وزارت علوم تحقیقات و فناوری

گروه مهندسی کامپیوتر

پایان نامه برای دریافت درجه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی

عنوان:

تشخیص نواحی کد کننده پروتئینی سرطانی روی توالی‌های DNA

استاد راهنما:

دکتر بهزاد مظفری تازه‌کند

استاد مشاور:

دکتر میر هادی سیدعربی

پژوهش‌گر:

امین خدائی

۱۳۹۳

چکیده:

با وجود تمام پژوهش‌ها و پیشرفت‌هایی که در حوزه‌های مختلف علمی در راستای تشخیص و درمان بیماری‌ها صورت گرفته است، میزان ابتلا و مرگ و میر برخی بیماری‌ها، به ویژه سرطان‌ها، همچنان در حال افزایش است. ساختار شکل‌گیری و روند انتشار سرطان، محققان را به مطالعه ابعاد مختلف آن ترغیب می‌کند. منشا ژنتیکی سرطان، لزوم بررسی اجزای درون سلولی آن را بیان می‌کند. با در نظر گرفتن این نکته که تشخیص هر چه سریع‌تر سرطان، در بهبود و درمان آن تاثیر به سزایی دارد، کاوش توالی‌های DNA به عنوان منشا اصلی ژنتیکی سلول، می‌تواند موفقیت بزرگی در درمان این بیماری پنهان باشد. از این رو برای تشخیص ویژگی‌های مبتنی بر ژن توالی‌های DNA که تحت عنوان نواحی کدکننده پروتئینی شناخته می‌شوند، از روش‌های متعددی که مبتنی بر روش‌های پردازش سیگنال است، می‌توان استفاده کرد. در سال‌های اخیر و با ظهور و توسعه علم انفورماتیک زیستی به عنوان دانشی بین‌رشته‌ای، این تحقیقات گسترش یافته‌اند. انفورماتیک زیستی، روش‌ها و الگوریتم‌های علوم محاسباتی و کامپیوتری را در کنار مطالعات ژنتیکی و زیستی به کار می‌گیرد.

در این پژوهش سعی شده است تا با بهره‌گیری از روش‌های تحلیل توالی‌های DNA و با در نظر گرفتن ویژگی ژنتیکی منحصر به فرد تناوب-۳ نواحی کد شده پروتئینی توالی‌های DNA، موارد سرطانی و غیرسرطانی به طور مجزا تحلیل و تفکیک شوند. در مرحله پیش‌پردازش، از روش نگاشت منحنی Z به منظور تبدیل رشته‌ای به عددی توالی‌های DNA بهره گرفته شده است. در روش‌های پیشنهادی ارائه شده که به صورت یک سیستم شناسایی آماری الگو هستند، از روش‌های پردازش سیگنال مدل پیشگوی خطی (LPC)، تبدیل فوریه گسسته (DFT) و فیلترهای دیجیتال به منظور استخراج ویژگی استفاده شده است. همچنین در مرحله استخراج و انتخاب ویژگی‌های مفید و کمینه، از روش‌های محاسباتی و آماری استفاده شده است. در نهایت مقایسه برخی پارامترهای آماری، با طبقه بندی کننده ماشین بردار پشتیبان (SVM)، تمایز خوبی بین نمونه‌های سرطانی و غیرسرطانی نشان داده است. نتایج حاصل و آزمون‌های ارزیابی و اعتبارسنجی نتایج به دست آمده، صحت و دقت بالای روش‌های ارایه شده را نشان می‌دهد که به نوعی ماهیت جهش ژنتیکی بیماری سرطان را هم تصدیق می‌کند.

واژه‌های کلیدی:

بیماری سرطان، توالی DNA، ناحیه کد شده پروتئینی، روش‌های پردازش سیگنال، شناسایی آماری الگو

فهرست مطالب

عنوان **صفحه**

فصل اول: مقدمه ۱
۱-۱- مروری بر کارهای انجام شده ۲
۱-۲- ارتباط موضوع با کارهای پیشین ۶
۱-۳- هدف تحقیق ۶
۱-۴- ساختار پایان‌نامه ۷

فصل دوم: مفاهیم پایه ۹
۲-۱- سلول ۱۰
۲-۲- ساختار DNA ۱۱
۲-۲-۱ ساختار نوکلئوتید ۱۳
۲-۲-۲ قواعد حاکم بر نوکلئوتیدها ۱۴
۲-۳- بیان ژن و ویژگی‌های DNA ۱۵
۲-۳-۱ همانندسازی ۱۶
۲-۳-۲ رونویسی ۱۷
۲-۳-۳ ترجمه ۱۸
۲-۴- کد ژنتیکی و ویژگی تناوب-۳ ..... ۱۸
۲-۵- سرطان ۲۰
۲-۶- انفورماتیک زیستی ۲۲
۲-۶-۱ مروری بر تاریخچه انفورماتیک زیستی ۲۳
۲-۶-۲ شاخه‌های انفورماتیک زیستی ۲۴
۲-۶-۳ فناوری‌های DNA ۲۵
۲-۷- خلاصه فصل ۲۷

فصل سوم: روش بررسی ۲۸
۳-۱- نمونه‌برداری و پایگاه داده‌های زیستی ۲۹
۳-۲- تبدیل رشته‌های DNA به سیگنال عددی ۳۰
۳-۲-۱ روش Voss ۳۲
۳-۲-۲ روش چهارسطحی ۳۲
۳-۲-۳ روش منحنی Z ۳۳
۳-۲-۴ روش اعداد مختلط ۳۳
۳-۲-۵ روش چهارگانه ۳۴
۳-۲-۶ روش پتانسیل فعل و انفعال الکترون-یون (EIIP) ۳۴
۳-۲-۷ فاصله بین نوکلئوتیدی ۳۵
۳-۲-۸ فرکانس پیشامد نوکلئوتیدی ۳۵

عنوان	صفحه

۳-۲-۹ سایر روش‌ها ۳۵

۳-۳- روش‌های استخراج ویژگی با تحلیل توالی‌های DNA ۳۶
۳-۳-۱ تبدیل فوریه و روش‌های مبتنی بر تبدیل فوریه گسسته (DFT) ۳۷
۳-۳-۲ فیلترهای دیجیتال ۳۹
۳-۳-۳ الگوریتم پیشگوی خطی (LPC) ۴۲

۳-۴- طبقه‌بندی ۴۴
۳-۴-۱ روش طبقه‌بندی ماشین بردار پشتیبان ۴۵
۳-۴-۲ روش‌های طبقه‌بندی مشهور دیگر ۴۷

۳-۵- خلاصه فصل ۴۷

فصل چهارم: روش پیشنهادی ۴۸
۴-۱- نمونه‌برداری ۴۹
۴-۲- پیش‌پردازش ۵۰
۴-۳- الگوریتم پیشنهادی مبتنی بر پیشگوی خطی و تجزیه مقدار منفرد ۵۴
۴-۳-۱ استخراج ویژگی با الگوریتم LPC پنجره‌بندی شده ۵۵
۴-۳-۲ انتخاب ویژگی با ماتریس کواریانس و SVD ۵۶
۴-۳-۳ طبقه‌بندی داده‌ها با SVM غیرخطی ۵۸

۴-۴- الگوریتم پیشنهادی مبتنی بر تبدیل فوریه ۵۹
۴-۴-۱ استخراج ویژگی با الگوریتم DFT پنجره‌بندی شده لغزان ۵۹
۴-۴-۲ انتخاب ویژگی با ماتریس کواریانس ۶۰
۴-۴-۳ طبقه‌بندی با SVM غیرخطی ۶۳

۴-۵- الگوریتم پیشنهادی مبتنی بر فیلتر Anti-notch ۶۳
۴-۵-۱ استخراج ویژگی با فیلتر میان‌گذر Anti-notch ۶۳
۴-۵-۲ انتخاب ویژگی با تشکیل ماتریس کواریانس ۶۴
۴-۵-۳ طبقه‌بندی ۶۶

۴-۶- خلاصه فصل ۶۷

فصل پنجم: نتایج پیاده‌سازی ۶۸
۵-۱- معیارهای ارزیابی ۶۹
۵-۱-۱ حساسیت، ویژه بودن و منحنی مشخصه عملکرد (ROC) ۷۰
۵-۱-۲ معیارهای اعتبارسنجی طبقه‌بندی‌کننده‌ها ۷۳

۵-۲- نتایج پیاده‌سازی الگوریتم پیشنهادی مبتنی بر مدل LPC ۷۵
۵-۲-۱ ارزیابی روش‌های مختلف نگاشت عددی DNA ۷۵
۵-۲-۲ ارزیابی مقدار مرتبه مدل LPC و طول پنجره انتخابی ۷۶
۵-۲-۳ ارزیابی مقدار هم‌پوشانی پنجره‌های مجاور ۷۷

عنوان	صفحه

۵-۲-۴ نتایج طبقه‌بندی در قالب نمودارهای میانگین-انحراف‌معیار ۷۸
۵-۲-۵ ارزیابی زمان اجرای الگوریتم ۸۰

۵-۳- نتایج پیاده‌سازی الگوریتم پیشنهادی مبتنی بر تبدیل DFT ۸۱
۵-۳-۱ ارزیابی اندازه پنجره انتخابی برای پنجره‌بندی DFT ۸۱
۵-۳-۲ ارزیابی مقدار هم‌پوشانی پنجره‌های مجاور ۸۲
۵-۳-۳ نتایج طبقه‌بندی در قالب نمودارهای نقطه بیشینه-انحراف‌معیار ۸۲
۵-۳-۴ ارزیابی زمان اجرای الگوریتم ۸۵

۵-۴- نتایج الگوریتم مبتنی بر فیلتر Anti-notch ۸۶
۵-۵- مقایسه روش‌های پیشنهادی ۸۸
۵-۵-۱ مقایسه مبتنی بر معیار فاصله اقلیدسی ۸۹
۵-۵-۲ مقایسه مبتنی بر ضریب تشابه کسینوسی ۹۰
۵-۵-۳ مقایسه مبتنی بر نمودار و ضریب نیم‌رخ ۹۱
۵-۵-۴ نمودارهای مقایسه و ارزیابی ۹۲

۵-۶- خلاصه فصل ۹۳

فصل ششم: بحث و نتیجه‌گیری ۹۴
۶-۱- پیشنهادها و راهکارهای آتی ۹۷

پیوست الف – علایم اختصاری ۹۹
پیوست ب – واژه‌نامه انگلیسی به فارسی ۱۰۱
پیوست ج – بانک اطلاعات زیستی NCBI ۱۰۴

منابع ۱۰۷

فصل ششم

بحث و نتیجه‌گیری

جداول و نمودارهای فصل پیش نشان دادند که الگوریتم‌های ارایه شده که بـه منظـور کـاوش نـواحی کدشده پروتئینی داده‌ها استفاده شدند، می‌توانند الگوها و ویژگـی‌هـای دو دسـته متمـایز سـرطانی و غیرسرطانی را از هم تفکیک سازند. البته همان‌طور که در بخش استخراج ویژگی این الگوریتم‌ها بحث شد، اندازه بسیار بزرگ توالی‌های DNA، روش‌هایی به منظور کاهش ابعاد را می‌طلبد که در هـر یـک از روش‌های ارایه شده، از روش‌های محاسباتی و آماری برای انتخاب ویژگی به منظور تمایز نمونه‌های سرطانی از غیرسرطانی استفاده شد.

الگوریتم پیشنهادی مبتنی بر مدل LPC که از روش‌های تجزیه مقدار منفرد و ماتریس کواریـانس برای انتخاب ویژگی سود می‌برد، نشان داد که با فشرده‌سازی مبتنی بر پیش‌گویی سیگنال‌ها می‌تواند نمونه‌های سرطانی را از نمونه‌های غیرسرطانی جدا سازد. این الگـوریتم بـه دلیـل دقـت بـالای آن در پیش‌بینی نواحی کد شده پروتئینی در مرتبه‌های پایین‌تر، دقت بیشتری هم در طبقه‌بندی دو دسـته موجود به همراه داشت. از جمله مزایای دیگر این روش، ماهیت پنجره‌بندی شـده آن اسـت؛ چـرا کـه داده‌های این حوزه به صورتی هستند که توانایی پردازش آن در یک اجرا میسر نیست. از سوی دیگـر، استفاده از تکنیک پنجره‌بندی سبب می‌شود تا تمام سیگنال‌ها پردازش شوند. در نظر گـرفتن مقـادیر هم‌پوشانی بیشتر پنجره‌ها، علاوه بر بررسی مرز پنجره‌ها به کسب اطمینان از بررسی همبسـتگی‌هـای سیگنال با سیگنال‌های مختلف و اجرای صحیح الگوریتم کمک می‌کند.

الگوریتم پیشنهادی مبتنی بر تبدیل DFT پنجره‌بندی شده، با وجود صرف زمان بیشـتر، از لحـاظ دقت و ساختار فضای ویژگی، نتایج بهتری ارایه داد. تبدیل DFT که پیش از این هم کاربرد زیادی در تحلیل توالی‌های DNA داشت، نشان داد که می‌تواند ویژگی‌های متمایزی روی توالی‌های سـرطانی و غیرسرطانی، بر اساس ویژگی تناوب-۳ نواحی کد شده پروتئینی آشکار سازد. استفاده از این الگـوریتم که همراه با استفاده از ماتریس کواریانس برای آشکارسازی نقطه بیشینه مربوط بـه ویژگـی تنـاوب-۳ همراه بود، به خوبی جهش توالی‌های سرطانی را با پاسخ فرکانسی نشان داد. این ویژگی که با انتخـاب مناسب مقدار طول پنجره بهتر هم نمود پیدا کرد، مولفه بسیار خوبی برای هر نوع طبقـه‌بنـدی‌کننـده برای تفکیک موارد سرطانی از غیرسرطانی است.

الگوریتم پیشنهادی مبتنی بر فیلتر میان‌گذر Anti-notch به دلیل ماهیت میان‌گذر فیلتـر آن، بـه خوبی قابلیت آشکارسازی یک نقطه خاص (ویژگی تناوب-۳ توالی‌های DNA) را داراست. در واقع این روش بر مبنای ماهیت فیلتر Anti-notch طراحی شده است و به دلیل این کـه بـدون هـیچ پـردازش اضافی ویژگی مناسبی استخراج می‌کند، می‌تواند کارکرد بیشتری هم داشته باشد. این فیلتـر، اگرچـه این ویژگی را در هر دو نوع سرطانی و غیرسرطانی ظاهر می‌سازد، ولی همان طـور کـه مشـاهده شـد خروجی آن در موارد سرطانی به حدی زیـاد اسـت کـه پـس از نرمـال‌سـازی نتـایج، مقـادیر بیشـینه نمونه‌های غیرسرطانی نزدیک به صفر نشان داده می‌شوند. مزیت اصلی این روش، علاوه بـر سـادگی و کم خرج بودن آن، صرف زمان کم در عین حصول نتیجه با دقت بالاست.

روش‌های پیشنهادی ارایه شده در این پژوهش، به خوبی ویژگی جهش ژنی به وجود آمده در توالی‌های DNA را به تصویر کشیدند. این الگوریتم‌ها با وجود این که به طور مستقیم به این موضوع اشاره ندارند، ولی اگر به تفاوت مقادیر میانگین و انحراف‌معیار به وجود آمده مدل LPC، یا مقادیر بیشینه تیز به وجود آمده در روش مبتنی بر تبدیل DFT پنجره‌بندی شده و یا اختلاف فاحش مقادیر مقادیر بیشینه حاصل از اعمال فیلتر Anti-notch در نمونه‌های سرطانی و غیرسرطانی دقت شود، مشخص است که این تغییرات مفهوم زیستی جهش ژنی را به خوبی نشان می‌دهند.

شایان ذکر است که در این پژوهش، تلاش شد تا حد امکان هیچ پیش‌داوری در مورد نتایج مورد انتظار صورت نگیرد و متغیرهای موجود در هر الگوریتم مورد سنجش قرار گیرند تا معیار ارزش‌یابی الگوریتم‌ها، فقط نتایج، جداول و نمودارهای ارزیابی و اعتبارسنجی باشند.

همچنین استفاده از دو مجموعه داده مختلف در این پژوهش، در ابتدا برای اطمینان از درستی نتایج و تصادفی نبودن آن بود، ولی در طول پژوهش برخی نتایج دیگر را هم در بر داشت. با این که این داده‌ها از لحاظ ماهیت ژنتیکی با هم تفاوت داشتند و هر یک به ژن‌هایی از کروموزوم‌های متفاوت از اندام‌های خاص بدن انسان اشاره داشتند، اما مساله اصلی ماهیت سرطانی یا غیرسرطانی بودن آنها بود. جداول ۴-۱ و ۴-۲ که نمونه‌های مربوط به سرطان سینه بود، به دلیل این که از نواحی یکسانی از کروموزوم‌های سلول استخراج شده بودند، قابلیت مقایسه و بررسی را ایجاد نمودند. نمونه‌های جدول ۴-۳ هم که از ژن‌های کروموزوم‌های سلولی مرتبط با خون بودند، این قابلیت را مهیا ساختند. اما نتیجه کلی استفاده از دو مجموعه مجزای سرطانی و غیرسرطانی، مقایسه توالی‌های DNA در شرایط یکسان روی ژن‌های مشابه با ویژگی‌های زیستی یکسان است. بدین معنا که نمی‌توان به عنوان مثال یک نمونه توالی یک ژن غیرسرطانی مرتبط با خون را با یک نمونه سرطانی مرتبط با سینه، ریه یا هر عضو دیگری را مقایسه کرد، چرا که از نواحی کروموزومی یکسانی توالی‌یابی نشده است.

در کنار این نتایج مثبت به دست آمده، نباید از برخی مشکلات و محدودیت‌های ناشی از علوم زیستی و ژنتیک، چشم‌پوشی کرد تا از اعتماد بیش از حد و داشتن انتظارات بیهوده جلوگیری کرد. پیش‌گویی‌های انفورماتیک زیستی در هر زمینه و مفهومی، به طور حتم و دقیق اثباتی برای آن مفهوم نیستند، چرا که کیفیت این پیش‌بینی‌ها به کیفیت داده‌ها و قابلیت‌های روش‌های استفاده شده بستگی دارند. داده‌های توالی موجود در بانک‌های اطلاعاتی یا هر جای دیگری، ممکن است به دلایل مختلف و به دلیل مراحل زمان‌بر و با خطاپذیری بالای آن، همراه با خطا یا نویز باشند. به تعبیری می‌توان استفاده از علم انفورماتیک زیستی در تحقیقات ژنومیک و زیست‌شناسی را به جمع‌آوری اطلاعات از طریق جاسوسی در میدان جنگ تشبیه کرد. اطلاعات جمع‌آوری شده همان‌قدر که می‌توانند در جهت‌دهی تاکتیک و راهکار موردنظر موفقیت‌آمیز باشند، از سوی دیگر اعتماد بیش از حد به این اطلاعات هم، می‌تواند اشتباهات زمان‌بر و پرهزینه‌ای به بار آورد.

در مورد مشکلات و چالش‌های این پژوهش، می‌توان به مراحل ابتدایی مرتبط با پیش‌پردازش و نمونه‌برداری توالی‌های DNA اشاره کرد. از یک سو نمونه برداری، مراحل زمان‌بری به همراه دارد و هزینه آن هم به دلیل استفاده از تجهیزات پیشرفته بالاست؛ از سوی دیگر نتایج حاصل از توالی‌یابی، تضمین‌شده نیست، چرا که تنها یک اشتباه کوچک در نمونه‌برداری می‌تواند باعث حذف یک نوکلئوتید از یک توالی با طول چند صد میلیون نوکلئوتید شود که این اشتباه در ظاهر ناچیز، می‌تواند منجر به تحلیل و پردازش اشتباه شود.

همچنین می‌توان به چالش دیگری که به پیشرفت علم ژنتیک مربوط است، اشاره کرد. همان طور که در طول تحقیق اشاره شد، دانشمندان و محققان ژن‌های متعددی روی ژنوم انسان شناسایی کرده‌اند، اما با وجود این تعداد ژن، با پیشرفت علم همچنان ژن‌های جدیدی شناسایی می‌شوند. پس ممکن است داده‌هایی که در این پژوهش، غیرسرطانی فرض شده‌اند، از نوع ژن‌هایی باشند که فعلا ژن آنها کشف نشده است.

۶-۱- پیشنهادها و راهکارهای آتی

راهکارهای ارایه شده را می‌توان از دو زمینه کلی نگاه کرد. یکی این که می‌توان از حیث زیستی و پزشکی به این پژوهش نگاه کرد. یک مورد هم می‌توان به موقعیت‌ها و شرایط این پژوهش از حیث بهینه‌سازی یا استفاده از روش‌های دیگر یا ترکیب با روش‌های دیگر حوزه پردازش سیگنال، به مطالعه و بررسی این ساختارهای ریزمولکولی مرموز پرداخت.

از حیث زیستی، روند بررسی و پردازش الگوریتم‌های ارایه شده به نحوی است که استفاده از این روش‌ها می‌تواند در نمونه‌های ژنومیکی دیگر هم مورد بررسی و مطالعه قرار گیرد. به دلیل وجود بیماری‌های ژنتیکی متعدد دیگر، می‌توان الگوریتم‌های ارایه شده را روی انواع دیگر بیماری سرطان و بیماری‌های ژنتیکی دیگر بررسی کرد. به ویژه این که چون ژنوم انسان با گذر زمان تغییر می‌یابد، شاید بهترین زمان برای بررسی بیماری‌های ژنتیکی، بعد از تولد است. یعنی می‌توان با اعمال این الگوریتم‌ها روی نمونه DNA نوزادان، هر گونه بیماری ژنتیکی را تشخیص داد.

از دیدگاه دیگر، می‌توان روش‌های دیگر تحلیل سیگنال‌های ژنومیکی را برای استخراج ویژگی به کار برد. از این روش‌ها که در طول تحقیق هم بدان اشاره شد، می‌توان به مدل مارکوف، تبدیل موجک و یا درخت‌های تصمیم‌گیر اشاره کرد. همچنین می‌توان به بهبود عملکرد الگوریتم‌های ارایه شده پرداخت. به عنوان مثال در الگوریتم پیشنهادی مبتنی بر مدل LPC که نشان داده شد مقدار مرتبه و طول پنجره تاثیرگذار است، می‌توان با بهره‌گیری از تکنیک‌های بهینه‌سازی، مقادیر بهینه این مولفه‌ها را به دست آورد. به ویژه این که اگر این الگوریتم روی توالی‌های بیشتری و با طول بیشتر اعمال شود، به دلیل وجود نواحی کد شده پروتئینی بیشتر، انتخاب این مقادیر تاثیرگذاری بیشتری

فهرست شکل‌ها

عنوان **صفحه**

شکل ۲-۱: ساختار درونی یک سلول با اجزای داخلی آن ۱۰
شکل ۲-۲: موقعیت مولکول‌های DNA در سلول ۱۲
شکل ۲-۳: ساختار تشکیل دهنده DNA به دو صورت نردبانی و مارپیچی شکل ۱۳
شکل ۲-۴: ساختار شیمیایی نوکلئوتیدهای سازنده DNA ۱۳
شکل ۲-۵: تعداد جفت باز و ژن هر یک از کروموزوم‌های سلول‌های بدن انسان ۱۵
شکل ۲-۶: مراحل بیان ژن تبدیل DNA به پروتئین ۱۶
شکل ۲-۷: مراحل همانندسازی DNA ۱۷
شکل ۲-۸: مراحل رونویسی بیان ژن DNA ۱۷
شکل ۲-۹: مراحل ترجمه بیان ژن DNA ۱۸
شکل ۲-۱۰: تمام حالات تشکیل ORF در ترجمه رشته‌های DNA ۱۹
شکل ۲-۱۱: نحوه رشد و انتشار سلول‌های سرطانی با گذشت زمان ۲۱
شکل ۲-۱۲: انواع جهش‌های نوکلئوتیدی و تاثیر آن روی mRNA و اسیدآمینه حاصل ۲۲
شکل ۲-۱۳: جایگاه علم انفورماتیک زیستی در ارتباط با سایر علوم ۲۳
شکل ۲-۱۴: شاخه‌های علم انفورماتیک زیستی از بزرگ به کوچک ۲۴
شکل ۲-۱۵: مراحل تشکیل آرایه‌های DNA ۲۶
شکل ۲-۱۶: مراحل تشکیل ریزآرایه‌های DNA ۲۷
شکل ۳-۱: مقایسه مراحل طراحی یک سیستم پردازش سیگنال‌های حیاتی با سیستم شناسایی الگو ۲۹
شکل ۳-۲: تاثیر معیار چرخش طیفی در تبدیل فوریه روی نواحی اکسونی ۳۹
شکل ۳-۳: تاثیر اعمال تبدیل فوریه روی توالی با شماره F56F11.4a ۳۹
شکل ۳-۴: تاثیر اعمال معیار SC مختلط تبدیل فوریه روی توالی با شماره F56F11.4a ۴۰
شکل ۳-۵: انواع فیلتر از نظر عبور دادن فرکانس ورودی ۴۰
شکل ۳-۶: نحوه عملکرد فیلتر دیجیتال در شناسایی نواحی اکسونی ۴۱
شکل ۳-۷: تاثیر مقدار R در خروجی فیلتر Anti-notch ۴۱
شکل ۳-۸: نحوه عملکرد فیلتر Anti-notch در آشکارسازی ویژگی تناوب-۳ ۴۲
شکل ۳-۹: نحوه عملکرد مدل LPC در فشرده‌سازی اطلاعات ۴۳
شکل ۳-۱۰: تاثیر روش مبتنی بر مدل LPC در پیش‌بینی نواحی اکسونی روی نمونه F56F11.4 ۴۴
شکل ۳-۱۱: انواع پراکندگی و مرز تفکیک دو نوع داده در فضای ویژگی ۴۵
شکل ۳-۱۲: نحوه جداسازی الگوریتم SVM در تفکیک نمونه‌های دو کلاسه ۴۶
شکل ۴-۱: قالب fasta نمونه سرطانی NM_012403 ۵۳
شکل ۴-۲: نمایش سیگنالی نمونه سرطانی NM_012403 با روش منحنی Z ۵۴
شکل ۴-۳: نمایش سیگنالی نمونه سرطانی NM_012403 با روش منحنی Z بهبود یافته ۵۴
شکل ۴-۴: فلوچارت الگوریتم پیشنهادی مبتنی بر مدل LPC و SVD ۵۵

پیوست الف – علایم اختصاری

اختصارات	معادل انگلیسی
A	Adenine
AC	Approximate Correlation
ACP	Average Conditional Probability
AN	Actual Negative
AP	Actual Positive
AR	AutoRegressive
AUC	Area Under Curve
BLAST	Basic Local Alignment Search Tool
BP	Base Pair
C	Cytosine
CC	Correlation Coefficient
cDNA	Complementary DeoxyRibonucleic Acid
DFT	Discrete Fourier Transform
DNA	Deoxy RiboNucleic Acid
EIIP	Electron Ion Pseudo Potential
FASTA	Fast All
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FN	False Negative
FP	False Posotive
G	Guanine
GSP	Genomic Signal Processing
HGP	Human Genome Project
IIR	Infinite Impulse Response
ISM	Informational Spectra Method
LPC	Linear Predictive Coding
mRNA	messenger RiboNucliec Acid
NCBI	National Center for Biotechnology Information
NIH	National Institute of Health
ORF	Open Reading Frame
PN	Predicted Negative
PP	Predicted Positive
PSD	Power Spectral Density
RNA	RiboNucliec Acid
ROC	Receiver Operating Characteristic
SC	Spectral Content
SNP	Single Nucleotide Polymorphism
SR	Spectral Rotation
STFT	Short Time Fourier Transform
SVD	Singular Value Decompisition
SVM	Support Vector Machines
Sn	Sensitivity
Sp	Specificity

T	Thiamine
TN	True Negative
TP	True Positive
U	Uracil