فشرده‌سازی سیگنال‌های ژنوم با کمک حسگری فشرده و کاربرد آن در مقایسه دنباله‌های ژنی

محمود طوماری[۱]، کارشناس ارشد مهندسی پزشکی؛ سپیده جباری[۲]، استادیار

۱-گروه برق و کامپیوتر - دانشکده مهندسی - دانشگاه زنجان - زنجان - ایران - m.toumari@znu.ac.ir
۲- گروه برق و کامپیوتر - دانشکده مهندسی - دانشگاه زنجان - زنجان - ایران - sjabbari@znu.ac.ir

چکیده: تحلیل توالی‌های ژنی نقطه شروع درک عملکرد ارگانیسم‌های بیولوژیکی است. در سال‌های اخیر، هزینه‌های توالی‌برداری ژن به‌شدت کاهش یافته است و مقدار زیادی از داده‌های ژنومی درحال تولید هستند. ازطرفی، هزینه حافظه ذخیره‌سازی، پردازش و انتقال این داده‌ها درحال افزایش است. پردازش این حجم عظیم اطلاعات بیشتر توسط روش‌های کاراکتر مبنا صورت می‌گیرد که زمان‌بر است. ظرفیت‌های بالقوّه فراوانی برای مقابله با این چالش‌ها در حوزه پردازش سیگنال وجود دارد. بنابراین، نگاه سیگنالی به دنباله‌های ژنی، پردازش سیگنال ژنوم و فشرده‌سازی آن می‌تواند مفید واقع شود. فشرده‌سازی سیگنال‌ها هزینه برای آنالیز، فضای حافظه برای ذخیره‌سازی، پهنای باند برای مبادله و زمان مورد نیاز برای تحلیل را کاهش می‌دهد. در این مقاله ابتدا دنباله‌های ژنی کاراکتری به‌صورت سیگنالی بیان شدند. سپس، سیگنال‌های ژنومی حاصل توسط روش حسگری فشرده مبتنی‌بر یادگیری بیزین فشرده‌سازی شدند. توانایی روش در بازسازی سیگنال‌های فشرده شده با استفاده از معیارهای PRD و NMSE مورد بررسی قرار گرفت. سپس به‌منظور مقایسه و بررسی مشابهت دنباله‌ها، درختچه فیلوژنتیک از روی سیگنال‌های فشرده شده با نرخ ۷۵٪ رسم شد. نتایج نشان دادندکه مقایسه دنباله‌ها با روش سیگنال مبنا با صرف زمان بسیار کمتر ۱/۲۸۵۳ ثانیه در مقایسه با روش کاراکتر مبنا با زمان ۱۲۶ ثانیه صورت می‌گیرد.

واژه‌های کلیدی: توالی ژن، فشرده‌سازی، حسگری فشرده، پردازش سیگنال ژنوم، یادگیری بیزین، درختچه فیلوژنتیک.

Genomic Signals Compression by Compressed Sensing and Its Application in Sequences Comparison

Mahmoud Toumari[1], MSc in Biomedical Engineering; Sepideh Jabbari[2], Assistant Professor

1- Electrical Engineering Department, Faculty of Engineering, University of Zanjan, Zanjan, Iran, Email: m.toumari@znu.ac.ir
2-Electrical Engineering Department, Faculty of Engineering, University of Zanjan, Zanjan, Iran, Email: sjabbari@znu.ac.ir

Abstract: The analysis of gene sequences is fundamentally important for exploring biological functions. Recently, the cost of gene sequencing has dropped sharply, thereby resulting in the production of considerable genomic data. However, the costs of saving, processing, and transferring these data are rising. At present, processing this massive volume of information is done by character based method which is highly time - consuming. Alternative methods challenge these problems in the realm of signal processing. Accordingly, the signal outlook to the genome, signal processing of the genome and compression of the genome are presently hot issues which are practically in demand. Compression reduces the cost, memory space, bandwidth for exchange, and the time required for analysis.

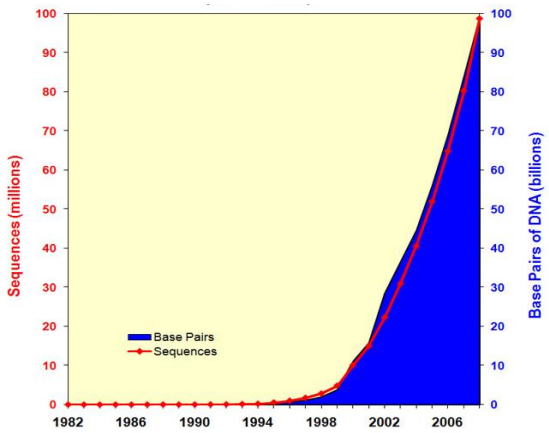
In this study, the character genes were firstly represented as signals. Then, these genomic signals were compressed by compressed sensing. Consequently, they were reconstructed by bayesian learning method. Adopted criteria for reconstruction were PRD and NMSE, respectively. Then, signals were selected with a compression rate of 75% for comparison. Meanwhile, the same cluster analysis was run with character based method. The results indicated that the time needed for signal based method was considerably lower than the character based method.

Keywords: Gene sequence, Compression, Compressed sensing, Genomic signal processing, Bayesian learning, Phylogenetic tree

تاریخ ارسال مقاله: ۱۳۹۶/۱۲/۲۶
تاریخ اصلاح مقاله: ۱۳۹۷/۰۶/۰۶
تاریخ پذیرش مقاله: ۱۳۹۷/۰۶/۱۳
نام نویسنده مسئول: سپیده جباری
نشانی نویسنده مسئول: ایران، زنجان،دانشگاه زنجان، دانشکده فنی، گروه برق، اتاق ۱۰۵، کدپستی ۳۸۷۹۱-۴۵۳۷۱.

۱ -مقدمه

دئوکسی ریبونوکلئیک اسید (DNA[۱]) یک اسید نوکلئیک شامل دستورالعمل‌های ژنتیکی مورد استفاده در رشد و عملکرد همه ارگانیسم‌های زنده شناخته شده است. بنابراین، تجزیه و تحلیل توالی‌های DNA نقطه شروع مهمی برای درک و فهم عملکردهای بیولوژیکی است [۱]. از طرفی، پیشرفت‌های اخیر در تکنولوژی توالی‌برداری[۲] سبب رشد نمایی در حجم داده‌های دنباله‌های ژنومی در دسترس گردیده است [۲]. شکل ۱ وضعیت رو به رشد حجم داده‌های ژنی در طول بیش از دو دهه را نمایش می‌دهد. یک راهکار مناسب جهت ذخیره، مدیریت، تحلیل و مقایسه این داده‌های حجیم فشرده‌سازی آنها می‌باشد. فشرده‌سازی منجر به دیسک سخت کمتر جهت ذخیره‌سازی، پهنای باند کمتر جهت مبادله داده و هزینه کمتر جهت تحلیل و مقایسه می‌شود. بر این اساس، فشرده‌سازی دنباله‌های ژنی از زمینه‌های تحقیقاتی موردتوجه در سال‌های اخیر بوده است [۴].



شکل ۱: وضعیت رشد داده‌های ژنی از سال ۱۹۸۲-۲۰۰۸ [۳]

۱ ـ۱ -ضرورت پردازش سیگنال ژنوم

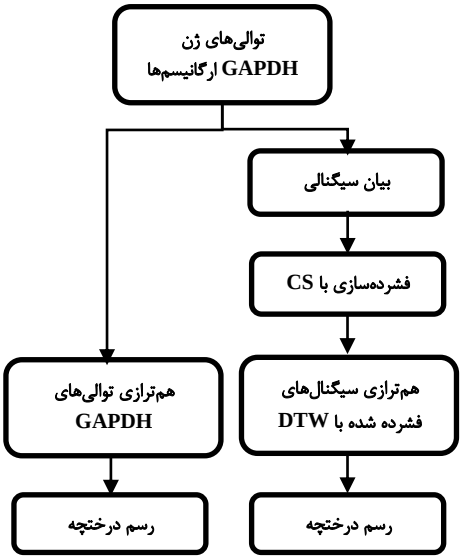
از زمان تکمیل برنامه کل ژنوم انسان، تجزیه و تحلیل و مقایسه اطلاعات موجود در حجم رو به رشد پایگاه داده‌های مربوط به توالی DNA انسان و سایر ارگانیسم‌ها مدنظر قرار گرفت [۵]. امروزه، مقایسه دنباله‌های ژنومی با روش‌های مبتنی بر کاراکتر صورت می‌گیرد. هم ترازی[۳] چندین دنباله و پیچیدگی روش‌های مقایسه مبتنی بر کاراکتر، به‌کارگیری این روش‌ها را محدود به مجموعه داده‌های با طول کم می‌کند و فقط بخش کوچکی از دنباله‌ها را می‌توان در زمانی معقول مورد مقایسه قرار داد [۶]. از طرفی، بیان کاراکتر - مبنای دنباله‌های DNA فقط می‌تواند تفاوت‌ها و شباهت‌های محلی بین توالی‌ها را آشکار کند و مقایسه دنباله‌های با طول زیاد برای چشم انسان ناراحت کننده است.

یک روش جایگزین جهت تحلیل داده‌ها، پردازش سیگنال ژنوم حاصل از دنباله‌های‌کاراکتری است. پردازش سیگنال ژنوم (GSP[۴]) اشاره به استفاده از روش‌های پردازش سیگنال برای تحلیل داده‌های ژنومی دارد. با بیان سیگنالی، ویژگی‌های توالی‌ها در مقیاس بزرگتری دیده می‌شود و اطلاعات برای چشم انسان قابل رؤیت است. بعلاوه اینکه، نتایج تحقیقات اخیر کارآیی روش را در هم‌ترازی دنباله‌ها و مقایسه ارگانیسم‌ها با یکدیگر تأیید می‌کنند. نتیجه اینکه بیان سیگنالی و پردازش سیگنال ژنوم از زمینه‌های رو به رشد مطرح در حوزه بیوانفورماتیک است [۶].

روش‌های مطرح در GSP تبدیل یا نگاشت دنباله‌های کاراکتری ژنی به یک بیان عددی را می‌طلبند. تا به امروز، چندین بیان عددی متناظر با دنباله‌های DNA پیشنهاد شده‌اند که از میان آنها می‌توان به روش‌های نگاشت تک مقداری، بیان چند بُعدی و بیان انباره‌ای اشاره کرد [۵].

فشرده‌سازی سیگنال ژنومِ حاصل از دنباله‌های کاراکتری برای اولین بار توسط سدلار و همکارانش مطرح شد [۷]. آنها از تبدیل موجک گسسته (DWT[۵]) با موجک مادر هار بدین منظور استفاده کردند و روش پیشنهادی خود را بر روی مجموعه‌ای از ژنوم ۱۵ باکتری اعمال نمودند. طبقه‌بندی سیگنال‌های فشرده شده با آنالیز خوشه‌ای نشان دهنده توانایی روش در مقایسه دنباله‌های ژنی بود.

ما در این مقاله روش حسگری فشرده (CS[۶]) را به منظور فشرده‌سازی سیگنال‌های ژنوم به‌کار گرفته‌ایم. در این روش، با انتخاب میزان خطای بازسازی مشخص می‌توان سیگنال را با نرخ دلخواه فشرده و سپس بازسازی کرد. در مرحله بعد، خوشه‌بندی ارگانیسم‌ها را با بکارگیری این سیگنال‌های فشرده شده مدنظر قرار داده‌ایم. به‌منظور مقایسه، عمل خوشه‌بندی را با استفاده از دنباله‌های کاراکتری DNA ارگانیسم‌ها و با کمک نرم‌افزار CLC S.V.[۷] نیز انجام داده‌ایم. بلوک‌دیاگرام شکل ۲ روند روش پیشنهادی در کنار روش متداول کاراکتر مبنا را نشان می‌دهد.



شکل ۲: بلوک‌دیاگرام روش پیشنهادی

۲ -مواد و روش‌ها

۲ ۱ -دادگان مورد استفاده

جهت بررسی قابلیت فشرده‌سازی سیگنال‌های ژنومی، ما از دنباله‌های کاراکتری مربوط به دوازده ژن گلیسرالدهید -۳- فسفات دهیدروژناز (GAPDH[۸]) متعلق به ارگانیسم‌های مختلف استفاده نمودیم. دنباله‌های فوق از پایگاه داده GenBank در سایت NCBI[۹] در دسترس هستند [۸]. توضیحات مربوط به دنباله‌ها در جدول ۱ آورده شده است. این ژن به‌عنوان یک ژن خانه‌دار[۱۰] در سلول‌های تمام ارگانیسم‌ها حضور داشته و عاملی مهم در حفظ و نگه‌داری عملکردهای اصلی سلول است. دلیل انتخاب یک ژن خانه‌دار، بیان ثابت آن در تمام یا اغلب سلول‌ها جهت واکنش‌های متابولیک سلول است. چنانچه الگوریتم در خوشه‌بندی ارگانیسم‌ها با استفاده از این ژن موفق عمل کند، جداسازی ارگانیسم‌ها مبتنی‌بر ژن‌های فرمانروا[۱۱] که دارای بیان افتراقی در سلول‌های مختلف هستند به‌راحتی امکان‌پذیر خواهد بود.

۲ ۲ -استخراج سیگنال ژنوم از دنباله‌های کاراکتری

تاکنون روش‌های مختلفی جهت بیان سیگنالی توالی‌های کاراکتری DNA و پروتئین‌ها ارائه شده‌است [۹-۱۱]. از جمله این روش‌ها، که تا حد امکان خصوصیات بیولوژیکی دنباله کاراکتری مورد بررسی را حفظ می‌کند، روش نمایش فاز انباره‌ای[۱۲] است. مزیت اصلی این روش یک بُعدی بودن سیگنال حاصل و مناسب بودنش برای فرآیند فشرده‌سازی و نیز هم‌ترازی است [۶، ۷].

در بیان فاز انباره‌ای، هر کدام ازچهار نوکلئوتید تشکیل دهنده توالی DNA شامل آدنین، سیتوزین، تیمین و گوآنین [۱۳](A, C, T, G)، به یک عدد مختلط در صفحه مختلط به‌فرم $A \to [1,\ j],\ C \to [-1,\ -j],\ G \to [-1,\ j],\ T \to [1,\ -j]$ نگاشت می‌شوند.

سیگنال فاز انباره‌ای از جمع نمودن فاز این اعداد مختلط در طول نمونه‌ها مطابق با معادلات زیر ایجاد می‌گردد:

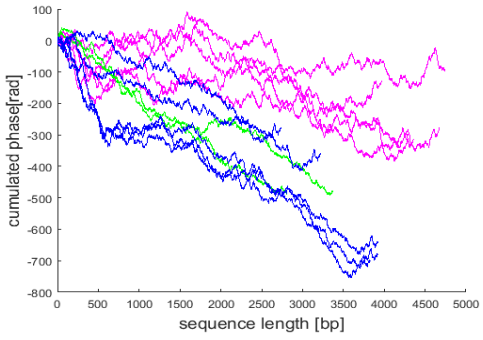
$$\{\varphi_A,\ \varphi_C,\ \varphi_G,\ \varphi_T\} = \left\{\frac{\pi}{4}, -\frac{3\pi}{4}, \frac{3\pi}{4}, -\frac{\pi}{4}\right\} \quad (۱)$$

$$\theta_{cum} = \frac{\pi}{4}\left[3(n_G - n_C) + (n_A - n_T)\right] \quad (۲)$$

که φ_X فاز مربوط به نوکلئوتید X، n_X تعداد نوکلئوتیدهای X از ابتدا تا موقعیت کنونی در توالی و θ_{cum} سیگنال فاز انباره‌ای حاصل است [۶]. شکل ۳ سیگنال‌های متناظر با دوازده دنباله کاراکتری اشاره شده در جدول ۱ را نمایش می‌دهد.

۲ ۳ -فشرده‌سازی با روش CS

حسگری فشرده یک چارچوب جدید جهت نمونه‌برداری و فشرده‌سازی سیگنال به‌صورت هم‌زمان معرفی‌می نماید. شرط لازم جهت بازسازی مؤثر سیگنال اصلی از روی سیگنال فشرده، تنکی یا به‌عبارتی تعداد کم مؤلفه‌های غیر صفر سیگنال است [۱۲]. درCS با استفاده از ماتریسی که از آن به‌عنوان ماتریس حسگری[۱۴] یاد می‌شود، یک تبدیل خطی بر روی سیگنال اعمال کرده و بدین ترتیب آن را به‌منظور کاهش در مصرف انرژی و حجم ذخیره‌سازی و پهنای باند ارسالی، قبل از ارسال و در فرستنده فشرده می‌نمایند. سپس، به کمک همین ماتریس و با فرض تُنُکی سیگنال، سیگنال در گیرنده و از روی نمونه‌های دریافتی بازسازی می‌شود.

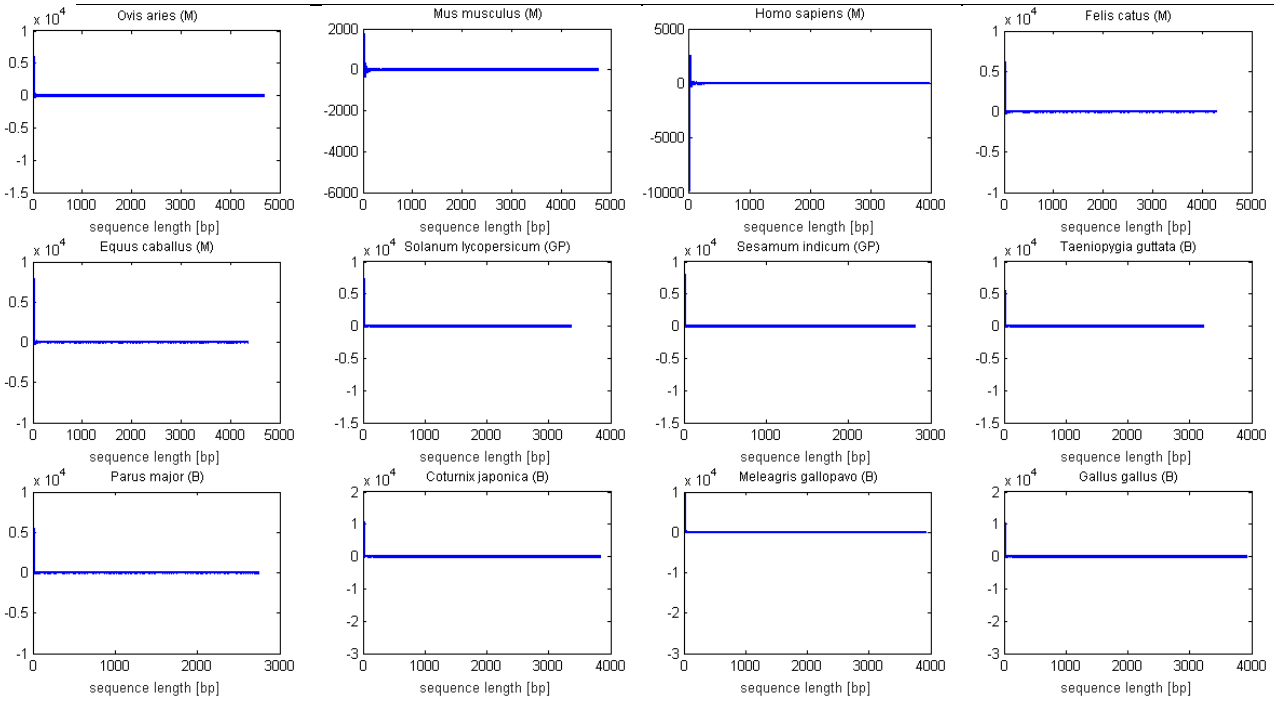


شکل۲: سیگنال‌های فاز انباره‌ای متعلق به ژن‌های GAPDH [پستانداران (ارغوانی)، گیاهان زمینی (سبز)، پرندگان (آبی)]

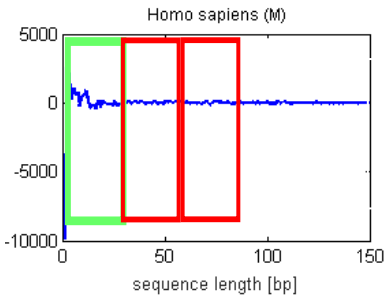
جدول ۱: مشخصات داده‌های مورد استفاده به تفکیک ارگانیسم‌ها

فرمانرو	طول کل توالی (bp)	محدوده توالی	شماره دسترسی	ارگانیسم
پرندگان	۳۸۲۸	69305341-69309168	NC_029516.1	Coturnix japonica
پرندگان	۳۹۱۸	76952888-76956805	NC_006088.4	Gallus gallus
پرندگان	۳۹۲۴	75800689-75804612	NC_015011.2	Meleagris gallopavo

Parus major	NC_031768.1	87583441-87586179	۷۲۳۹	پرندگان
Taeniopygia guttata	NC_011462.1	90242163-90245385	۳۲۲۳	پرندگان
Sesamum indicum	NC_026146.1	18171965-18174769	۲۸۰۵	گیاهان زمینی
Solanum lycopersicum	NC_015440.2	61708462-61711828	۳۳۶۷	گیاهان زمینی
Equus caballus	NC_009149.2	34074401-34078757	۴۳۵۷	پستانداران
Felis catus	NC_018729.3	41629972-41634269	۴۲۹۸	پستانداران
Homo sapiens	NC_000012.12	6534517-6538375	۳۸۵۹	پستانداران
Mus musculus	NC_000072.6	125161721-125166467	۴۷۴۷	پستانداران
Ovis aries	NC_019460.2	207594186-207598861	۴۶۷۶	پستانداران



(الف)



(ب)

شکل ۳: (الف) ضرایب DCT سیگنال‌های فاز انباره‌ای GAPDH ارگانیسم‌ها، (ب) بزرگ‌نمایی شده ضرایبDCT متعلق به GAPDH یکی از ارگانیسم‌ها [M: پستانداران، GP: گیاهان زمینی و B: پرندگان]

مدل ریاضی این تبدیل به صورت زیر است:

$$\mathbf{y}=\mathbf{\Phi x}+\mathbf{v} \qquad (۳)$$

که در آن $\mathbf{y}\in\mathbb{R}^{M\times 1}$ سیگنال فشرده شده، $\mathbf{\Phi}\in R^{M\times N}$، ماتریس حسگر، $\mathbf{x}\in\mathbb{R}^{N\times 1}$ سیگنال اصلی ($M<<N$) و $\mathbf{v}$ بردار نویز نامعلوم است. بازسازی سیگنال اولیه نیاز به حل معادله معکوس فرومعین[۱۵] با بی‌نهایت جواب برای $\mathbf{x}$ دارد. در صورت به اندازه کافی تنک بودن $\mathbf{x}$، می‌توان جوابی یکتا به‌طور دقیق یا با حداقل خطا یافت.

بازسازی سیگنال با حل مسأله بهینه‌سازی زیر قابل حصول است که $\|\mathbf{x}\|_p$ اشاره به نُرم صفر، یک یا دو دارند [۱۲]:

$$P:\ \hat{\mathbf{x}}=argmin\|\mathbf{x}\|_p \qquad (۴)$$
$$\text{subject to}\|\mathbf{y}-\mathbf{\Phi x}\|_2\le\varepsilon$$

الگوریتم‌های رایج CS، تنها در صورت به اندازه کافی تنک بودن سیگنال از کارآیی مناسب در بازسازی برخوردار هستند. بعد از توسعه الگوریتم‌های مبتنی‌بر یادگیری بیزین و ورود آن به حوزه CS، حتی در صورت غیر تنک بودن سیگنال، امکان فشرده‌سازی - بازسازی با دقت مناسب فراهم شد [۱۵-۱۳]. در این روش مسأله بازسازی تنک در CS

به‌صورت یک مسأله تخمین قابل حل در چهارچوب بیزین بازنویسی می‌شود.

۲ ۴ -چهارچوب یادگیری بیزین تنکی بلوکی[۱۶] (BLBS)

روش BSBL اولین بار توسط آقای ژانگ برای بازسازی سیگنال‌های قلبی فشرده شده با CS ارائه شد [۱۶]. این روش، به‌عنوان یک چهارچوب جدید، ساختار زمانی و دینامیکی سیگنال را برای حل مسأله CS بکار می‌گیرد. به‌عبارتی، استفاده از ساختارهای موجود در سیگنال منجر به نتایج بهتر بازسازی می‌شود. در مفهوم تنکی بلوکی، سیگنال $\mathbf{x}$ را طبق رابطه زیر بلوک‌بندی می‌کنیم:

$$\mathbf{x}=[\underbrace{x_1,\cdots,x_{d_1}}_{\mathbf{x}_1^T},\cdots,\underbrace{x_{d_{g-1}+1},\cdots,x_{d_g}}_{\mathbf{x}_g^T}]^T \quad (۵)$$

که $\mathbf{x}_i\in\mathbf{R}^{d_i\times 1}$، و $d_i(i=1,...,g)$ها، لزوماً یکسان نیستند. از میان g بلوک، تنها k بلوک غیر صفر می‌باشد که $g>>k$ است. لازم به ذکر است که در بسیاری از مواقع $\mathbf{x}$ به‌طور مستقیم تنک بلوکی نیست ولی حوزه‌ای وجود دارد که با انتقال به آن، تنک می‌شود. به‌عبارتی، $\mathbf{x}=\mathbf{\Psi\theta}$ که $\mathbf{\Psi}\in\mathbb{R}^{N\times N}$ یک ماتریس پایه اورتونرمال فضای تبدیل و $\mathbf{\theta}$ بردار ضرایب تنک است. در نتیجه مدل (۳) به‌صورت زیر نوشته می‌شود:

$$\mathbf{y}=\mathbf{\Phi\Psi\theta}+\mathbf{v}=\mathbf{\Omega\theta}+\mathbf{v} \quad (۶)$$

که $\mathbf{\Omega}=\mathbf{\Phi\psi}$. از آنجائیکه $\mathbf{\theta}$ تنک است، الگوریتم CS ابتدا $\mathbf{\theta}$ را با استفاده از $\mathbf{y}$ و $\mathbf{\Omega}$ بازسازی می‌کند، و سپس $\mathbf{x}$ از رابطه $\mathbf{x}=\mathbf{\Psi\theta}$ به‌دست می‌آید.

ما نیز در این کار با گرفتن تبدیل کسینوسی گسسته (DCT[۱۷]) از سیگنال‌های فاز انبارهای ارگانیسم‌ها آن‌ها را به‌صورت تنک بلوکی یافتیم. شکل ۴- (الف)، DCT سیگنال‌ها را نمایش می‌دهد. در شکل ۴- (ب)، در ارگانیسم Homo، کادر سبز رنگ بلوک غیرصفر و کادرهای قرمز رنگ متناظر با بلوک‌های صفر است.

در روش بازسازی BSBL، فرض بر این است که هر بلوک $\mathbf{x}_i$ یک توزیع گوسی چند متغیره دارد:

$$p(\mathbf{x}_i;\gamma_i,\mathbf{B}_i)\sim\mathcal{N}(0,\gamma_i\mathbf{B}_i),\quad i=1,...,g \quad (۷)$$

که در آن γ_i یک پارامتر غیرمنفی کنترل‌کننده تنکی بلوکی $\mathbf{x}$ است. هرگاه $\gamma_i=0$ گردد، آنگاه بلوک i ام صفر درنظر گرفته می‌شود. $\mathbf{B}_i\in\mathbb{R}^{d_i\times d_i}$ یک ماتریس مثبت معین است که ساختار همبستگی بلوک i ام را ضبط می‌کند. با فرض اینکه بلوک‌ها دو به دو ناهمبسته باشند، احتمال پیشین $\mathbf{x}$ به‌صورت $p(\mathbf{x})\sim\mathcal{N}(0,\Sigma_0)$، که Σ_0 ماتریس بلوک قطری به‌صورت زیر است:

$$\Sigma_0=\begin{bmatrix}\gamma_1\mathbf{B}_1 & & & \\ & \gamma_2\mathbf{B}_2 & & \\ & & \ddots & \\ & & & \gamma_g\mathbf{B}_g\end{bmatrix} \quad (۸)$$

همچنین فرض بر آن است که نویز دارای توزیع گوسی چند متغیره است:

$$p(\mathbf{v};\lambda)\sim\mathcal{N}(0,\lambda\mathbf{I}) \quad (۹)$$

که λ یک اسکالر مثبت است. احتمال پسین $\mathbf{x}$ به‌صورت زیر تعریف می‌شود:

$$p(\mathbf{x}|\mathbf{y};\lambda,\{\gamma_i,\mathbf{B}_i\}_{i=1}^g)\sim\mathcal{N}(\mu_x,\Sigma_x) \quad (۱۰)$$

که در این رابطه:

$$\mu_x=\Sigma_0\mathbf{\Phi}^T(\lambda\mathbf{I}+\mathbf{\Phi}\Sigma_0\mathbf{\Phi}^T)^{-1}\mathbf{y} \quad (۱۱)$$

$$\Sigma_x=(\Sigma_0^{-1}+\frac{1}{\lambda}\mathbf{\Phi}^T\mathbf{\Phi})^{-1} \quad (۱۲)$$

هنگامی که پارامترهای $\lambda,\{\gamma_i,\mathbf{B}_i\}_{i=1}^g$ تخمین زده می‌شوند، تخمین بیشینه احتمال پسین (MAP[۱۸]) از $\mathbf{x}$ که با $\hat{\mathbf{x}}$ نشان داده می‌شود، می‌تواند به‌صورت مستقیم از میانگین احتمال پسین، یعنی $\mathbf{x}=\mu_{\mathbf{x}}$، به‌دست آید [۱۶].

۳ -نتایج

برای ارزیابی عملکرد بازسازی و نیز به‌منظور مشخص نمودن بیشترین امکان کاهش نمونه، از معیار درصد جذر مربع تفاضل (PRD[۱۹]) و میانگین خطای مربع نرمالیزه‌شده (NMSE[۲۰]) بین سیگنال اصلی و سیگنال بازسازی شده استفاده نمودیم:

$$\mathrm{PRD}=\sqrt{\frac{\sum_{i=1}^{n}(x(i)-x_r(i))^2}{\sum_{i=1}^{n}(x(i)-\bar{x})^2}}\times 100 \quad (۱۳)$$

و

$$\mathrm{NMSE}=\|\mathbf{x}-\mathbf{x}_r\|_2^2/\|\mathbf{x}\|_2^2 \quad (۱۴)$$

که $x(i)$ و $x_r(i)$ به‌ترتیب، اشاره به نمونه‌های سیگنال اصلی و بازسازی شده و $\bar{x}$ اشاره به میانگین سیگنال اصلی دارد. نرخ فشرده‌سازی (CR[۲۱]) نیز به‌صورت زیر تعریف می‌شود:

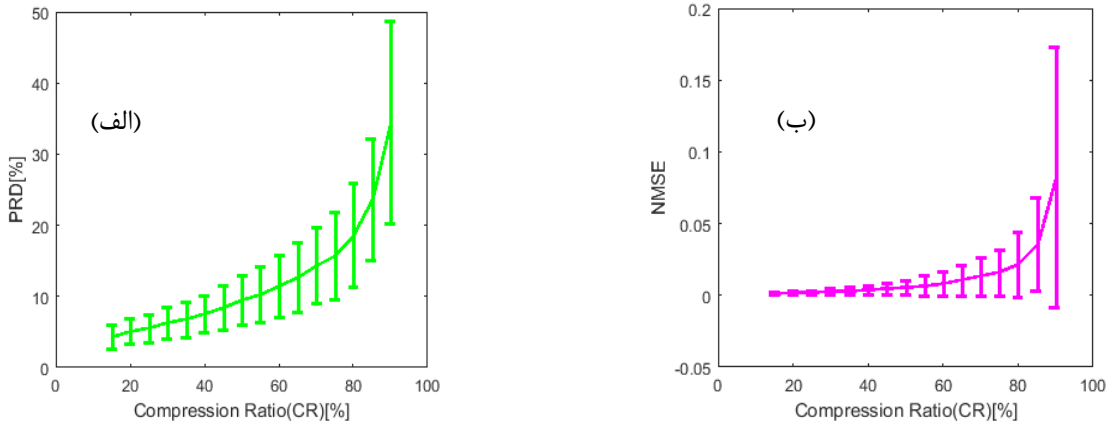
$$\mathrm{CR}=\frac{N-M}{N}\times 100 \quad (۱۵)$$

که N تعداد نمونه‌های داده اصلی و M تعداد نمونه‌های سیگنال فشرده شده است. جهت انجام شبیه‌سازی‌ها از سیستم کامپیوتری دارای ۴

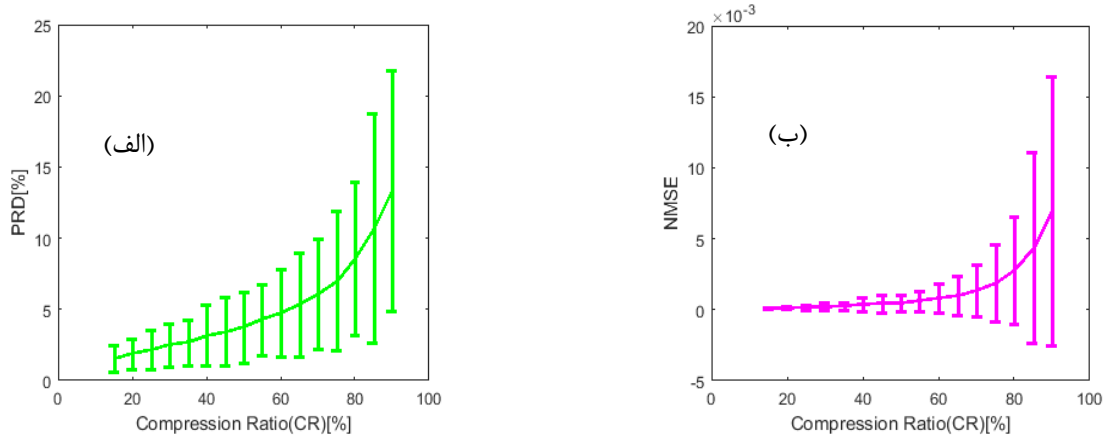
گیگابایت رم و پردازشگر دو هسته‌ای ۱/۸۷ گیگاهرتز و نرم‌افزار متلب ۲۰۱۷ استفاده شده است.

۳ ـ۱ ـ فشرده‌سازی سیگنال‌ها

شکل ۵ مقادیر متوسط PRD و NMSE تمام سیگنال‌ها را برحسب نرخ فشرده‌سازی نشان می‌دهد. عملیات فشرده‌سازی با استفاده از یک ماتریس حسگر باینری تنک انجام شده است. هر ستون ماتریس حسگر شامل ۱۵ مؤلفه ۱ در موقعیت‌های تصادفی است و باقی مولفه‌ها صفر می‌باشند [۱۷]. واضح است که با افزایش CR خطای بازسازی بیشتر می‌شود. ما CR=75% را برای ادامه کار برگزیدیم که نتایج حاصل از بازسازی این نرخ فشرده‌سازی در شکل ۷ آورده شده است. از نظر ریاضی، میانگین PRD و NMSE مربوط به CR = 75% به‌ترتیب کم‌تر از ۱۵٪ و ۰/۰۰۵ است که مناسب هستند. به‌منظور مقایسه، فشرده‌سازی سیگنال‌های ژنوم با روش DWT مطابق با روش ارائه شده در [۷] را نیز مدنظر قرار دادیم. بدین‌ترتیب که پس از انتخاب سطح تجزیه موردنظر و یک سطح آستانه مناسب، بازسازی سیگنال با ضرایب موجک بزرگتر از آستانه انتخابی انجام شد. شکل ۶ مقادیر PRD و NMSE برحسب نرخ فشرده‌سازی برای روش مبتنی‌بر DWT را نشان می‌دهد. واضح است که در این روش افزایش نرخ فشرده‌سازی متناسب با سطح تجزیه بالاتر است.



شکل ۵: ارزیابی عملکرد بازسازی با کمک BSBL بر حسب CR، (الف)معیار PRD (ب) معیار NMSE



شکل ۶: ارزیابی عملکرد بازسازی با کمک DWT بر حسب CR، (الف)معیار PRD (ب) معیار NMSE

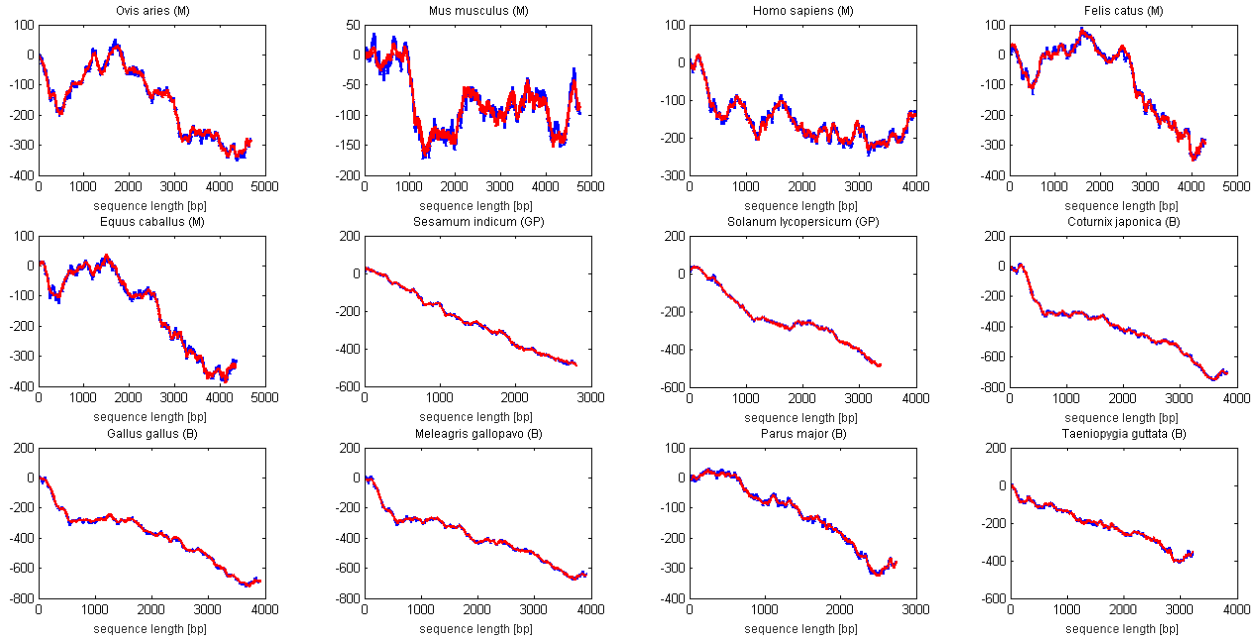
همان‌طور که از شکل مشخص است، به ازای نرخ فشرده سازی یکسان الگوریتم مبتنی‌بر DWT مقادیر PRD و NMSE کمتری نسبت به روش مبتنی‌بر CS داشته است. در جدول ۲ زمان اجرای برنامه فشرده‌سازی - بازسازی برای یک سیگنال ژنوم به‌طول ۳۹۷۱ نمونه در هر دو روش و به ازای CR برابر ۷۵٪ آورده شده است. دلیل تفاوت چشم‌گیر زمان اجرا در این است که روش مبتنی‌بر CS با تقسیم سیگنال به بلوک‌های با طول کمتر (۳۸۸ نمونه در سیگنال فوق)، هر بلوک را به تنهایی فشرده‌سازی - بازسازی کرده و با کنار هم قرار دادن آنها سیگنال اولیه را ارائه می‌دهد. در صورتیکه روش مبتنی‌بر DWT، فرآیند را به کل سیگنال اعمال می‌کند. مزیت دیگر روش پیشنهادی ما عدم وابستگی آن به طول سیگنال است. چنانچه در [۶] اشاره شده است، DWT قادر به فشرده‌سازی سیگنال‌های با طول کم نیست. زیرا طول کم سیگنال امکان تجزیه تا سطوح بالاتر و رسیدن به CR مناسب را نمی‌دهد. نتیجه اینکه روش مبتنی‌بر CS علی‌رغم PRD کمی بالاتر، به دلیل پیچیدگی‌کمتر و کاهش زمان اجرایCPU، بهترین راندمان کلی را دارا است.

جدول ۲: مقایسه زمان لازم برای فشرده‌سازی – بازسازی با دو روش مبتنی‌بر CS و DWT

	CS	DWT
نرخ فشرده‌سازی (٪)	۷۵	۷۵
زمان اجرای برنامه (ثانیه)	۰/۷	۱۶

۳ ۲ -هم‌ترازی سیگنال‌ها

مقایسه توالی‌ها در قلب آنالیزهای بیوانفورماتیک قرار دارد. لازمه این امر هم‌ترازی دو به دو دنباله‌های DNA و یا سیگنال‌های استخراج شده به روش فاز انباره‌ای است. هم‌ترازی دو به دو توالی‌ها درکی از شباهت و خویشاوندی دو توالی تحت مطالعه را فراهم می‌کند و نیز پایه جستجوی تشابهات در پایگاه داده‌ها و هم‌ترازی‌های چندگانه است [۱۸]. ما از الگوریتم DTW که الگوریتمی برای هم‌ترازی سیگنال‌ها است و به‌مثابه هم‌ترازی کلی عمل می‌کند استفاده کردیم. الگوریتم فوق به سیگنال‌های فشرده‌شده مرحله قبل اعمال شد.



شکل ۷: سیگنال‌های اصلی (آبی) و بازسازی شده از روی فشرده‌های با CR = 75% (قرمز) متعلق به GAPDH ارگانیسم‌ها M: پستانداران، GP: گیاهان زمینی، B: پرندگان

از آن‌جایی که جفت سیگنال‌ها بعد از هم‌ترازی کلی، دارای طول یکسان n می‌شوند، فاصله آن‌ها از هم به‌کمک رابطه زیر محاسبه می‌شود:

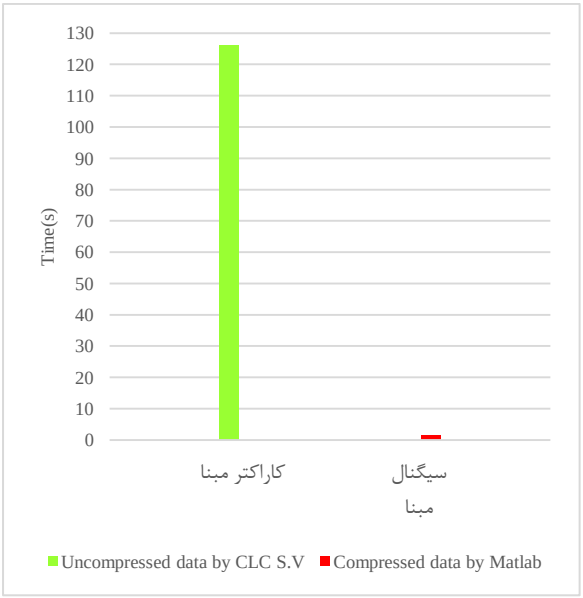
$$d = \sqrt{\sum_{i=1}^{n} [x(i) - y(i)]^2} \qquad (۱۶)$$

که $x(n)$ و $y(n)$ سیگنال‌های فشرده شده هم‌تراز شده هستند و d فاصله دو سیگنال از هم است [۷]. نتایج حاصل از هم‌ترازی سیگنال‌ها به صورت ماتریسی تحت عنوان ماتریس فاصله ثبت می‌گردد. با کمک این ماتریس، امکان تحلیل خوشه‌ای دنباله‌ها فراهم می‌شود که در بخش بعدی مفصل‌تر توضیح داده خواهد شد.

جدول ۳ ماتریس فاصله نرمالیزه شده ارگانیسم‌ها به روش سیگنال مبنا و با استفاده از رابطه ۱۵ را نشان می‌دهد. همچنین جهت مقایسه، ماتریس فاصله بین دنباله‌های کاراکتری محاسبه شده با استفاده از نرم افزار CLC نیز در جدول ۴ آورده شده است.

در هر دوماتریس فاصله مذکور، عدد صفر بیان‌گر کم‌ترین فاصله و عدد یک بیان‌گر بیشترین فاصله دو سیگنال یا دنباله کاراکتری از هم است. بدیهی است که فاصله یک سیگنال (ژن) از خودش، صفر است. این ماتریس‌ها در مرحله بعد به‌منظور رسم درختچه‌های فیلوژنتیکی مورد استفاده قرار خواهند گرفت. زمان لازم برای هم‌ترازی سیگنال‌های فشرده شده و نیز دنباله‌های کاراکتری در شکل ۸ آورده شده است. همان‌طور که از شکل مشاهده می‌شود زمان صرف شده برای هم‌ترازی به روش سیگنالی و از روی سیگنال‌های فشرده‌شده

۱/۲۸۵۳ ثانیه، در مقایسه با زمان لازم برای هم‌ترازی دنباله‌های کاراکتری با نرم افزار CLC با زمان ۱۲۶ ثانیه بوده است.



شکل ۸: زمان صرف شده برای هم‌ترازی GAPDH ارگانیسم‌ها به تفکیک روش‌ها و برنامه‌ها [محور عمودی زمان برحسب ثانیه و محور افقی روش‌ها و برنامه‌ها]

جدول۳: فاصله اقلیدسی سیگنال‌های فشرده GAPDH ارگانیسم‌ها [پستانداران (ارغوانی)، گیاهان زمینی (سبز)، پرندگان (آبی)]

	Ovis	Mus	Homo	Felis	Equus	Sesamum	Solanum	Coturnix	Gallus	Meleagris	Parus	Taeniopygia
Ovis	۰											
Mus	۰/۳۸	۰										
Homo	۰/۴۶	۰/۳۷	۰									
Felis	۰/۴۶	۰/۳۲	۰/۳۸	۰								
Equus	۰/۵۶	۰/۴۲	۰/۴۶	۰/۳۵	۰							
Sesamum	۰/۸۳	۰/۷۳	۰/۶۱	۰/۷۲	۰/۷۲	۰						
Solanum	۰/۶۳	۰/۵۵	۰/۵۰	۰/۵۶	۰/۶۰	۰/۶۹	۰					
Coturnix	۰/۹۰	۰/۸۵	۰/۷۶	۰/۸۷	۰/۹۲	۰/۹۶	۱	۰				
Gallus	۰/۸۲	۰/۸۰	۰/۶۸	۰/۸۳	۰/۸۷	۰/۹۶	۰/۸۸	۰/۷۴	۰			
Meleagris	۰/۹۲	۰/۸۱	۰/۷۹	۰/۸۵	۰/۸۸	۰/۹۸	۰/۸۷	۰/۹۵	۰/۸۰	۰		
Parus	۰/۶۰	۰/۴۶	۰/۴۲	۰/۴۶	۰/۵۴	۰/۴۱	۰/۵۰	۰/۸۹	۰/۸۱	۰/۸۵	۰	
Taeniopygia	۰/۶۵	۰/۵۰	۰/۵۰	۰/۵۲	۰/۶۰	۰/۵۲	۰/۵۹	۰/۹۱	۰/۸۷	۰/۸۲	۰/۴۱	۰

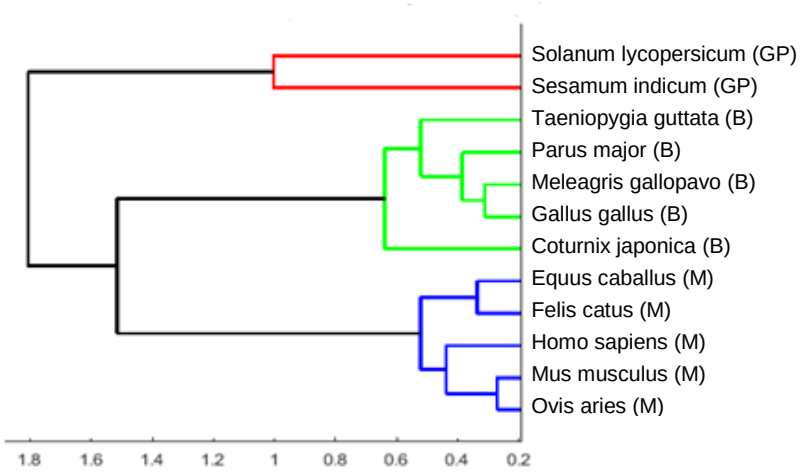
جدول ۴: فاصله کاراکتری GAPDH ارگانیسم‌ها با نرم‌افزار CLC

	Felis	Equus	Ovis	Homo	Mus	Gallus	Meleagris	Coturnix	Parus	Taeniopygia	Sesamum	Solanum
Felis	۰											
Equus	۰/۲۰	۰										
Ovis	۰/۲۷	۰/۲۷	۰									
Homo	۰/۲۸	۰/۲۹	۰/۳۳	۰								
Mus	۰/۳۴	۰/۳۴	۰/۴۳	۰/۳۸	۰							
Gallus	۰/۶۶	۰/۶۷	۰/۷۰	۰/۶۴	۰/۶۹	۰						
Meleagris	۰/۶۸	۰/۶۹	۰/۷۱	۰/۶۵	۰/۷۱	۰/۰۹	۰					
Coturnix	۰/۶۹	۰/۷۱	۰/۷۲	۰/۶۸	۰/۷۴	۰/۱۲	۰/۱۳	۰				
Parus	۰/۵۵	۰/۵۸	۰/۵۷	۰/۵۲	۰/۵۳	۰/۲۹	۰/۲۹	۰/۳۰	۰			
Taeniopygia	۰/۶۲	۰/۶۳	۰/۶۴	۰/۵۹	۰/۶۱	۰/۳۰	۰/۳۰	۰/۳۲	۰/۰۸	۰		
Sesamum	۱/۱۱	۱/۱۲	۱/۱۶	۱/۰۷	۱/۱۲	۱/۰۲	۰/۹۸	۱/۰۲	۰/۹۸	۱/۰۷	۰	
Solanum	۱/۱۲	۱/۱۲	۱/۲۳	۱/۱۷	۱/۱۵	۱/۱۴	۱/۰۹	۱/۱۲	۱/۱۶	۱/۱۷	۰/۴۷	۰

۳ ۳ -خوشه‌بندی ارگانیسم‌ها

فیلوژنتیک، مطالعه تاریخ تکاملی ارگانیسم‌های زنده با استفاده از دیاگرام‌های مشابه درخت است تا شجره‌نامه این ارگانیسم‌ها نمایش داده شود. الگوهای شاخه‌ای درخت که نشان‌دهنده انشعابات تکاملی است، فیلوژنی نامیده می‌شود. خوشبختانه اطلاعات مولکولی به صورت توالی‌های DNA یا پروتئین می‌توانند جنبه‌های تکاملی مفیدی از ارگانیسم‌های موجود فراهم کنند. زیرا وقتی ارگانیسم نمو می‌کند مواد ژنتیکی جهش‌هایی را انباشته می‌کند که در طول زمان موجب تغییرات فتوتیپی می‌شود [۱۸].

جهت بررسی امکان به‌کارگیری سیگنال‌های فشرده شده برای مقایسه ارگانیسم‌ها و تحلیل خوشه‌ای آنها، درختچه فیلوژنتیک با استفاده از ماتریس فاصله مرحله قبل رسم شد. بدین منظور، روش جفت گروه بی‌وزن با میانگین ریاضی(UPGMA[۲۲]) برای دستیابی به بهترین فاصله خوشه‌ای استفاده شده است. در شکل ۹ خوشه‌ها به‌طور مناسبی از هم جدا شده‌اند و ارگانیسم‌های مربوط به یک فرمانرو در کنار هم قرار گرفته‌اند. لذا، روش پیشنهادی ما توانسته ارگانیسم‌های مختلف را با استفاده از یک ژن خانه‌دار به‌طور مناسبی خوشه بندی نماید. همان‌طور که مشاهده می‌شود، ارگانیسم‌های هر فرمانرو در مجاورت هم قرار گرفته‌اند. به‌علاوه، فرمانرو پستانداران و پرندگان دارای ریشه مشترک هستند.



شکل ۹: درختچه فیلوژنتیکی حاصل از سیگنال‌های فشرده شده GAPDH ارگانیسم‌ها با CR=75% به وسیله نرم‌افزار متلب، [M: پستانداران، GP: گیاهان زمینی، B: پرندگان]

۴ -نتیجه

جوامع علمی و علی‌الخصوص غربی سرمایه‌گذاری‌ها و پژوهش‌های زیادی بر روی ژنوم، رمزگشایی و تحلیل آن به‌منظور استفاده در کارکردهای مختلف پزشکی، زیستی، نظامی و... انجام می‌دهند. تاکنون اکثر این مطالعات با روش‌های کاراکتر مبنا بوده است. به نظر می‌آید با وجود طیف عظیمی از الگوریتم‌های متنوع پردازش سیگنال و مزایایی که برای بیان سیگنالی ژن ذکر گردید، می‌توان از ظرفیت‌های این حوزه برای مقابله با چالش‌های موجود در مطالعات ژنومی استفاده کرد. با توجه به نتایج، جای امید است که پردازش سیگنالی ژن موجب شناخت بیشتر دنیای پر رمز و راز ژنوم گردد.

همانطور که از نتایج مشاهده می‌گردد تحلیل خوشه‌ای به روش سیگنالی و از روی سیگنال‌های فشرده‌شده با CR = 75% هم‌پای تحلیل خوشه‌ای به روش کاراکتری عمل کرده است. در حالی که زمان خیلی کم‌تری را صرف کرده است، و دوماً با بیان سیگنالی، می‌توان ویژگی‌های توالی‌ها را در مقیاس بزرگ دید. سوماً با فشرده‌سازی هزینه پردازش، ذخیره‌سازی و انتقال کاهش می‌یابد.

مراجع

[1] B. S. Jeong, A. T. M. G. Bari, M. R. Reaz, S. Jeon, C. G. Lim, and H. J. Choi, "Codon-based encoding for DNA sequence analysis," Methods, vol. 67, no. 3, pp. 373–379, 2014.

[2] P. Hanus, J. Dingel, G. Chalkidis, and J. Hagenauer, "Compression of whole genome alignments," IEEE Trans. Inf. Theory, vol. 56, no. 2, pp. 696–705, 2010.

[3] GenBank Growth Statistics, http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

[4] S.Wandelt, M. Bux, and U. Leser, "Trends in genome compression," Curr. Bioinform., vol. 9, no. 3, pp. 315–326, 2014.

[5] S. Arniker and H. K. Kwan, "Numerical representation of DNA sequences," IEEE Int. Con. on Electrro/Information Technology, pp. 307–310, 2009.

[6] K. Sedlar, H. Skutkova, M. Vitek, and I. Provaznik, "Set of rules for genomic signal downsampling," Comput. Biol. Med., vol. 69, pp. 308–314, 2016.

[7] K. Sedlar, H. Skutkova, M. Vitek, and I. Provaznik, "Prokaryotic DNA signal downsampling for fast whole genome comparison," Inf. Technol. Biomed., vol. 3, pp. 373–383, 2014.

[8] http://www.ncbi.nlm.nih.gov/genbank/

[9] G. Mendizabal-Ruiz, I. Román-Godínez, S. Torres-Ramos, R. A. Salido-Ruiz, and J. A. Morales, "On DNA numerical representations for genomic similarity computation," PLoS One, vol. 12, no. 3, pp. 1–27, 2017.

[10] T. Hoang, Ch. Yin, and S. Yau, "Numerical encoding of DNA sequences by chaos game representation with application in similarity comparision," Genomics, vol. 108, no. 3-4, pp. 134–142, 2016.

[11] D. Anastassiou, "Genomic signal processing," Signal Processing Magazine, vol. 18, pp. 8–20, 2001.

[۱۲] هادی شکری و محمدحسین کهایی، «حسگری فشرده تصاویر ابرطیفی با دسته‌بندی طیفی و بازسازی با تنظیم‌کننده تغییرات کلی طیفی-مکانی»، مجله مهندسی برق دانشگاه تبریز، دوره ۴۷، شماره ۴، صفحه ۱۵۱۳-۱۵۲۱، زمستان ۱۳۹۶.

[13] Z. Zhang and B. D. Rao, "Extension of SBL algorithms for the recovery of block sparse signals with intra-block correlation," IEEE Trans. Signal Process., vol. 61, no. 8, pp. 2009–2015, 2013.

[14] Z. Zhilin and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning," IEEE J. Selected Topics in Signal Process., vol. 5, no. 5, pp. 912–926, 2011.

[15] Z. Zhang, *Sparse signal recovery exploiting spatiotemporal correlation*, Ph.D. Thesis, University of California, San Diego, 2012.

[16] Z. Zhang, T. Jung and S. Makeig, "Compressed sensing for energy-efficient wireless telemonitoring of noninvasive fetal ECG via block sparse bayesian learning," IEEE Trans. on Biomedical Engineering, vol. 60, no. 2, pp. 300–309, 2013.

[۱۷] محمدمهدی محدث و محمدحسین کهایی، «ساخت ماتریس‌های نمونه‌برداری یقینی براساس توابع هش»، مجله مهندسی برق دانشگاه تبریز، دوره ۴۶، شماره ۴، صفحه ۳۰۷-۳۱۵، زمستان ۱۳۹۵.

[۱۸] رمضانی, مجیدی, مهدی‌زاده, ملکی و پوربرخورداری, *اصول بیوانفورماتیک*، زنجان، جامعه نگر, ۱۳۹۳.

زیرنویس‌ها

[1] Deoxyribonucleic acid
[2] Sequencing
[3] Alignment
[4] Genomic signal processing
[5] Discrete wavelet transform
[6] Compressed sensing
[7] CLC sequence viewer
[8] Glyceraldehyde 3-phosphate dehydrogenase
[9] https://www.ncbi.nlm.nih.gov/gene/?term=GAPDH
[10] Housekeeping gene
[11] Master gene``
[12] Cumulated phase
[13] Adenine, Cytosine, Guanine, Thymine
[14] Sensing matrix
[15] Underdetermined problem
[16] Block sparse bayesian learning
[17] Discrete cosine transform
[18] Maximum a posterior
[19] Percentage root mean square difference
[20] Normalized mean square error
[21] Compression ratio
[22] Unweighted pair group method with arithmetic mean