و غیرسرطانی موجود، می‌تواند معیار مناسبی برای تفکیک و طبقه‌بندی باشد. در واقع در این پژوهش، روشی برای مشخص ساختن انواع ژن روی ژنوم انسان ارائه شد که با تعمیم و توسعه آن روی انواع دیگر ژن‌ها، می‌توان یک دنباله با طول بسیار زیادی از نوکلئوتیدها را در یک یا دو عدد خلاصه کرد و به نوعی عمل فشرده‌سازی انجام داد. نکته حائز اهمیت این عمل، مستقل بودن تحلیل این توالی‌های عظیم نسبت به طول آن‌ها و مهم‌تر از آن قابلیت تفکیک انواع ژن (سرطانی-غیرسرطانی) بر حسب تغییرات ژنتیکی حاصل است که به صورت حدود آستانه‌ای خاص و مشخص می‌توان برای هر نوع ژن و حتی هر ژن خاص تعیین کرد.

منابع

[1] Howlader, N., et al, *SEER Cancer Statistics Review 1975-2008*, National Cancer Institute. Bethesda, 2011, Available on: http://seer.cancer.gov/archive/csr/1975_2008/.

[2] American Cancer Society. *Cancer Facts & Figures*, American Cancer Society Inc., Atlanta, USA, 2009, Available on: http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2009/index.

[3] Brown, M. S., Goldin, J. G., Suh, R. H., Mcnitt-Gray, M. F., Sayre, J. W., Aberle, D. R., *"Lung Micronodules: Automated Method for Detection at Then-Section CT-Initial Experience"*, Radiology, Vol. 226, No. 1, pp. 256-262, 2003.

[4] Akhtar, M., Epps, J., Ambikairajah, E., *"On DNA Numerical Representations for Period-3 based Exon Prediction"*, The 2th IEEE International Workshop on Genomic Signal Processing and Statistics, pp. 1-4, Tuusula, Finland, 2007.

[5] Abo-Zahhad, M., Ahmed, S. M., Abd-Elrahman, S. A., *"Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Technique"*, International Journal of Information Technology & Computer Science, Vol. 4, No. 8, pp. 22-36, 2012.

[6] Khare, A., Nigam, A., Saxena, M., *"Identification of DNA Sequences by Signal Processing Tools in Protein Coding Regions"*, Vol. II, No. 2, pp. 44-49, 2011.

[7] Saberkari, H., Shamsi, M., Heravi, H., Sedaaghi, M.H., *"A Novel Fast Algorithm for Exon Prediction in Eukaryotic Genes Using Linear Predictive Coding Model and Goertzel Algorithm based on the Z-Curve"*, International Journal of Computer Applications, Vol. 67, No. 17, pp. 25- 38, 2013.

[8] Akhtar, M., Ambikairajah, E., Epps, J., *"Detection of Period-3 Behavior in Genomic Sequences Using Singular Value Decomposition"*, The 1th IEEE International Conference on Emerging Technologies, pp. 13-17, Islamabad, Pakistan, 2005.

[9] Satapathi, G.N., Srihari, P., Jyothi, Ch.A., Lavanya, S., *"Prediction of Cancer Cell Using DSP Techniques"*, The 3th IEEE International Conference on Communication and Signal Processing, pp.149-153, Melmaruvathur, India, 2013.

[10] Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*, 1th ed., New York, Garland Publishing Inc., 1997.

[11] Watson, J. D., Crick, F. C. H., *"Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid"*, Nature, Vol. 171, No. 4356, pp. 737-738, 1953.

[12] Reece, J. B., Taylor, M. R., Simon, E. J., Dickey, J. L., *Campbell Biology: Concepts and Connections*, 6th ed., Benjamin Cummings Inc., 2009.

[13] Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R., *"Prediction of Probable Genes by Fourier Analysis of Genomic Sequence"*, Bioinformatics, Vol. 13, No. 3, pp. 263-270, 2007.

[14] National Center for Biotechnology Information, *Genbank Nucleotide Database*, Available on: http://www.ncbi.nlm.nih.gov.

[15] Zhang, Ch. T., Zhang, R., Ou H. Y., *"The Z curve Database: A Graphic Representation of Genome Sequences"*, Bioinformatics, Vol. 19, No. 5, pp. 593-599, 2003.

[16] Oppenheim, A. V., Schafer, R. W., *Discrete-Time Signal Processing*, Prentice Hall Inc., 1999.

[17] Vaidyanathan, P. P., Yoon, B. J., *"The Role of Signal-Processing Concepts in Genomics and Proteomics"*, Journal of Franklin Institute, Special Issues on Genomics, Vol. 341, No. 1, pp. 111-135, 2004.

[18] Anastassiou, D., *"Frequency-Domain Analysis of Biomolecular Sequences"*, Bioinformatics, Vol. 16, No. 12, pp. 1073-1081, 2002.

[19] Datta, S., Asif, A., *"A Fast DFT based Gene Prediction Algorithm for Identification of Protein Coding Regions"*, The 30th EEE International Conference on Acoustics, Speech and Signal Processing, pp. 653-656, Philadelphia, USA, 2005.