تخمین مکان نواحی کدکنندهٔ پروتئین در توالی عددی DNA با استفاده پنجره با طول متغیر بر مبنای منحنی سه‌بعدی Z

حمیدرضا صابرکاری*، موسی شمسی و محمد حسین صداقی

دانشکده مهندسی برق، دانشگاه صنعتی سهند تبریز، تبریز، ایران

چکیده

تخمین دقیق نواحی کدکنندهٔ پروتئین در ژن‌ها با استفاده از ابزارهای پردازش سیگنال، در سال‌های اخیر به چالشی در بیوانفورماتیک تبدیل شده است. بسیاری از روش‌های پردازش سیگنال‌های ژنومیک بر مبنای خاصیت تناوب-۳ بازهای موجود در رشته‌های DNA متمرکز بوده و سپس تحلیل‌های طیفی به‌منظور یافتن موقعیت مؤلفه‌های متناوب بر روی توالی‌های عددی DNA اعمال می‌شود. در این مقاله با استفاده از پنجره با طول متغیر و بر مبنای منحنی Z، الگوریتمی به‌منظور تعیین نواحی کدکنندهٔ پروتئین ارائه می‌کنیم. منحنی Z، یک منحنی سه‌بعدی منحصربه‌فرد برای نمایش توالی DNA می‌باشد که توصیف کاملی از رفتار بیولوژیکی توالی DNA را به‌دست می‌دهد. الگوریتم پیشنهادی به‌دلیل استفاده از پنجره گوسی با طول قابل تنظیم، از وضوح و دقت بسیار بالایی در تخمین نواحی ژنی برخوردار بوده و نواحی غیرپروتئینی در آن به‌طورکامل حذف می‌شود. همچنین به‌منظور استخراج مؤلفهٔ تناوب-۳ از یک فیلتر میان‌گذر باند- محدود با فرکانس مرکزی $\frac{2\pi}{3}$ استفاده می‌کنیم. الگوریتم پیشنهادی ابتدا بر روی توالی F56F11.4 در C.elegans اعمال و نتایج آن با سایر روش‌های موجود مقایسه شده و سپس آن را به‌ترتیب برروی ژن‌های موجود در دو پایگاه داده HMR195 و BG570 اعمال می‌کنیم.

واژگان کلیدی: نواحی کدکنندهٔ پروتئین، تناوب-۳، DNA، پنجره با طول متغیر، فیلتر باند- محدود.

Estimation of protein-coding regions in numerical DNA sequences using variable length window method based on 3D Z-curve

Hamidreza Saberkari*, Mousa Shamsi & Mohammad Hossein Sedaaghi

Department of Electrical Engineering, Sahand University of Technology, Tabriz, Iran

Abstract

In recent years, estimation of protein-coding regions in numerical deoxyribonucleic acid (DNA) sequences using signal processing tools has been a challenging issue in bioinformatics, owing to their 3-base periodicity. Several digital signal processing (DSP) tools have been applied in order to Identify the task and concentrated on assigning numerical values to the symbolic DNA sequence, then applying spectral analysis tools such as the discrete Fourier transform (DFT) to locate the periodicity components. Despite of many advantages of Fourier transform in detection of exotic regions, this approach has some restrictions, such as high computational complexity and disability in locating the small length coding regions. In this paper, we improve the performance of the conventional DFT in estimating the protein coding regions utilizing a

* نویسندهٔ عهده‌دار مکاتبات

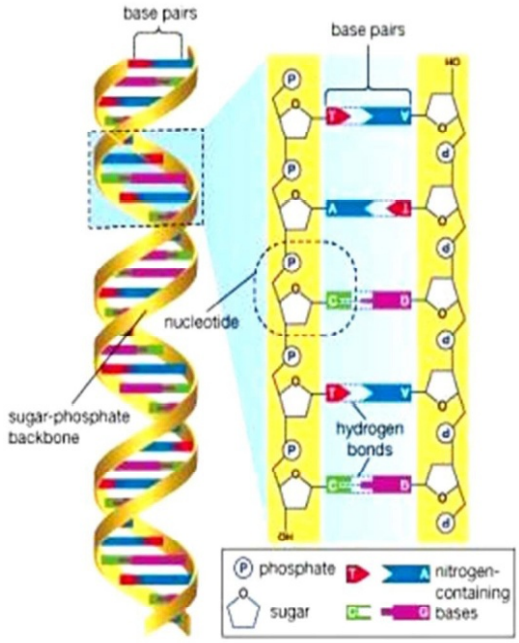
Gaussian window with variable length. First, the DNA strands are converted to numerical signals via the 3-D Z-curve method. Z curve is a robust, independent, less redundant approach, and has clear biological interpretation which can be regarded as a useful visualization technique for DNA analysis of any length. In the second stage, non-coding regions besides the background noise components are completely suppressed using the Gaussian variable length window. Also, we use a narrow-band band-pass filter in order to extract the period-3 components with $\frac{2\pi}{3}$ central frequency. Performance of the proposed algorithm is tested on F56F11.4 from C.elegans chromosome III, also two eukaryotic datasets, HMR195 and BG570, is compared with other state-of-the-art methods based on the nucleotide evaluation metrics such as sensitivity, specificity, approximation correlation, and precision. Results revealed that, the area under the receiver operating characteristic (ROC) curve is improved from 4% to 40%, in HMR195 and BG570 datasets compared to other methods. Furthermore, the proposed algorithm reduces the number of incorrect nucleotides which are estimated as coding regions.

Keywords: Protein coding regions, Period-3, Digital signal processing, DNA, Variable length window, Band-limited filter.

۱- مقدمه

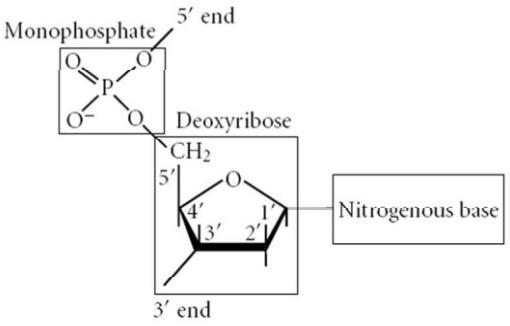
هدف اصلی علم ژنتیک، درک ماهیت اطلاعات نهفته در ژن‌ها و نقش آن‌ها در تعیین عملکردهای خاصی است که توسط ژن‌ها بیان می‌شود. یکی از گام‌های اصلی برای رسیدن به این هدف، تعیین موقعیت ژن‌ها و به‌طور دقیق‌تر، تعیین نواحی کدکنندهٔ پروتئین در رشته‌های DNA [1] است. یک رشته DNA، ملکول طویلی از خانواده بیوپلیمرهاست که اطلاعات ژنتیکی را حمل می‌کند و دارای وظایف بیولوژیکی مهمی از جمله ذخیره و انتقال اطلاعات ژنتیک است. این ملکول، از دو رشته پلیمرخطی تشکیل شده و متشکل از واحدهای منومری به‌نام نوکلئوتید است [1]. شکل۱ ساختار ملکول DNA را نشان می‌دهد. همان‌طورکه مشاهده می‌شود، هر ملکول، حاوی نوکلئوتیدهای مختلف بوده و هر نوکلئوتید از سه بخش تشکیل شده است: یک قند پنج کربنه (قند دئوکسی ریبوز) ، یک تا سه گروه فسفات (PO_4^-) و نیز یک باز آلی نیتروژن‌دار. درشکل(۲)، ساختار یک نوکلئوتید در توالی DNA نشان داده شده است [2].

بازهای موجود در نوکلئوتیدها چهار نوع مختلف دارند؛ آدنین (A) و گوانین (G)، از دسته پورین‌ها بوده و دو حلقه‌ای هستند و تیمین (T)، سیتوزین (C) از دسته پیریمیدین‌ها بوده و تک‌حلقه‌ای هستند. شکل (۳) ساختار بازهای موجود در ملکول DNA را نشان می‌دهد [2]. در DNA بازهای مکمل (آدنین با تیمین و سیتوزین با گوانین) دوبه‌دو با پیوند هیدروژنی به‌هم متصل شده و تشکیل جفت باز[2] می‌دهند [2]. بنابراین، دو رشته DNA مکمل هم بوده و چون جهت‌های عکس هم دارند، موازی معکوس هستند. علاوه براین، رشته‌های DNA جهت‌دار هستند (از انتهای دارای کربن '5 به طرف انتهای دارای کربن '3)، زیرا اتصال به نوکلئوتیدهای جدید همواره از جهت کربن '3 انجام یافته و در نتیجه تشکیل توالی جدید DNA در جهت '5 به '3 انجام می‌گیرد. همچنین به هنگام رمزدهی ساخت پروتئین، رشته DNA در این جهت پویش می‌شود.



(شکل-۱): ساختار ملکول DNA [2]
(Figure-1): DNA molecule structure [2]

[1] Deoxyribonucleic Acid
[2] Base-Pair



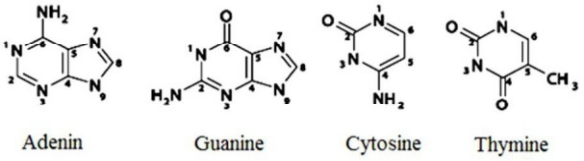
(شکل-۲): ساختار یک نوکلئوتید در ملکول DNA
(Figure-2): Structure of a nucleotide in DNA molecule

هر سه نوکلئوتید متوالی در توالی DNA، کدون[1] نامیده می‌شود. هرکدون یک اسیدآمینه خاص را فرا می‌خواند. پروتئین‌ها به‌عنوان واحدهای اساسی در بسیاری از فرآیندهای بیولوژیکی در داخل سلول، بافت و یا ارگانیسم‌های زنده، از آمینواسیدها تشکیل می‌شوند. سنتز پروتئین‌ها با استفاده از کدهای ژنتیکی صورت می‌گیرد؛ به‌گونه‌ای که ۶۴ کدون در توالی DNA به ۲۰ آمینواسید متناظر نگاشت می‌یابند؛ پس نگاشت از کدون‌ها به اسیدآمینه‌ها یک نگاشت چند به یک است؛ یعنی ممکن است هر اسیدآمینه توسط یک یا چند کدون فراخوانده شود. سه کدون نیز به‌عنوان کدون‌های توقف[2] (کدون‌های TAA، TAG و TGA) وجود دارند که نشان‌دهنده پایان فرآیند پروتئین‌سازی است [3].

در یوکاریوت‌ها، DNA به دو ناحیه ژنی و بین ژنی تقسیم می‌شود. تنها ناحیه ژنی یا به‌اختصار ژن، اطلاعات را برای سنتز پروتئین‌ها حمل می‌کند. هر ژن نیز به نوبهٔ خود متشکل از نواحی اکسون[3] و انترون[4] است که در شکل (۴) نشان داده شده است. بنابراین، اکسون‌ها کدهای لازم را برای تولید پروتئین حمل می‌کنند؛ ازاین‌رو به آن‌ها نواحی کدکنندهٔ پروتین می‌گویند [4]. با مشاهده از انتهای '5 از توالی DNA به‌طرف انتهای '3 می‌توان دریافت به‌طور کلی سه ناحیه کدکننده پروتئین در توالی DNA وجود دارند (شکل۴) که عبارتند از:

- اکسون ابتدایی[5] که از کدون آغازی (کدون ATG) شروع و تا نخستین ناحیهٔ دهندهٔ پروتئین ادامه دارد.

[1] Codon
[2] Stop Codons
[3] Exon
[4] Intron
[5] Initial Exon



(شکل-۳): ساختار بازهای موجود در ملکول DNA [2]
(Figure-3): Structure of bases in DNA molecule [2]

- اکسون داخلی[6] که از یک ناحیهٔ پذیرندهٔ پروتئین شروع و تا نخستین ناحیهٔ دهندهٔ پروتئین بعدی ادامه می‌یابد.
- اکسون انتهایی[7] که از آخرین ناحیهٔ پذیرندهٔ پروتئین شروع شده و تا کدون انتهایی (یکی از کدون‌های TAA، TAG و TGA) ادامه می‌یابد.

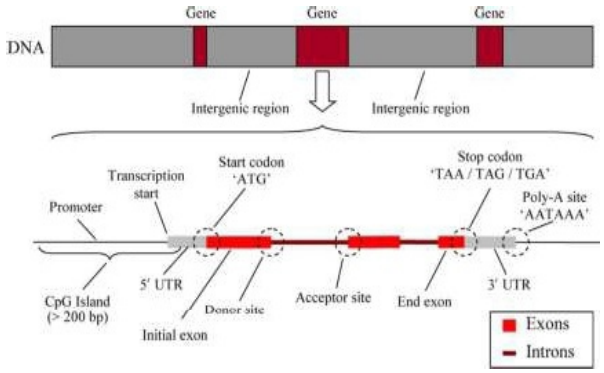
نواحی کدکننده پروتئین دارای مشخصه تناوب-۳[8] هستند که در دیگر قسمت‌های ملکول DNA مشاهده نمی‌شود [5]. این پدیده می‌تواند به‌علت استفاده غیرهمگن از کدون‌ها باشد. این بدین معنی است که با وجود اینکه چند کدون ممکن است یک اسیدآمینهٔ خاص را رمزدهی کنند، همه آن‌ها با احتمال یکسانی در موجودات زنده ظاهر نمی‌شوند. به‌عنوان نوکلئوتید G در موقعیت‌های خاصی در کدون‌های نواحی اکسون جا می‌گیرد [7]-[6]. مشخصه تناوب-۳ می‌تواند به‌عنوان شناساگری برای تشخیص نواحی ژنی بکار رود.

حضور نوفه پس‌زمینه در توالی DNA یک مشکل اساسی در تعیین نواحی ژنی به‌شمار می‌آید [9]-[8]. همچنین به‌دلیل ماهیت پیچیده این نواحی، به‌طورمعمول نیاز به یک ابزار قدرتمندی است که بتواند به‌طور مؤثر مشخصات نواحی کدکننده پروتئین را نمایش دهد. تاکنون روش‌های مختلفی برای رفع این مشکل پیشنهاد شده که در یک تقسیم‌بندی کلی می‌توان آن‌ها را به دو دسته تقسیم-بندی کرد؛ الگوریتم‌های وابسته به مدل[9] یا روش‌های نظارتی[10] و الگوریتم‌های مستقل از مدل یا روش‌های مبتنی بر فیلتر[11]. روش‌های وابسته به مدل مثل مدل مخفی مارکوف[12] [10] و شبکه‌های عصبی[13] [11]، که بر اساس برخی از اطلاعات اولیه جمع‌آوری شده از داده‌های موجود

[6] Donor Site
[7] Acceptor Site
[8] Period-3 Property
[9] Model-Dependent
[10] Supervised Methods
[11] Filter-based Methods
[12] Hidden Markov Model (HMM)
[13] Neural Networks

می‌باشند، به‌طور موفقیت‌آمیزی در تخمین اکسون‌ها در ژن‌ها مورد استفاده قرار می‌گیرند. با این وجود مشکل اصلی این روش‌ها این است که یک ناحیه کدکنندهٔ پروتئین ممکن است، در توالی یک ارگانیسم وجود داشته باشد؛ ولی در پایگاه دادهٔ مورد بررسی نمایش داده نشود. برای رفع این مشکل الگوریتم‌های مبتنی بر فیلتر که بر اساس تحلیل‌های طیفی [13]-[12] استوار هستند، در سال‌های اخیر به ابزارهای مفیدی در ژن‌یابی تبدیل شده است. الگوریتم‌های مختلفی برای تعیین نواحی ژنی بر مبنای خاصیت تناوب-۳ در مراجع پیشنهاد شده است [14]. از تبدیل فوریه برای این‌منظور استفاده کرده است. این روش با انتخاب یک پنجره با طول ثابت و لغزاندن آن بر روی توالی عددی DNA و سپس اعمال تبدیل گسسته فوریه و محاسبه انرژی طیف حاصل‌شدهٔ، نواحی ژنی را تعیین می‌کند [15]. از فیلتر معکوس میان‌گذر تیز[1] با فرکانس مرکزی $2\pi/3$ به‌منظور حذف نواحی غیرپروتئینی استفاده کرده است [16]. از یک الگوریتم جدید بر مبنای تبدیل فوریه و با به‌کارگیری پنجره بارتلت[2] برای تعیین نواحی ژنی استفاده کرده است [17]. از الگوریتم‌های زمانی برای تعیین نواحی کدکننده پروتئینی استفاده کرده است [18]. از ترکیب الگوریتم گورتزل[3] و تبدیل گسسته موجک[4] به‌منظور تعیین نواحی ژنی و حذف نویز پس‌زمینه استفاده کرده است. در تمامی الگوریتم‌های مطرح‌شده از پنجره با طول ثابت ۳۵۱ استفاده شده است. با استفاده از پنجره با طول ثابت نواحی پروتئینی با ابعاد کوچک قابل شناسایی نمی‌باشند و از این‌رو استفاده از الگوریتم‌های مطرح با محدودیت مواجه می‌شوند.

در این مقاله روشی بر مبنای پنجره با طول متغیر جهت تعیین نواحی کدکننده پروتئین در توالی DNA ارائه شده است. استفاده از این الگوریتم منجر به افزایش دقت تخمین نواحی کدکنندهٔ پروتئین و حذف نوفه پس‌زمینه به مقدار قابل توجهی می‌شود. پیکربندی مقاله به‌صورت زیر است؛ در بخش۲، تبدیل داده‌های نویسه‌ای DNA به سیگنال‌های دیجیتال بررسی می‌شود. در بخش۳ تبدیل فوریه کوتاه‌مدت معرفی و مشکل آن مطرح می‌شود. نمودار الگوریتم پیشنهادی در بخش۴ مطرح و مراحل مختلف آن توضیح داده می‌شود. ایده به‌کارگیری از فیلترهای چندطبقه به‌منظور استخراج مؤلفه تناوب-۳ در بخش۵ توضیح داده می‌شود. معیارهای ارزیابی سطح نوکلئوتیدی به‌منظور مقایسه الگوریتم پیشنهادی و سایر روش‌های موجود در بخش۶ بیان می‌شود. نتایج پیاده‌سازی در محیط بسته نرم‌افزاری طراحی‌شده در بخش۷ ارائه شده و درنهایت بخش۸ شامل جمع‌بندی و نتیجه‌گیری از مقاله است.



(شکل-۴): نواحی اکسون و انترون در DNA یوکاریوتی [17]
(Figure-4): Exon/Intron regions in Eukaryotic DNA [17]

۲- تبدیل رشته‌های DNA به سیگنال‌های دیجیتال

تبدیل رشته‌های DNA به سیگنال‌های دیجیتال، امکان اعمال تکنیک‌های پردازش سیگنال را به‌منظور تحلیل داده‌های ژنومیک و همچنین آشکارسازی ویژگی‌های کروموزوم‌ها نتیجه می‌دهد. ابزارهای تحلیل سیگنال‌های ژنومیک در حال حاضر توانایی آشکارسازی سطح وسیعی از ویژگی‌های یک دنباله DNA را در فواصل بیش از $10^8 - 10^6$ جفت باز که شامل هر دو ناحیهٔ پروتئینی و غیرپروتئینی است، در کلیه ژنوم‌ها یا کروموزوم‌ها به‌دست می‌دهند [20]-[19].

با اجرای هرچه بیشتر پروژه‌های ژنی، تعداد زیادی از توالی‌های ژنی در پایگاه‌های عمومی داده در دسترس است که نیاز ضروری به یافتن راه‌حل‌های جدید ریاضی به‌منظور تحلیل این توالی‌ها را نتیجه می‌دهد [21].به‌طورکلی، دو روش پایه ریاضی در مطالعه تئوری‌های فیزیکی وجود دارد؛ یک روش، روش جبری و دیگری روش هندسی است. در بیش‌تر موارد این دو روش مکمل یکدیگر هستند. امروزه، در زمینه مطالعه ژن‌ها، روش جبری در تحلیل توالی ژنی به‌طور گسترده‌ای استفاده شده است، درحالی‌که روش هندسی برای تحلیل توالی ژنی مدت‌زمان زیادی است که صرف‌نظر شده است. منحنی Z یک منحنی فضایی سه‌بعدی است که به‌منظور نمایش توالی DNA به‌کار می‌رود. بر اساس این منحنی، هر توالی DNA می‌تواند به‌طور مجزا با سه توزیع

[1] Anti-Notch Filter
[2] Bartlett Window
[3] Geortzel Algorithm
[4] Discrete Wavelet Transform

غیر وابسته X_n، Y_n، و Z_n توصیف شود که هر یک از این سه توزیع عبارتند از [22]:

$$\begin{cases} x_i = (a_i + g_i) - (c_i + t_i) \\ y_i = (a_i + c_i) - (g_i + t_i) \\ z_i = (a_i + t_i) - (g_i + c_i) \end{cases} \quad x_i, y_i, z_i \in [-1, 1], \quad i = 1, 2, 3. \qquad (۱)$$

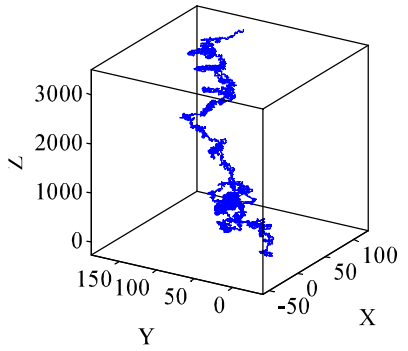
که در آن a_i، c_i، g_i و t_i به‌ترتیب، توزیع‌های تجمعی نوکلئوتیدهای A، C، G و T است. در روابط یادشده مقادیر اولیه a_i، c_i، g_i و t_i صفر فرض شده است ($a_0=c_0=t_0=g_0=0$) [22]. هر یک از این سه توزیع دارای ویژگی‌های بیولوژیکی خاصی هستند. X_n، نماینده حضور نوکلئوتیدهای A یا G که نشان‌دهنده تفاوت بین بازهای پیریمیدین/ پورین (R/Y) در طول توالی DNA است، می‌باشد. به‌طور مشابه، Y_n نشان‌دهنده توزیع بازهای کتو/آمینو (M/K) در طول رشته DNA و Z_n، نشان‌دهنده توزیع پیوندهای هیدروژنی ضعیف / قوی است [23]. برای زیرتوالی تشکیل‌شده از باز نخست تا باز nام از توالی DNA، وقتی بازهای پورین (A یا G) بیش‌تر از بازهای پیریمیدین(T یا C) باشد، $X_n > 0$ و در غیراین‌صورت $X_n < 0$ است. وقتی مقدار بازهای پورین و پیریمیدین مساوی باشد، $X_n = 0$ است. به‌طور مشابه، وقتی بازهای آمینو(A یا G) بیشتر از بازهای کتو (C یا T) باشد، $Y_n > 0$ و در غیراین‌صورت $Y_n < 0$ است و اگر این بازها برابر باشند، $Y_n = 0$ است و درنهایت وقتی بازهای پیوند هیدروژنی قوی (A یا T) بیشتر از بازهای پیوند هیدروژنی ضعیف (G یا C) باشد، $Z_n > 0$ و درغیراین‌صورت $Z_n < 0$ است و درصورتی‌که مقدار این بازها باهم برابر باشد، $Z_n = 0$ است. بنابراین منحنی Z حاوی همه اطلاعاتی است که توالی DNA متناظر با خود دارد. تحلیل توالی DNA می‌تواند بااستفاده از مطالعه منحنی Z متناظر انجام شود. برخی از ویژگی‌های منحنی Z عبارتند از [22]:

- برای یک توالی DNA، یک منحنی Z منحصربه‌فرد متناظر با آن وجود دارد و بالعکس.
- مختصات انتهایی یک منحنی Z برای یک توالی DNA فقط به موقعیت نوکلئوتیدها در آن بستگی دارد و مستقل از ترکیب توالی بازهای DNA است.
- اگر احتمال وقوع فرکانس‌های نوکلئوتیدها برای چهار باز در رشته DNA یکسان باشد، منحنی Z متناظر با آن تشکیل یک حلقه می‌دهد که از مبدأ شروع و به مبدأ ختم می‌شود.
- طول منحنی Z با تعداد بازها در توالی DNA متناسب و مستقل از ترتیب توالی یا ترکیب بازها در توالی است.
- یک ویژگی منحصربه‌فرد منحنی Z این است که در آن سه توزیع X_n، Y_n و Z_n مستقل از یکدیگرند که این امر پیچیدگی محاسباتی مربوط به طیف DNA را کاهش می‌دهد.
- میانگین X_n و Y_n یا X_n و ($-Y_n$) به‌ترتیب اختلاف AT یا اختلاف GC[1] نامیده می‌شود که میزان افزایش A در مقابل T و یا G در مقابل C در زیرتوالی تشکیل‌شده از باز نخست تا باز nام توالی مورد مطالعه را نمایش می‌دهد.
- وقتی یک باز به توالی DNA مورد مطالعه اضافه شود، مختصات x، y و z از منحنی z متناظر باید به اندازه ۱+ یا ۱- افزایش یابد؛ صرف‌نظر از این‌که باز اضافه‌شده A، C، G ویا T باشد.

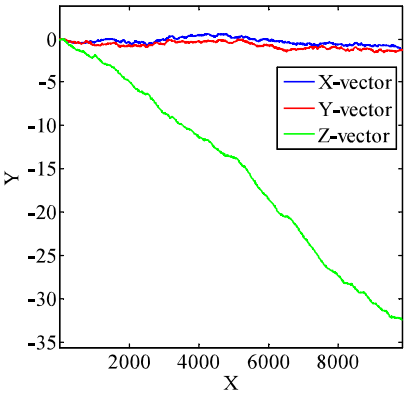
شکل۵-الف، منحنی Z را برای این توالی *F56F11.4* نشان می‌دهد. همان‌طور که مشاهده می‌شود، با استفاده از منحنی Z می‌توان برخی از ویژگی‌های عمومی و محلی مربوط به موقعیت نوکلئوتیدها در یک ژنوم را به‌طوری‌که برای مشاهده‌کننده به‌صورت شهودی قابل درک باشد، نمایش داد؛ درحالی‌که نمایش چنین توالی‌هایی با طول زیاد به‌صورت نویسه‌ای و همچنین استخراج ویژگی از آن‌ها بسیار دشوارمی‌باشد. در شکل ۵-ب، منحنی یک بعدی حاصل از افکنش[1] منحنی Z به‌ترتیب روی محورهای x، y و z نشان داده شده است.

شکل(۶) منحنی اختلاف AT و GC را نشان می‌دهد. این منحنی‌ها به‌ترتیب میزان افزایش نوکلئوتید A نسبت به T و G نسبت به C را در توالی DNA نشان می‌دهد و برای باز نخست تا باز nام یک توالی به‌ترتیب به‌صورت $(x_n+y_n)/2$ و $(x_n-y_n)/2$ تعریف می‌شوند. در یک تقسیم‌بندی کلی منحنی‌های اختلاف AT و GC به سه طبقه تقسیم می‌شوند؛ در طبقهٔ نوع ۱(نوع- AT)، $A_n \approx T_n$ و $G_n \neq C_n$. در طبقهٔ نوع ۲(نوع- GC)، $A_n \neq T_n$ و $G_n \approx C_n$ و در طبقهٔ نوع ۳ (نوع ترکیبی)، $A_n \neq T_n$ و $G_n \neq C_n$ است. همان‌طورکه مشاهده می‌شود، توالی *F56F11.4* در طبقهٔ نوع ۲ قرار داشته که در آن، $A_n \neq T_n$ و $G_n \approx C_n$ (خط آبی) و به‌صورت خطی (خط قرمز) کاهش می‌یابد.

[1] Projection



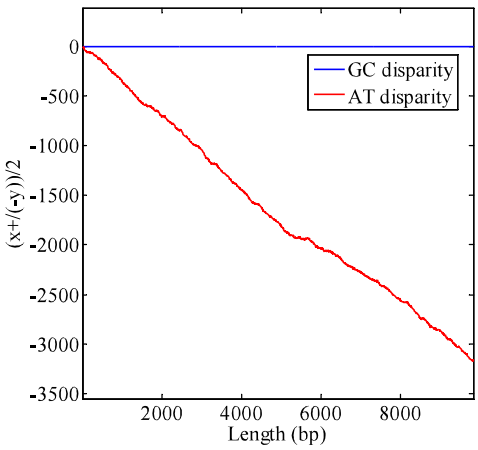
(الف)



(ب)

(شکل-۵): الف. منحنی سه بعدی Z برای توالی F56F11.4. ب. منحنی یک بعدی حاصل از افکنش منحنی Z به‌ترتیب روی محورهای x، y و z

(Figure-5): (a). 3-D Z-Curve for F56F11.4 sequence, (b). 1-D projected curve on x, y, and z axis, respectively.



(شکل-۶): منحنی اختلاف AT و GC در توالی F56F11.4

(Figure-6): AT and GC disparities curve for F56F11.4 sequence

۳- استفاده از تبدیل فوریه کوتاه‌مدت در تعیین نواحی کدکنندهٔ پروتئین و محدودیت آن

بسیاری از روش‌های پردازش سیگنال‌های ژنومیک بر مبنای خاصیت تناوب-۳ بازهای موجود در رشته‌های DNA متمرکز بوده و سپس تحلیل‌های طیفی از جمله تبدیل فوریه کوتاه‌مدت[1] به‌منظور یافتن موقعیت مؤلفه‌های متناوب بر روی توالی‌های عددی DNA اعمال می‌شود. تبدیل STFT، روشی است که در آن یک سیگنال غیرایستا[2] به تعدادی از بلوک‌های کوچک‌تر تجزیه شده و سپس برای هر بلوک، تبدیل فوریه اعمال می‌شود. با استفاده از این روش، تحلیل سیگنال در حوزهٔ زمان- فرکانس به عرض پنجره بستگی خواهد داشت؛ یعنی پنجره‌های با طول زیاد منجر به ایجاد وضوح پایین در حوزه زمان خواهد شد و بالعکس، پنجره‌های با طول کم، منجر به ایجاد وضوح بالا در حوزه زمان می‌شود. این پدیده، مهم‌ترین محدودیت تبدیل STFT به‌شمار آمده و تحلیل سیگنال‌ها را در حوزه زمان با مشکل مواجه می‌سازد. این تبدیل یک تابع چندبعدی بوده، به‌طوری‌که همواره یک سیگنال یک‌بعدی را به صفحه دوبعدی (زمان- فرکانس) نگاشت می‌دهد [24].

۴- استفاده از تبدیل S در تعیین نواحی کدکننده پروتئین و الگوریتم پیشنهادی

تبدیل STFT از پنجره‌هایی با طول ثابت برای محاسبات سیگنال ورودی یک‌بعدی استفاده می‌کند. در بیش‌تر مقالات این طول برابر با ۳۵۱ است. انتخاب این طول نمی‌تواند نواحی پروتئینی با ابعاد کوچک را به‌خوبی شناسایی کند.

به‌منظور رفع این مشکل و بهبود عملکرد تخمین از تبدیل S استفاده می‌کنیم. این تبدیل حالت خاصی از تبدیل STFT است که در آن از یک پنجره گوسی باطول قابل تنظیم استفاده شده است. شکل(۷)، نمودار جعبه‌ای الگوریتم پیشنهادشده را نشان می‌دهد.

طبق تعریف، تبدیل فوریه کوتاه‌مدت برای یک سیگنال غیرایستان $h(t)$ به‌صورت زیر تعریف می‌شود [25]:

$$STFT(\tau,f)=\int_{-\infty}^{\infty}h(t)g(\tau-t)e^{-j2\pi ft}dt \qquad (۲)$$

که در آن τ، موقعیت طیف زمانی، f فرکانس تبدیل فوریه و $g(t)$ تابع پنجره گوسی[3] است.

تابع گوسی قابل تنظیم به‌صورت زیر توصیف می‌شود [25]،[26]:

[1] Short Time Fourier Transform
[2] Non-Stationary
[3] Gaussian Window

$$g(t)=\frac{|f|}{\sigma\sqrt{2\pi}}e^{\frac{-t^2f^2}{2}} \quad (۳)$$

که در آن σ عرض پنجره است. دو ویژگی مهم پنجره گوسی استفاده شده عبارت است از [25]:

- این پنجره در هر دو حوزه زمان و فرکانس دارای تقارن است؛ به‌عبارت دیگر، تبدیل فوریه آن نیز یک گوسی خواهد بود.
- این پنجره فقط دارای لوب اصلی[1] بوده و لوب‌های جانبی در آن وجود ندارد.

تبدیل S با جایگزینی تابع پنجره $g(t)$ با تابع گوسی قابل تنظیم تعریف‌شده در رابطه (۳) به‌صورت زیر به‌دست می‌آید:

$$S(\tau,f)=STFT(\tau,f)=\int_{-\infty}^{\infty}h(t)\frac{f}{\sigma\sqrt{2\pi}}e^{\frac{-(\tau-t)^2f^2}{2}}e^{-j2\pi ft}dt \quad (۴)$$

مطابق رابطه (۴)، تازمانی‌که عرض و ارتفاع پنجره با فرکانس کنترل شود، وضوح فرکانسی در کمترین فرکانس‌ها و متناظر با آن وضوح زمانی در بیش‌ترین فرکانس‌ها به‌ترتیب با پهن‌تر و باریک‌ترشدن پنجره گوسی به‌دست می‌آید [24]،[27].

(شکل-۷): نمودار جعبه‌ای الگوریتم پیشنهادی
(Figure-7): Block diagram of the proposed algorithm

۴-۱- خاصیت خطینگی S و تبدیل و تأثیر نوفه در آن

تبدیل S، یک عملگرخطی روی سری‌های زمانی به‌شمار می‌رود. به‌طور کلی داده‌ها را می‌توان به‌صورت مجموع دو مؤلفهٔ سیگنال و نوفه مدل کرد [28]:

$$\text{داده } t=\text{سیگنال } t+\text{نویز} \quad (۵)$$

با اعمال تبدیل S به رابطه (۵) داریم:

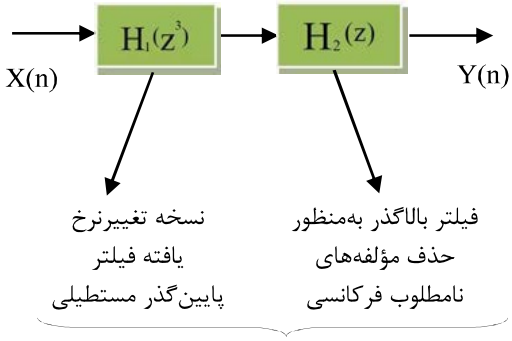
$$S\{\text{داده}\}=S\{\text{سیگنال}\}+S\{\text{نویز}\} \quad (۶)$$

از این‌رو مؤلفه‌های نوفه‌ای به‌دلیل خاصیت خطینگی این تبدیل به‌طور هم‌زمان از سیگنال اصلی تفکیک شده و سیگنال و نوفه با یکدیگر ترکیب نمی‌شوند. این امر باعث افزایش دقت و بهبود عملکرد تخمین نواحی کدکننده پروتئین نسبت به روش‌های پیشین خواهد شد.

به‌طور کلی مزیت استفاده از تبدیل S بجای STFT در این است که در تبدیل S، عرض پنجره σ تابعی از f است. درحالی‌که در تبدیل STFT، این عرض مقدار ثابتی است. این امر باعث افزایش وضوح فرکانسی در تبدیل S می‌شود.

۵- استخراج مؤلفهٔ تناوب-۳ با به‌کارگیری فیلتر چندطبقه

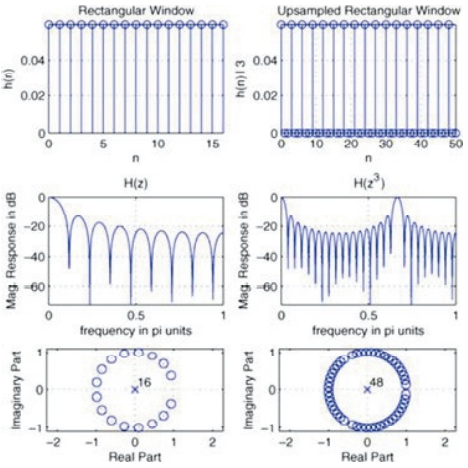
به‌منظور استخراج مولفه تناوب-۳ از فیلتر میان‌گذر باند-محدود استفاده می‌کنیم. فیلتر میان‌گذر استفاده‌شده در این مقاله از نوع فیلترهای چندطبقه است. فیلتر چندطبقه [17] روش مؤثری در طراحی فیلتر میان‌گذر است که با افزایش تعداد ضرب‌کننده‌ها منجر به افزایش تضعیف باند توقف و بهبود عمل فیلتر می‌شود. ایده اصلی این فیلتر در شکل(۸) نشان داده شده که در آن $H_1(z)$، یک فیلتر پایین‌گذر مستطیلی و $H_1(z^3)$، نمونهٔ تغییرنرخ یافته آن است. شکل (۹)، پاسخ ضربه، پاسخ دامنه و نمودار صفر- قطب فیلترهای $H_1(z)$ و $H_1(z^3)$ را نشان می‌دهد. $H_1(z^3)$ باعث ایجاد دو باند عبور نامطلوب در $\omega=0$ و $\omega=\frac{4\pi}{3}$ می‌شود. به‌منظور حذف این دو باند، $H_1(z^3)$ را با فیلتر بالاگذر $H_2(z)$ سری می‌کنیم. شکل (۱۰) پاسخ دامنهٔ فیلتر $H_2(z)$ را نشان می‌دهد. فیلتر H(z)، یک فیلتر میانگذر با فرکانس مرکزی $\frac{2\pi}{3}$ بوده که برای جداسازی مولفه تناوب-۳ در توالی *DNA* استفاده می‌شود. پاسخ دامنه این فیلتر در شکل(۱۱) نشان داده شده است.



$$H(z) = H_1(z^3).H_2(z)$$

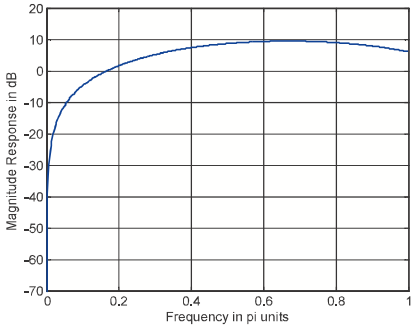
(شکل-۸): ایده فیلترینگ چندطبقه

(Figure-8): Multistage filtering idea



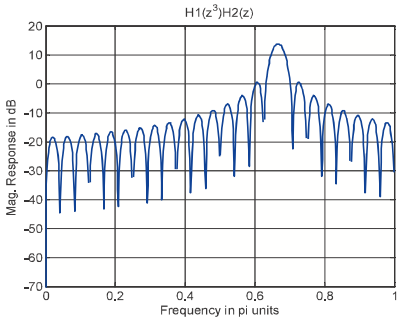
(شکل-۹): پاسخ ضربه، پاسخ دامنه و نمودار صفر- قطب فیلترهای $H_1(z)$ و $H_1(z^3)$ [17]

(Figure-9): Impulse response, amplitude response, and zero-pole diagram of $H_1(z)$ and $H_1(z^3)$ filter [17]



(شکل-۱۰): پاسخ دامنه $H_2(z)$

(Figure-10): Amplitude response of $H_2(z)$



(شکل-۱۱): پاسخ دامنه فیلتر میانگذر $H(z) = H_1(z^3)\,H_2(z)$

(Figure-11): Amplitude response of band-pass filter

۶- معیارهای ارزیابی سطح نوکلئوتیدی

به‌منظور مقایسه دقیق روش‌های گوناگون در تخمین نواحی کدکننده پروتئین، ارزیابی در سطح نوکلئوتیدی انجام گرفته است. درتعیین نواحی ژنی با استفاده از تکنیک‌های پردازش سیگنال، برخی از پارامترها باتغییر سطح آستانه در طیف خروجی تعریف می‌شوند. در این بخش بااستفاده از شکل (۱۲)، به معرفی این پارامترها که امکان مقایسه را به‌دست می‌دهند، می‌پردازیم. در این شکل، TP تعداد نوکلئوتیدهایی است که به‌درستی به‌عنوان اکسون تخمین زده شده‌اند و TN تعداد نوکلئوتیدهایی است که به‌درستی به‌عنوان انترون تخمین زده شده‌اند. به‌طور مشابه، تعداد نوکلئوتیدهای انترون که به‌عنوان اکسون پیشگویی شده‌اند را با FP و تعداد نوکلئوتیدهای اکسون که به‌عنوان انترون پیشگویی شده‌اند را با FN نشان می‌دهیم. با تعریف چهار کمیت فوق، پارامترهای حساسیت[1] (Sn) و ویژگی[2] (Sp) به‌صورت زیر تعریف می‌شوند [29]:

$$Sn = \frac{TP}{TP + FN}$$
$$Sp = \frac{TP}{TP + FP} \qquad (۷)$$

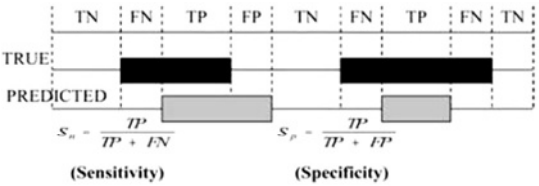
در رابطه (۷)، Sn به‌صورت نسبت نوکلئوتیدهای اکسون که به‌درستی به‌عنوان اکسون شناسایی شده‌اند و Sp، به‌صورت نسبت نوکلئوتیدهای اکسون تخمین زده شده که واقعاً در نواحی ژنی موجودند، تعریف می‌شود.

این پارامترها به‌تنهایی برای ارزیابی مناسب نیستند؛ زیرا در حساسیت بالا مقدار ویژگی پایین بوده و برعکس. بنابراین پارامتر دیگری به‌نام همبستگی تقریبی[3] مطرح کرده که با ترکیب Sn و Sp به‌صورت زیر تعریف می‌شود [29]:

$$ACP = \frac{1}{4}\left(\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FN} + \frac{TN}{TN+FP}\right) \qquad (۸)$$
$$AC = (ACP - 0.5) * 2$$

در اعمال تکنیک‌های پردازش سیگنال به‌منظور جستجوی نواحی ژنی، پارامترهای دیگری نیز تعریف می‌شوند. یک معیار ارزیابی متداول، منحنی (ROC)[4] است که با انتخاب سطوح آستانهٔ مختلف، مقادیر TP برای یک FP داده شده در هر سطح به‌دست می‌آید. ناحیه زیرمنحنی ROC نیز به‌عنوان یک معیار ارزیابی مورد استفاده قرار می‌گیرد؛ به‌طوری‌که مقادیر بزرگ‌تر AUC دقت بالای الگوریتم را نتیجه می‌دهد [30].

[1] Sensitivity
[2] Specificity
[3] Approximation Correlation
[4] Receiver Operating Characteristic



(شکل-۱۲): معیارهای ارزیابی سطح نوکلئوتیدی [29]
(Figure-12): Nucleotide level evaluation criteria [29]

۷- نتایج پیاده‌سازی و بحث

در این بخش نتایج الگوریتم پیشنهادی و سایر روش‌های ارائه‌شده در مراجع با استفاده از معیارهای ارزیابی مطرح شده در بخش ۶، مقایسه شده است. این روش‌ها عبارتند از: *AMDF*[1]، *TDP*[2] [17]، فیلتر معکوس میان‌گذر تیز [4]، تبدیل فوریه گسسته در زمان [14] و *Asif* [16].

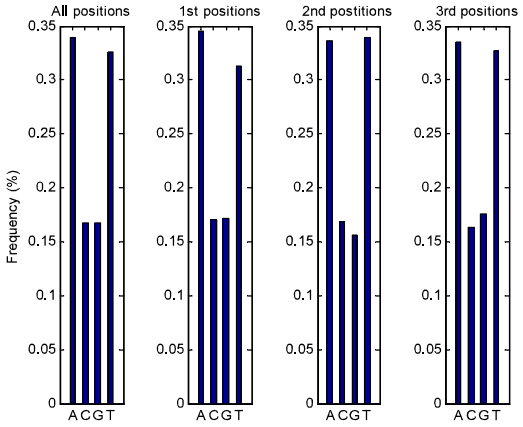
الگوریتم پیشنهادی ابتدا بر روی توالی *F56F11.4* در *C.elegans chromosome III* اعمال شده است. این ژن در *Caenorhabditis elegance* که یک انگل روده‌ای است که طولی در حدود ۱mm داشته و به‌طور طبیعی در محیط‌های معتدل خاکی زندگی می‌کند. طول بازهای آن ۱-۴۲۷۹۹ است و ۸۰۰۰ نوکلئوتید از موقعیت ۷۲۰۱ در آن وجود دارد [31].

شکل (۱۳) نمودار میله‌ای درصد حضور نوکلئوتیدها را برحسب فرکانس آن‌ها در ژن *F56F11.4* نشان می‌دهد. از این نمودار می‌توان به‌منظور محاسبه عددی مقدار محتوای *G+C* در نواحی پروتئینی استفاده کرد. درجدول۱، تعداد نوکلئوتیدها و همچنین نسبت آن‌ها در توالی ژن *F56F11.4* نشان داده شده است. مقدار محتوای *G+C* در این توالی برابر ۰/۳۴ است.

شکل (۱۴)، بسته نرم‌افزاری طراحی‌شده در محیط نرم‌افزار *MATLAB* به‌منظور تحلیل توالی‌های *DNA* را نشان می‌دهد. این بسته از دو بخش اصلی تشکیل یافته است؛ صفحه نمایش گرافیکی و ابزارهای پردازش سیگنال برای تحلیل توالی‌های *DNA*. صفحه نمایش گرافیکی به کاربر این اجازه را می‌دهد که ساختار توالی فراخوانده‌شده را هم به‌صورت گرافیکی و هم در قالب نمادین مشاهده کرده و الگوهای خاصی را در توالی جستجو کند. این بسته امکان فراخوانی انواع توالی‌های *DNA* را در فرمت‌های مختلف نظیر *FASTA*، *TXT*، *PHY* دارد. همچنین ابزارهای پردازش سیگنال، امکان اعمال انواع تبدیل‌های طیفی و عددی را به‌منظور تحلیل توالی‌های *DNA* به‌دست می‌دهند.

برخی از قابلیت‌های بستهٔ نرم‌افزاری طراحی‌شده عبارتند از:

- فراخوانی انواع توالی‌های *DNA* در فرمت‌های مختلف و نمایش آن‌ها به‌صورت گرافیکی و نمادین،
- تبدیل توالی *DNA* به سیگنال‌های عددی با استفاده از روش‌های *EIIP*[3]، دودویی، نگاشت *Z*،
- اعمال روش‌های پردازش سیگنال نظیر *DFT*، تبدیل ویولت، تبدیل *S*،
- مشاهده طیف رنگی انواع ژنوم‌های مختلف و ابزارهای موجود به‌منظور بررسی رفتار آن‌ها،
- جستجوی الگوهای خاص در توالی (نظیر کدون‌های ابتدایی و انتهایی در هر قطعه از توالی *DNA*)،
- پیش‌گویی مکان تقریبی نواحی ژنی با استفاده از روش‌های پردازشی و غیرپردازشی.

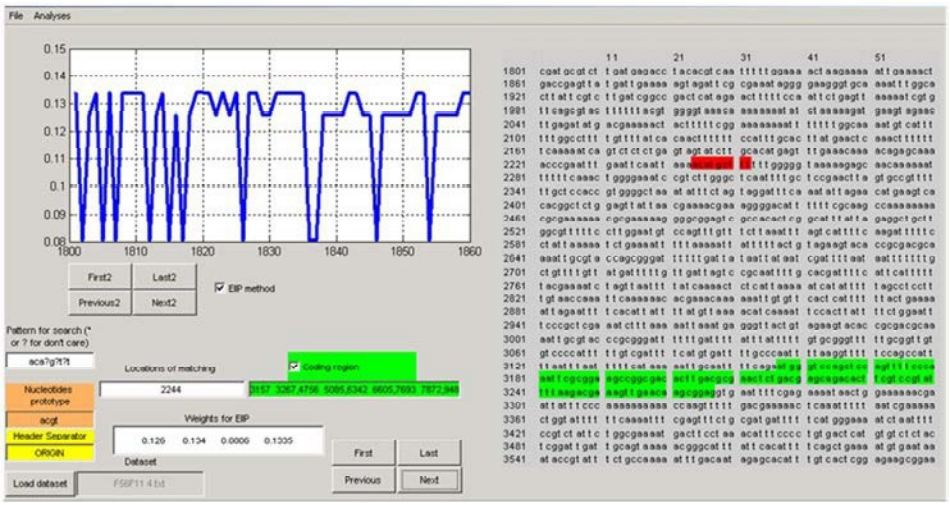


(شکل-۱۳): نمودار میله‌ای درصد حضور نوکلئوتیدها در ژن F56F11.4
(Figure-13): Bar chart of the percentage of nucleotides in F56F11.4 gene sequence

(جدول-۱): تعداد و نسبت نوکلئوتیدها در توالی ژن F56F11.4
(Table-1): Number and proportion on nucleotides in F56F11.4 gene sequence

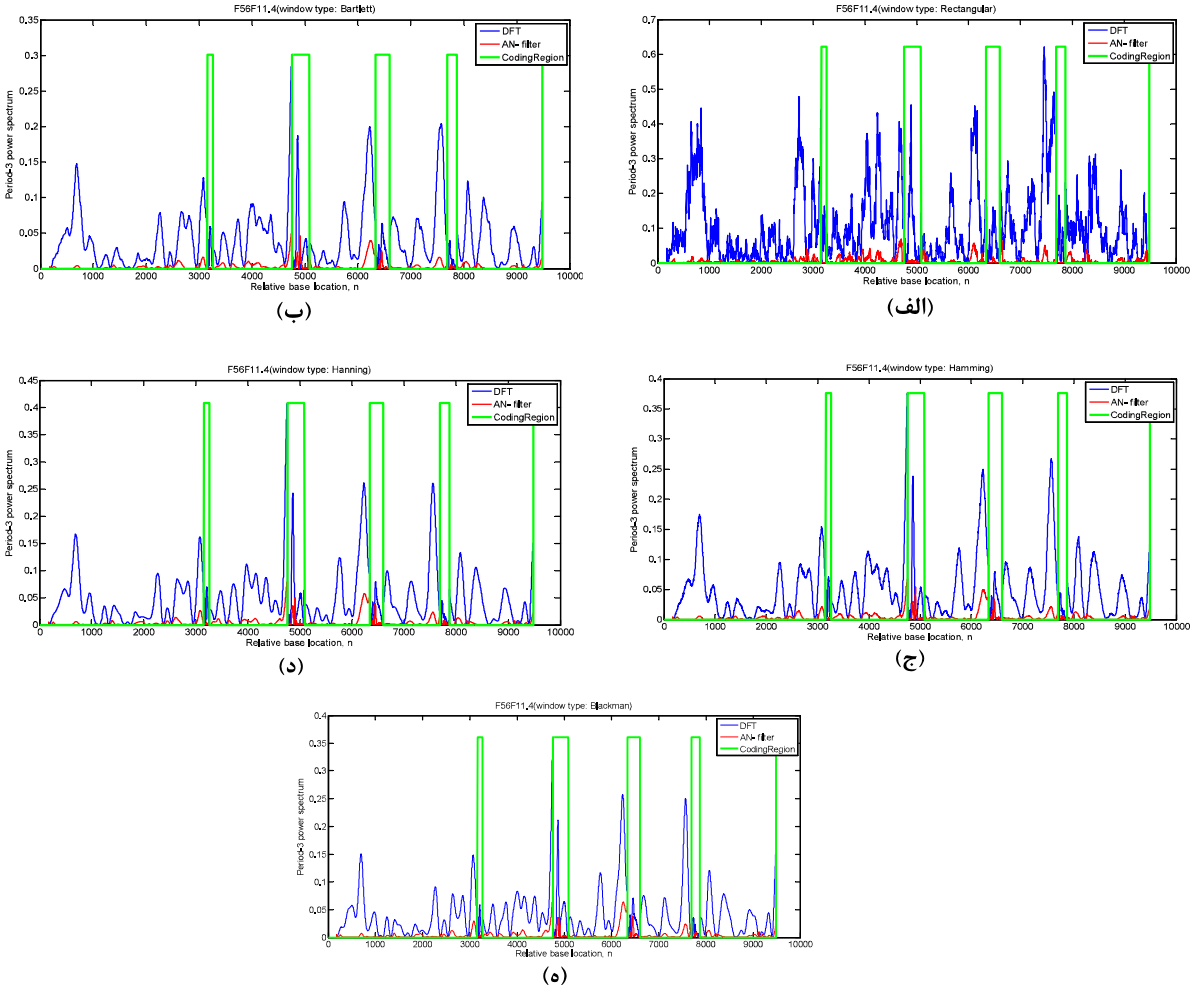
نوکلئوتید	تعداد نوکلئوتیدها	نسبت نوکلئوتیدهادر توالی(٪)
A	3332	٪34
C	1647	٪17
T	1647	٪17
G	3206	٪33

[1] Average Magnitude Difference Function
[2] Time Domain Periodgram
[3] Electron Ion Interaction Potential



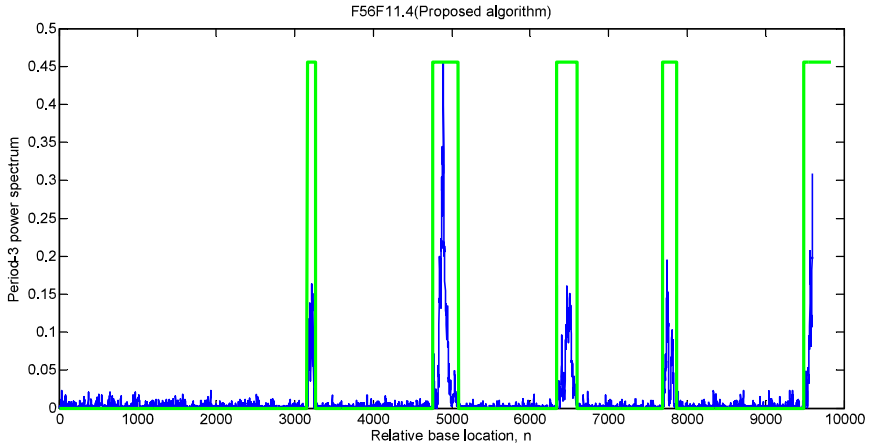
(شکل-۱۴): بسته نرم‌افزاری طراحی‌شده به‌منظور تحلیل عددی و طیفی توالی‌های DNA

(Figure-14): Designed user friendly package in order to numerical and spectral analysis of DNA sequences



(شکل-۱۵): روش‌های DFT و فیلتر معکوس میانگذر تیز در تعیین نواحی کدکننده پروتئین با انتخاب پنجره با طول ثابت ۳۵۱ در حالات (الف): پنجره مستطیلی و (ب): پنجره بارتلت (ج): پنجره همینگ (د): پنجره هنینگ (ه): پنجره بلکمن

(Figure-15): DFT and Anti-notch filter approaches in locating the protein coding regions by selecting the window with the length of 351 in (a). Rectangular window, (b). Bartlett window, (c). Hamming window, (d). Hanning window, and (e). Blackman window



(شکل-۱۶): اعمال الگوریتم پیشنهادی با انتخاب پنجره گوسی با طول قابل تنظیم در ژن F56F11.4

(Figure-16): Applying the proposed algorithm by selecting the variable length window in F56F11.4 gene sequence

شکل‌های۱۵ ، (الف) تا (ه) به‌ترتیب نتایج حاصل از روش‌های *DFT* و فیلتر معکوس میان‌گذر تیز را برروی ژن *F56F11.4* با انتخاب پنجره‌های مختلف نشان می‌دهد. این پنجره‌ها عبارتند از: مستطیلی[1]، بارتلت، همینگ[2]، هنینگ[3] و بلکمن[4]. پنجره بلکمن به‌دلیل اختصاص‌دادن وزن زیاد به کدون‌های موجود در مرکز پنجره نسبت به سایر پنجره‌ها نتیجهٔ بهتری به‌دست می‌دهد. در تمام حالات از پنجره به‌طول ۳۵۱ استفاده شده است. انتخاب این طول نمی‌تواند نواحی پروتئینی با ابعاد کوچک (اکسون موجود در ناحیه ۳۰۰۰ جفت باز) را به‌خوبی شناسایی کند. درحالی‌که این مشکل در الگوریتم پیشنهادی وجود ندارد؛ همانطورکه در شکل (۱۶) نشان داده شده، مهم‌ترین ویژگی این الگوریتم، حذف نوفه پس‌زمینه به‌میزان بسیار زیاد و همچنین شناسایی نواحی کدکننده با ابعاد کوچک است.

درجدول (۲)، مقادیر *AC* و *Sp* به‌ازای یک مقدار ثابت *Sn* در الگوریتم پیشنهادی و سایر الگوریتم‌ها در توالی *F56F11.4* نشان داده شده است. همانطور که مشاهده می‌شود الگوریتم پیشنهادی بیشترین مقدار این دو پارامتر را دارا است.

الگوریتم پیشنهادی بر روی کروموزوم سوم از *Caenorhabditis* که در مجموع شامل ۱۳۷۸۳۶۸۱ نوکلئوتید با ۸۱۷۲ ناحیه کدکننده است [31]، اعمال و نتایج آن با دو روش *AN* و *TDP* مقایسه شده که در جدول (۳) نشان داده شده است. برتری الگوریتم پیشنهادی نسبت به دو روش دیگر به‌وضوح در این جدول مشاهده می‌شود؛ به طوریکه در *Sn* برابر ۲۰ درصد، مقدار *FP* در الگوریتم پیشنهادی برابر ۱۲۸ جفت باز است؛ درحالی‌که این مقدار برای روش‌های *AN* و *TDP* به‌ترتیب برابر ۱۵۷ و ۱۹۶ جفت باز است. همچنین الگوریتم پیشنهادی مقدار *AC* را نسبت به روش‌های *AN* و *TDP* به‌ترتیب به‌میزان ۳٪ و ۶٪ بهبود می‌بخشد.

درنهایت الگوریتم پیشنهادی بر روی مجموعه ژن‌های موجود در دو پایگاه داده *HMR195* [32] و *BG570* [29] اعمال و نتایج آن با سایر روش‌های مطرح‌شده مقایسه شده است. در جدول (۵) ویژگی‌های این دو پایگاه داده به‌طور خلاصه ذکر شده است. پایگاه *HMR195* حاوی ۱۹۵ توالی ژن مربوط به انسان، موش و موش صحرایی است که در سال ۲۰۰۱ توسط *Rogic* و همکارانش تهیه شد. پایگاه *BG570* نیز حاوی ۵۷۰ توالی ژن است که توسط *Burset* و *Guigo* تهیه شده است.

(جدول-۲): مقایسه الگوریتم‌های مختلف در توالی F56F11.4

(Table-2): Comparison of different algorithms in F56F11.4 gene sequence.

Method	Sn	Sp	AC
DFT	0.80	0.18	0.09
AN-filter	0.80	0.24	0.26
Asif	0.80	0.19	0.13
AMDF	0.80	0.21	0.20
TDP	0.80	0.50	0.56
Proposed	**0.80**	**0.83**	**0.79**

[1] Rectangular Window
[2] Hamming Window
[3] Hanning Window
[4] Blackman Window

(جدول-۳): مقایسه الگوریتم‌های مختلف در کروموزوم III از C.elegans

(Table-3): Comparison of different algorithms in chromosome 3 of C. elegans.

		Sn								
		%20			%40			%60		
Methods	**AUC**	**Fp**	**Sp**	**AC**	**Fp**	**Sp**	**AC**	**Fp**	**Sp**	AC
AN-filter	0.6471	157	71	0.17	372	66.3	0.21	727	60	0.20
TDP	0.6115	196	70	0.15	436	65	0.18	796	59	0.19
Proposed	**0.7091**	**128**	**80.5**	**0.24**	**293**	**76.5**	**0.29**	**615**	**61**	0.30

(جدول-۴): خلاصه‌ای از دو پایگاه داده HMR195 و BG570

(Table-4): Summarizing of two HMR195 and BG570 datasets

داده	ارگانیسم	# توالی ژن	# جفت باز	exon #	چگالی نواحی پروتئینی (٪)
HMR195	پستانداران	195	1383720	948	14
BG570	مهره‌داران	570	2892149	2649	15.37

(جدول-۵): مقایسه الگوریتم پیشنهادی با سایر روش‌های موجود در دو پایگاه داده BG570 و HMR195

(Table-5): Comparison of proposed algorithm and other different methods for two HMR195 and BG570 datasets.

	BG570							**HMR195**						
			Sn							**Sn**				
			%10		%30		%50			**%10**		**%30**		**%50**
Methods	AUC	Fp	Sp	Fp	Sp	Fp	Sp	AUC	Fp	Sp	Fp	Sp	Fp	Sp
DFT	0.6540	279	45.8	767	43.3	1412	34.3	0.6782	438	51.5	1184	45	2064	41.7
AN-filter	0.6765	121	55	499	49.7	1103	36.7	0.7615	151	64.4	526	57.4	1217	51.1
Asif	0.5748	140	34.2	330	31.7	554	29.1	0.6261	214	47.1	473	44.6	787	39.9
AMDF	0.6600	340	40.8	770	39.4	1309	35.3	0.6980	410	47.9	1010	46.8	1821	43.3
TDP	0.7560	160	62	408	56	805	49.4	0.7850	262	64.8	627	60.4	1128	56
Proposed	**0.8743**	**79**	**80.5**	**214**	**72**	**534**	**64**	**0.8850**	**134**	**76**	**392**	**72**	**828**	**64**

(جدول-۶): مقادیر همبستگی تقریبی برای داده‌های ژنی موجود در دو پایگاه داده BG570 و HMR195

(Table-6): Quantities of approximation correlation for gene sequences of two HMR195 and BG570 datasets

	BG570			HMR195		
Method	**Sn**	**Sp**	**AC**	**Sn**	**Sp**	AC
DFT	0.80	0.28	0.18	0.80	0.31	0.18
AN-filter	0.80	0.26	0.17	0.80	0.39	0.32
Asif	0.80	0.25	0.10	0.80	0.30	0.15
AMDF	0.80	0.29	0.20	0.80	0.37	0.27
TDP	0.80	0.37	0.31	0.80	0.44	0.38
Proposed	**0.80**	**0.48**	**0.45**	**0.80**	**0.51**	**0.53**

(جدول-۷): مقایسه کمی زمان اجرای الگوریتم پیشنهادی با سایر روش‌های موجود در پایگاه‌های داده GenBank، HMR195 و BG570

(Table-7): Comparison of execution time of proposed algorithm and other methods for gene sequences in GenBank, HMR195 and BG570 datasets.

Gene identifier	Sequence Length (bp)	Average Computational Time (Second)		
		Proposed method	AN-filter	DFT
F56F11.4	9833	**695.2015**	714.6968	718.4017
AF009962	7422	**450.0254**	712.2368	391.0609
AJ223321.1	5321	**639.3497**	710.5081	546.2907

نتایج حاصل‌شده از اعمال الگوریتم پیشنهادی و سایر روش‌ها در جدول (۵) برروی داده *HMR195* نشان داده شده است. همان‌طور که مشاهده می‌شود، الگوریتم پیشنهادی کمترین تعداد نوکلئوتیدهای نادرست شناخته‌شده به‌عنوان اکسون را دارد. در *Sn* برابر با ۳۰ درصد، تعداد نوکلئوتیدهای نادرست در الگوریتم پیشنهادی با ضریب ۱/۲۱ نسبت به بهترین روش مطرح‌شده در مراجع، یعنی روش *Asif*، بهبود می‌یابد. همچنین الگوریتم پیشنهادی، سطح زیر منحنی *ROC* را نسبت به روش‌های *DFT*، فیلتر *AN*، *Asif*، *AMDF* و *TPD* به‌ترتیب به‌میزان ۲۳٪، ۱۴٪، ۲۹٪، ۲۱٪ و ۱۱٪ بهبود می‌بخشد. برتری مشابهی نیز در الگوریتم پیشنهادی برای داده *BG570* وجود دارد که در جدول (۵) نشان داده شده است.

جدول (۶) مقدار *AC* را در الگوریتم پیشنهادی و سایر روش‌های مطرح‌شده نشان می‌دهد. در *Sn* برابر با ۸۰ درصد، مقدار *AC* در الگوریتم پیشنهادی برابر با ۴۵ درصد در داده *BG570* است؛ درحالی‌که مقدار آن در روش *TDP* برابر ۳۱درصد است.

در جدول (۷) زمان اجرای الگوریتم و مقایسه آن با سایر روش‌ها برای توالی‌های *F56F11.4*، *AJ223321.1* از پایگاه *HMR195* و *AF009962* از پایگاه *BG570* نشان داده شده است. لازم به ذکر است که پیچیدگی الگوریتم پیشنهادشده در این مقاله به‌دلیل استفاده از پنجره با طول متغیر، در حدود تبدیل فوریه گسسته در زمان و برابر $O(n^2)$ است. برای این منظور در توالی ژن *F56F11.4*، ابتدا این توالی به پنج زیرتوالی تقسیم و بر روی هر یک الگوریتم پیشنهادشده اعمال شده است. با انجام این کار پیچیدگی الگوریتم به $O(n\log n)$ تنزل پیدا می‌کند.

۸-نتیجه‌گیری

در این مقاله، یک الگوریتم جدید بابه‌کارگیری پنجره گوسی با طول قابل تنظیم و برمبنای منحنی سه‌بعدی *Z* به‌منظور افزایش دقت در تخمین نواحی کدکنندهٔ پروتئین ارائه شد. استفاده از ایده فیلترینگ چند طبقه به‌منظور استخراج مولفه تناوب-۳، امکان شناسایی نواحی ژنی را آسان‌تر می‌سازد. یک مزیت مهم الگوریتم پیشنهادی حذف نوفه پس‌زمینه به‌میزان بسیار زیاد در آن به‌علت استفاده از پنجره گوسی با طول قابل تنظیم است. توانایی آشکارسازی نواحی اکسون با ابعاد کوچک نیز یکی دیگر از مزایای این الگوریتم است. با مقایسه الگوریتم پیشنهادی با سایر روش‌های موجود، مشاهده می‌شود که این الگوریتم برای داده‌های *HMR195* و *BG570*، سطح زیر منحنی *ROC* را از ۴ درصد تا ۴۰ درصد بهبود می‌بخشد. روش پیشنهادی ما همچنین تعداد نوکلئوتیدهای نادرستی را که به‌عنوان نواحی کدکننده تخمین زده شده‌اند کاهش می‌دهد. این کاهش تعداد نوکلئوتیدهای نادرست منجر به افزایش میزان *Sp* خواهد شد. برای مثال، در *Sn* برابر ۳۰ درصد، میزان بهبود مقدار *Sp* در الگوریتم پیشنهادی نسبت به سایر روش‌ها از ۱۵ درصد تا ۸۵ درصد است.

9-Refrences ۹-مراجع

[1] D. P. Snustad, and M. J. Simmons, *Principles of Genetics*, John Wiley & Sons Inc, 2000.

[2] E. R. Dougherty, et al., *Genomic signal processing and statistics*, EURASIP Book Series on Signal Processing and Communications, 2005.

[3] D. L. Brutlag, Understanding the human genome, *Eds. New York: Scientific American*, 1994.

Applications (ISIEA), Bandung, Indonesia, pp. 354-359, September 2012.

[16] S. Datta, A. Asif, "A Fast DFT-Based Gene Prediction Algorithm for Identification of Protein Coding Regions," Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing 2005.

[17] M. Akhtar, J. Epps, E. Ambikairajah, "Signal Processing in sequence Analysis: advanced in Eukaryotic gene Prediction," *IEEE journal of selected topics in signal processing*, vol. 2, pp. 310-321, 2008.

[18] H. Saberkari, M. Shamsi, H. Heravi, and M. H. Sedaaghi, "A Fast Algorithm for Exonic Regions Prediction in DNA," *Journal of Medical Signals and Sensors*, vol. 3, no. 3, pp. 139-149, 2013.

[19] J. M. Claverie, "Computational methods for the identification of genes in vertebrate genomic sequences," *Hum. Mol. Genet*, vol. 6, no. 10, PP. 1735-1744, 1997.

[20] W. F. Doolittle, "Phylogenetic classification and the universal tree," *Science*, vol. 284, no. 5423, pp.2124–2128, 1999.

[21] F. Gao, and C. T. Zhang, "GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences," *Nucleic Acids Research*, vol. 34, pp. 686-691, 2006.

[22] C. T. Zhang, and R. Zhang, "analysis of distribution of bases in the coding sequences by a diagrammatic technique," *Nucleic Acids Research*, vol. 19, pp. 6313-6317, 1991.

[23] A. Rushdy, and J. Tuqan, "Gene Identification using the Z-Curve Representation," International Conference on Acoustics, Speech, and Signal Processing, pp. 1024-1027, 2006.

[24] R. G. Stockwell, Why Use the S-Transform? Boulder, CO: Fields Institute Communications, 2007.

[25] S. S. Sahu, G. Panda, "Identification of protein-coding regions in DNA sequences using a time-frequency approach," *Genomics Proteomics Bioinformatics*, vol. 9, pp. 45-55, 2011.

[26] M. K. Hota, and V. K. Srivastava. "DSP Technique for Gene and Exon Prediction Taking EIIP Indicator Sequence," In Proceedings of the Second International

[4] P. P. Vaidyanathan, and B. J. Yoon, "The role of signal processing concepts in genomics and proteomics," *Journal of Franklin Institute, special issue on Genomics*, 2004.

[5] E. N. Trifonov, and J. L. Sussman, "The pitch of chromatin DNA is reflected in its nucleotide sequence," *Proc. of the Nat. Acad. Sci.*, vol. 77, pp. 3816–3820, 1980.

[6] X. F. Wan, D. Xu, A. Kleinhofs and J. Zhou, "Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes," *BMC Evolutionary Biology*, vol. 4, no. 19, 2004.

[7] H. E. Herzel, N. Trifonov, O. Weiss, and I. Groβe, "Interpreting correlations in bio-sequences," *Physica A*, vol. 249, pp. 449–459, 1998.

[8] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Phy. Re. Lett*, vol. 85, pp. 1342-1345, 1992.

[9] C. A. Chatzidimitriou, and D. Larhammar, "Long-range correlations in DNA," *Nature*, vol. 361, pp. 212-213, 1993.

[10] J. Henderson, J, et al., "Finding genes in DNA with a Hidden Markov Model," *J. Comput. Biol*, vol. 4, pp. 127-141, 1997.

[11] C. H. Ding, and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, pp. 349-358, 2001.

[12] D. Anastassiou, "Genomic signal processing," *IEEE Sign. Proc. Mag,* vol. 18, pp. 8-20, 2001.

[13] T. W. Fox, and A. Carreira, "A digital signal processing method for gene prediction with improved noise suppression," *EURASIP J. Appl. Aign. Proc*, pp. 108-114, 2004.

[14] S. Tiwari S, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Comput Appl Biosci*, vol. 13, pp. 263-270, 1997.

[15] H. Saberkari, M. Shamsi, M. H. Sedaaghi, and F. Golabi, "Prediction of protein coding regions in DNA sequences using signal processing methods," 2012 IEEE Symposium on Industrial Electronics and

Conference on Information Processing, 117-123. Piscataway, NJ: IEEE Press, 2008.

[27] R. G. Stockwell, L. Mansinha, and R. P. Lowe. "Localization of the Complex Spectrum: The S-Transform, " *IEEE Trans. on Sig. Process.* vol. 44, no. 4, 998-1001, 1996.

[28] Yu. H. Wang, "The Tutorial: S-Transform," Accessed June 30, 2011.

[29] M. Burset, R. Guigo, "Evaluation of gene structure prediction programs," *Genomics*, pp. 353-367, 1996.

[30] M. Akhtar, E. Ambikairajah, J. Epps, "Detection of period-3 behavior in genomic sequences using singular value decomposition," Proceedings of the International Conference on Emerging technologies, 2005.

[31] National Center for Biotechnology Information, National Institutes of Health,National Library of Medicine, http://www.ncbi.nlm.nih.gov/GenebANk/index.html.

[32] S. Rogic S, A. K. Mackworth, and B. F. Ouellette, "Evaluation of gene-finding programs on mammalian sequences," *Genome*, vol. 11, pp. 817-832, 2001.

حمیدرضا صابرکاری تحصیلات مقطع کارشناسی خود را در سال ۱۳۹۰ در رشته مهندسی برق- الکترونیک در دانشگاه گیلان به پایان رساند. تحصیلات کارشناسی ارشد خود را در رشته مهندسی برق- مخابرات سیستم در دانشگاه صنعتی سهند تبریز ادامه داد و در سال ۱۳۹۲ مدرک کارشناسی ارشد خود را اخذ کرد. ایشان در حال حاضر در مقطع دکترای تخصصی رشته مهندسی برق- الکترونیک دانشگاه صنعتی سهند تبریز مشغول به تحصیل است. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: پردازش سیگنال و تصویر حیاتی، بیوانفورماتیک و حسگرهای زیستی.

نشانی رایانامه ایشان عبارت است از:

h_saberkari@sut.ac.ir

موسی شمسی مدرک کارشناسی و کارشناسی ارشد خود را به‌ترتیب در سال‌های ۱۳۷۵ و ۱۳۷۹ در رشته‌های مهندسی برق- الکترونیک و مهندسی پزشکی از دانشگاه تبریز و دانشگاه تهران اخذ و همچنین مدرک دکترای خود را در گرایش مهندسی پزشکی در سال ۱۳۸۷ از دانشگاه تهران دریافت کرد. زمینه‌های پژوهشی مورد علاقه ایشان پردازش تصویر و سیگنال حیاتی، شناسایی آماری الگو و الگوریتم‌های بهینه-سازی بوده و در حال حاضر عضو هیأت علمی با مرتبه دانشیاری در دانشگاه صنعتی سهند است.

نشانی رایانامه ایشان عبارت است از:

shamsi@sut.ac.ir

محمد حسین صداقی مدرک کارشناسی و کارشناسی ارشد خود را به‌ترتیب در سال‌های ۱۳۶۵ و ۱۳۶۶ در رشته مهندسی برق از دانشگاه صنعتی شریف اخذ و مدرک دکترای خود را در رشته مهندسی برق در سال ۱۳۷۷ از دانشگاه لیورپول انگلستان دریافت کرد. زمینه‌های پژوهشی مورد علاقه ایشان پردازش تصویر و سیگنال، شناسایی آماری الگو و بیومتریک بوده و در حال حاضر عضو هیأت علمی با مرتبه استادی در دانشگاه صنعتی سهند است.

نشانی رایانامه ایشان عبارت است از:

sedaaghi@sut.ac.ir