

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

محمود طوماری<sup>۱</sup>، کارشناس ارشد مهندسی پزشکی؛ سپیده جباری<sup>۲</sup>، 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 به شدت کاهش یافته است و مقدار زیادی از داده‌های ژنومی در حال تولید هستند. از طرفی، هزینه حافظه ذخیره‌سازی، پردازش و انتقال این داده‌ها در حال افزایش است. پردازش این حجم عظیم اطلاعات بیشتر توسط روش‌های کاراکتر مبنا صورت می‌گیرد که زمان‌بر است. ظرفیت‌های بالقوه 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 به صورت سیگنالی بیان شدند. سپس، سیگنال‌های ژنومی حاصل توسط روش حسگری فشرده مبتنی بر یادگیری بیزین فشرده‌سازی شدند. توانایی روش در بازسازی سیگنال‌های فشرده شده با استفاده از معیارهای PRD و NMSE مورد بررسی قرار گرفت. سپس به منظور مقایسه و بررسی مشابهت دنباله‌ها، درختچه فیلوژنتیک از روی سیگنال‌های فشرده شده با نرخ ۷۵٪ رسم شد. نتایج نشان دادند که 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<sup>1</sup>, MSc in Biomedical Engineering; Sepideh Jabbari<sup>2</sup>, 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 ۴۵۳۷۱-۳۸۷۹۱

## ۱- مقدمه

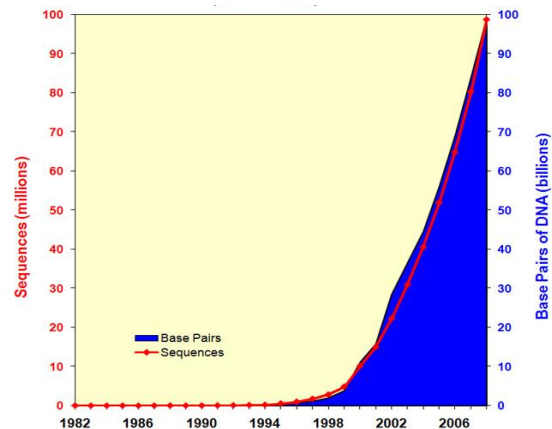
اشاره به استفاده از روش‌های پردازش سیگنال برای تحلیل داده‌های ژنومی دارد. با بیان سیگنالی، ویژگی‌های توالی‌ها در مقیاس بزرگتری دیده می‌شود و اطلاعات برای چشم انسان قابل رؤیت است. بعلاوه اینکه، نتایج تحقیقات اخیر کارآیی روش را در هم‌ترازی دنباله‌ها و مقایسه ارگانسیم‌ها با یکدیگر تأیید می‌کنند. نتیجه اینکه 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 می‌طلبد. 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 ژنوم حاصل از دنباله‌های کاراکتری برای اولین بار توسط سدلار و همکارانش مطرح شد [۷]. آنها از تبدیل موجک گسسته (DWT<sup>۵</sup>) 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<sup>۶</sup>) 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 ارگانسیم‌ها را با بکارگیری این سیگنال‌های فشرده شده مدنظر قرار داده‌ایم. به‌منظور مقایسه، عمل خوشه‌بندی را با استفاده از دنباله‌های کاراکتری DNA ارگانسیم‌ها و با کمک نرم‌افزار CLC S.V.<sup>۷</sup> نیز انجام داده‌ایم. بلوک‌دیگرام شکل ۲ روند روش پیشنهادی در کنار روش متداول کاراکتر مینا را نشان می‌دهد.

دئوکسی ریبونوکلیک اسید (DNA<sup>۱</sup>) یک اسید نوکلئیک شامل دستورات عمل‌های ژنتیکی مورد استفاده در رشد و عملکرد همه ارگانسیم‌های 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<sup>۲</sup> 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]

## ۱-۴ 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 ارگانسیم‌ها مدنظر قرار گرفت [۵]. امروزه، مقایسه دنباله‌های ژنومی با روش‌های مبتنی بر کاراکتر صورت می‌گیرد. هم‌ترازی<sup>۳</sup> 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 ناراحت‌کننده است.

یک روش جایگزین جهت تحلیل داده‌ها، پردازش سیگنال ژنوم حاصل از دنباله‌های کاراکتری است. پردازش سیگنال ژنوم (GSP<sup>۴</sup>)

در بیان فاز انباره ای، هر کدام از چهار نوکلئوتید تشکیل دهنده  $\varphi$  توالی DNA شامل آدنین، سیتوزین، تیمین و گوانین  $(A, C, T, G)$  به یک عدد مختلط در صفحه مختلط به فرم  $A \rightarrow [1, j], C \rightarrow [-1, -j], G \rightarrow [-1, j], T \rightarrow [1, -j]$  سیگنال فاز انباره ای از جمع نمودن فاز این اعداد مختلط در طول نمونه ها مطابق با معادلات زیر ایجاد می گردد:

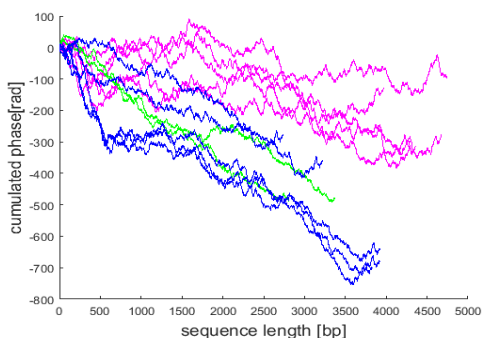
$$\{\varphi_A, \varphi_C, \varphi_G, \varphi_T\} = \left\{ \frac{\pi}{4}, -\frac{3\pi}{4}, \frac{3\pi}{4}, -\frac{\pi}{4} \right\} \quad (1)$$

$$\theta_{cum} = \frac{\pi}{4} [3(n_G - n_C) + (n_A - n_T)] \quad (2)$$

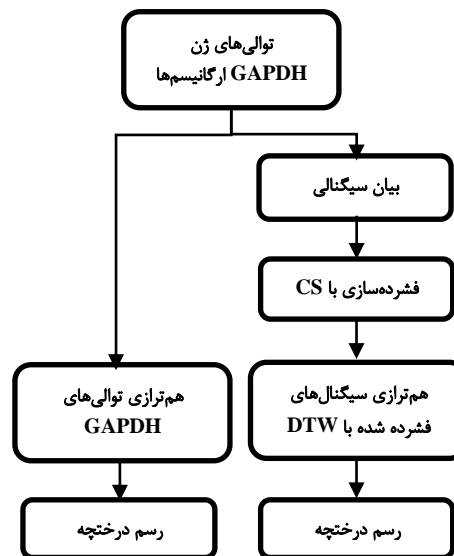
که  $\varphi_X$  فاز مربوط به نوکلئوتید  $X$ ،  $n_X$  تعداد نوکلئوتیدهای  $X$  از ابتدا تا موقعیت کنونی در توالی و  $\theta_{cum}$  سیگنال فاز انباره ای حاصل است [۶]. شکل ۳ سیگنال های متناظر با دوازده دنباله کاراکتری اشاره شده در جدول ۱ را نمایش می دهد.

### ۴ ۲ فشرده سازی با روش CS

حسگری فشرده یک چارچوب جدید جهت نمونه برداری و فشرده سازی سیگنال به صورت هم زمان معرفی می نماید. شرط لازم جهت بازسازی مؤثر سیگنال اصلی از روی سیگنال فشرده، تنگی یا به عبارتی تعداد کم مؤلفه های غیر صفر سیگنال است [۱۲]. در CS با استفاده از ماتریسی که از آن به عنوان ماتریس حسگری  $\Phi$  یاد می شود، یک تبدیل خطی بر روی سیگنال اعمال کرده و بدین ترتیب آن را به منظور کاهش در مصرف انرژی و حجم ذخیره سازی و پهنای باند ارسالی، قبل از ارسال و در فرستنده فشرده می نمایند. سپس، به کمک همین ماتریس و با فرض تنگی سیگنال، سیگنال در گیرنده و از روی نمونه های دریافتی بازسازی می شود.



شکل ۲: سیگنال های فاز انباره ای متعلق به ژن های GAPDH [استانداران (ارغوانی)، گیاهان زمینی (سبز)، پرندگان (آبی)]



شکل ۲: بلوک دیاگرام روش پیشنهادی

### ۲ مواد و روش ها

#### ۲ ۴ داده گان مورد استفاده

جهت بررسی قابلیت فشرده سازی سیگنال های ژنومی، ما از دنباله های کاراکتری مربوط به دوازده ژن گلیسرالدهید ۳- فسفات دهیدروژناز ( $GAPDH^3$ ) متعلق به ارگانسیم های مختلف استفاده نمودیم. دنباله های فوق از پایگاه داده GenBank در سایت NCBI<sup>۱</sup> در دسترس هستند [۸]. توضیحات مربوط به دنباله ها در جدول ۱ آورده شده است. این ژن به عنوان یک ژن خانه دار<sup>۱۱</sup> در سلول های تمام ارگانسیم ها حضور داشته و عاملی مهم در حفظ و نگهداری عملکردهای اصلی سلول است. دلیل انتخاب یک ژن خانه دار، بیان ثابت آن در تمام یا اغلب سلول ها جهت واکنش های متابولیک سلول است. چنانچه الگوریتم در خوشه بندی ارگانسیم ها با استفاده از این ژن موفق عمل کند، جداسازی ارگانسیم ها مبتنی بر ژن های فرمانروا<sup>۱۲</sup> که دارای بیان افتراقی در سلول های مختلف هستند به راحتی امکان پذیر خواهد بود.

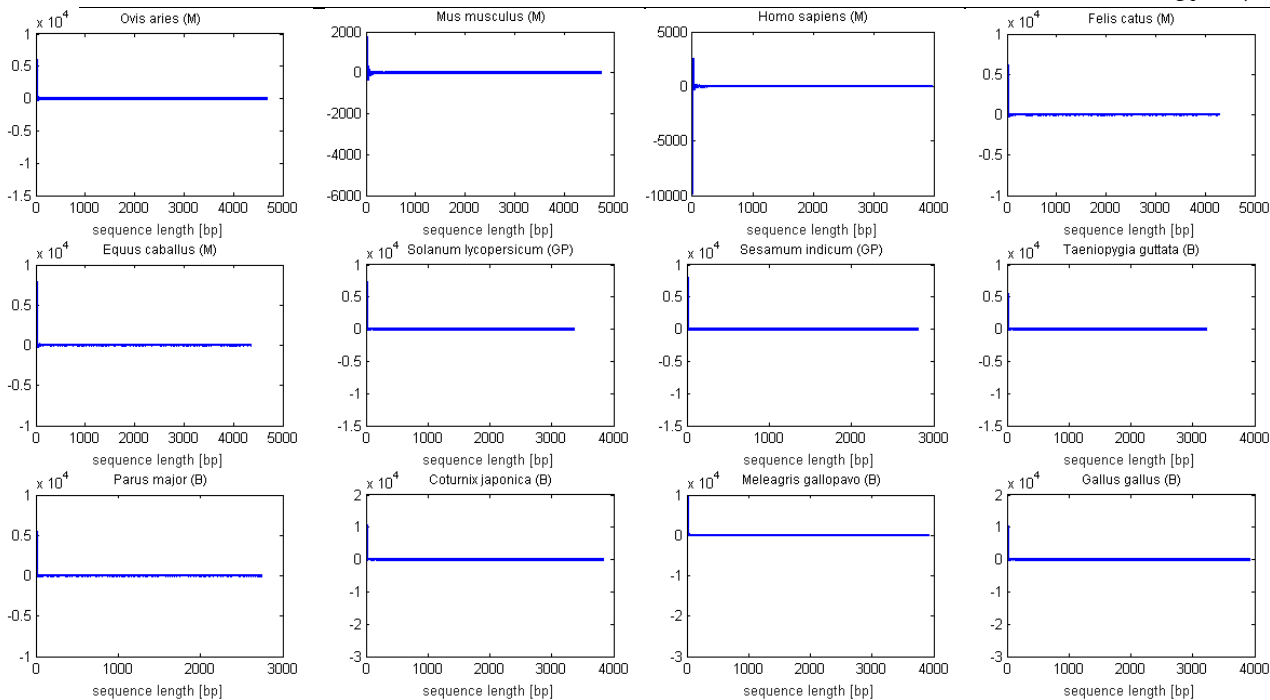
#### ۲ ۴ استخراج سیگنال ژنوم از دنباله های کاراکتری

تاکنون روش های مختلفی جهت بیان سیگنالی توالی های کاراکتری DNA و پروتئین ها ارائه شده است [۹-۱۱]. از جمله این روش ها، که تا حد امکان خصوصیات بیولوژیکی دنباله کاراکتری مورد بررسی را حفظ می کند، روش نمایش فاز انباره ای<sup>۱۲</sup> است. مزیت اصلی این روش یک بُعدی بودن سیگنال حاصل و مناسب بودنش برای فرآیند فشرده سازی و نیز هم ترازای است [۶، ۷].

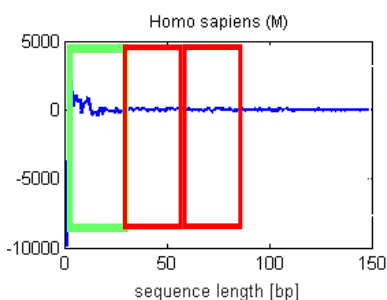
جدول ۱: مشخصات داده های مورد استفاده به تفکیک ارگانسیم ها

فرمانرو	طول کل توالی (bp)	محدوده توالی	شماره دسترسی	ارگانسیم
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	پرنده‌گان
Taeniopygia guttata	NC_011462.1	90242163-90245385	۳۲۲۳	پرنده‌گان
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 ارگانیزم‌ها، (ب) بزرگ‌نمایی شده ضرایب DCT متعلق به GAPDH یکی از ارگانیزم‌ها [M]: پستانداران، GP: گیاهان زمینی و B: پرنده‌گان

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

$$P: \hat{x} = \operatorname{argmin} \|x\|_p \quad (۴)$$

$$\text{subject to } \|y - \Phi x\|_2 \leq \varepsilon$$

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

مدل ریاضی این تبدیل به صورت زیر است:

$$y = \Phi x + v \quad (۳)$$

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

$$\Sigma_0 \square \begin{bmatrix} \gamma_1 \mathbf{B}_1 & & & \\ & \gamma_2 \mathbf{B}_2 & & \\ & & \ddots & \\ & & & \gamma_g \mathbf{B}_g \end{bmatrix} \quad (8)$$

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

$$p(\mathbf{v}; \lambda) \square \mathbf{N}(0, \lambda \mathbf{I}) \quad (9)$$

که  $\lambda$  یک اسکالر مثبت است. احتمال پسین  $\mathbf{x}$  به صورت زیر تعریف می شود:

$$p(\mathbf{x} | \mathbf{y}; \lambda, \{\gamma_i, \mathbf{B}_i\}_{i=1}^g) \square \mathbf{N}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}) \quad (10)$$

که در این رابطه:

$$\mu_{\mathbf{x}} = \Sigma_0 \Phi^T (\lambda \mathbf{I} + \Phi \Sigma_0 \Phi^T)^{-1} \mathbf{y} \quad (11)$$

$$\Sigma_{\mathbf{x}} = (\Sigma_0^{-1} + \frac{1}{\lambda} \Phi^T \Phi)^{-1} \quad (12)$$

هنگامی که پارامترهای  $\lambda, \{\gamma_i, \mathbf{B}_i\}_{i=1}^g$  تخمین زده می شوند، تخمین بیشینه احتمال پسین (MAP<sup>۱۸</sup>) از  $\mathbf{x}$  که با  $\hat{\mathbf{x}}$  نشان داده می شود، می تواند به صورت مستقیم از میانگین احتمال پسین، یعنی  $\mathbf{x} = \mu_{\mathbf{x}}$  به دست آید [۱۶].

### ۳- نتایج

برای ارزیابی عملکرد بازسازی و نیز به منظور مشخص نمودن بیشترین امکان کاهش نمونه، از معیار درصد جذر مربع تفاضل (PRD<sup>۱۹</sup>) و میانگین خطای مربع نرمالیزه شده (NMSE<sup>۲۰</sup>) بین سیگنال اصلی و سیگنال بازسازی شده استفاده نمودیم:

$$PRD = \sqrt{\frac{\sum_{i=1}^n (x(i) - x_r(i))^2}{\sum_{i=1}^n (x(i) - \bar{x})^2}} \times 100 \quad (13)$$

و

$$NMSE = \frac{\|\mathbf{x} - \mathbf{x}_r\|_2^2}{\|\mathbf{x}\|_2^2} \quad (14)$$

که  $x(i)$  و  $x_r(i)$  به ترتیب، اشاره به نمونه های سیگنال اصلی و بازسازی شده و  $\bar{x}$  اشاره به میانگین سیگنال اصلی دارد. نرخ فشرده سازی (CR<sup>۲۱</sup>) نیز به صورت زیر تعریف می شود:

$$CR = \frac{N - M}{N} \times 100 \quad (15)$$

که  $N$  تعداد نمونه های داده اصلی و  $M$  تعداد نمونه های سیگنال فشرده شده است. جهت انجام شبیه سازی ها از سیستم کامپیوتری دارای ۴

به صورت یک مسأله تخمین قابل حل در چهارچوب بیزین بازنویسی می شود.

### ۴-۲ چهارچوب یادگیری بیزین تنکی بلوکی<sup>۱۶</sup> (BLBS)

روش BSBL اولین بار توسط آقای ژانگ برای بازسازی سیگنال های قلبی فشرده شده با CS ارائه شد [۱۶]. این روش، به عنوان یک چهارچوب جدید، ساختار زمانی و دینامیکی سیگنال را برای حل مسأله CS بکار می گیرد. به عبارتی، استفاده از ساختارهای موجود در سیگنال منجر به نتایج بهتر بازسازی می شود. در مفهوم تنکی بلوکی، سیگنال  $\mathbf{x}$  را طبق رابطه زیر بلوک بندی می کنیم:

$$\mathbf{x} = [\underbrace{x_1, \dots, x_{d_1}}_{\mathbf{x}_1^T}, \dots, \underbrace{x_{d_{g-1}+1}, \dots, x_{d_g}}_{\mathbf{x}_g^T}]^T \quad (5)$$

که  $\mathbf{x}_i \in \mathbf{R}^{d_i \times 1}$  و  $d_i (i=1, \dots, g)$ ، لزوماً یکسان نیستند. از میان  $g$  بلوک، تنها  $k$  بلوک غیر صفر می باشد که  $g \gg k$  است. لازم به ذکر است که در بسیاری از مواقع  $\mathbf{x}$  به طور مستقیم تنک بلوکی نیست ولی حوزه های وجود دارد که با انتقال به آن، تنک می شود. به عبارتی،  $\mathbf{x} = \Psi \boldsymbol{\theta}$  که  $\Psi \in \mathbf{R}^{N \times N}$  یک ماتریس پایه اورتونرمال فضای تبدیل و  $\boldsymbol{\theta}$  بردار ضرایب تنک است. در نتیجه مدل (۳) به صورت زیر نوشته می شود:

$$\mathbf{y} = \Phi \Psi \boldsymbol{\theta} + \mathbf{v} = \Omega \boldsymbol{\theta} + \mathbf{v} \quad (6)$$

که  $\Omega = \Phi \Psi$ . از آنجائیکه  $\boldsymbol{\theta}$  تنک است، الگوریتم CS ابتدا  $\boldsymbol{\theta}$  را با استفاده از  $\mathbf{y}$  و  $\Omega$  بازسازی می کند، و سپس  $\mathbf{x}$  از رابطه  $\mathbf{x} = \Psi \boldsymbol{\theta}$  به دست می آید.

ما نیز در این کار با گرفتن تبدیل کسینوسی گسسته (DCT<sup>۲۲</sup>) از سیگنال های فاز انباره ای ارگانسیم ها آن ها را به صورت تنک بلوکی یافتیم. شکل ۴- الف)، DCT سیگنال ها را نمایش می دهد. در شکل ۴- ب)، در ارگانسیم Homo، کادر سبز رنگ بلوک غیر صفر و کادرهای قرمز رنگ متناظر با بلوک های صفر است.

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

$$p(\mathbf{x}_i; \gamma_i, \mathbf{B}_i) \square \mathbf{N}(0, \gamma_i \mathbf{B}_i), \quad i=1, \dots, g \quad (7)$$

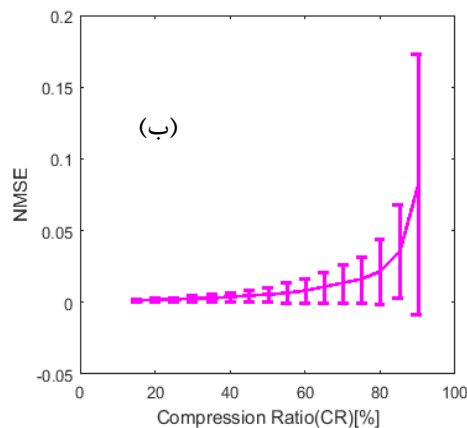
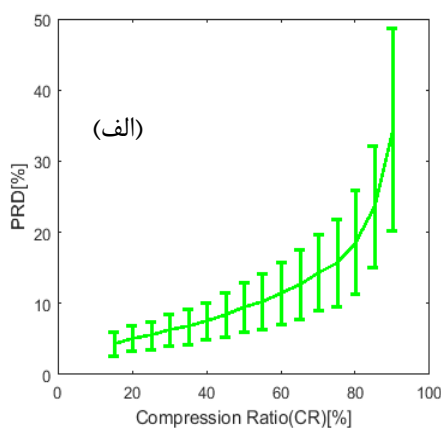
که در آن  $\gamma_i$  یک پارامتر غیر منفی کنترل کننده تنکی بلوکی  $\mathbf{x}$  است. هرگاه  $\gamma_i = 0$  گردد، آنگاه بلوک  $i$  ام صفر در نظر گرفته می شود.  $\mathbf{B}_i \in \mathbf{R}^{d_i \times d_i}$  یک ماتریس مثبت معین است که ساختار همبستگی بلوک  $i$  ام را ضبط می کند. با فرض اینکه بلوک ها دو به دو ناهمبسته باشند، احتمال پیشین  $\mathbf{x}$  به صورت  $p(\mathbf{x}) \sim \mathbf{N}(0, \Sigma_0)$ ، که  $\Sigma_0$  ماتریس بلوک قطری به صورت زیر است:

ریاضی، میانگین PRD و NMSE مربوط به  $CR = 75\%$  به ترتیب 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 مبتنی بر DWT را نشان می‌دهد. واضح است که در این روش افزایش نرخ فشرده‌سازی متناسب با سطح تجزیه بالاتر است.

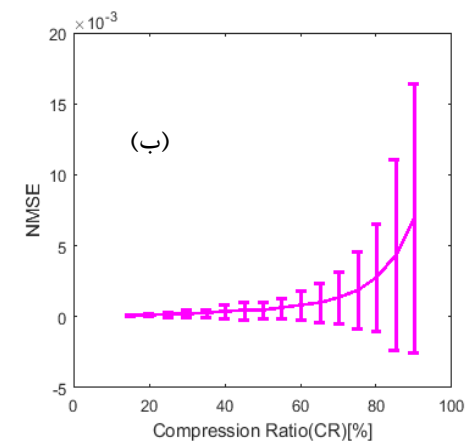
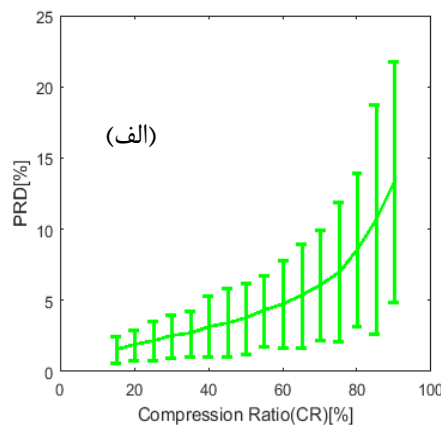
گیگابایت رم و پردازشگر دو هسته‌ای ۱/۸۷ گیگاهرتز و نرم‌افزار متلب ۲۰۱۷ استفاده شده است.

### ۳-۴ 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 مؤلفه‌ها 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



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

آنها سیگنال اولیه را ارائه می‌دهد. در صورتیکه روش مبتنی بر 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 مبتنی بر CS علی‌رغم PRD کمی بالاتر، به دلیل پیچیدگی کمتر و کاهش زمان اجرای CPU، بهترین راندمان کلی را دارا است.

همان‌طور که از شکل مشخص است، به ازای نرخ فشرده‌سازی یکسان الگوریتم مبتنی بر DWT مقادیر PRD و NMSE کمتری نسبت به روش مبتنی بر CS داشته است. در جدول ۲ زمان اجرای برنامه فشرده‌سازی - بازسازی برای یک سیگنال ژنوم به طول 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 مبتنی بر 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

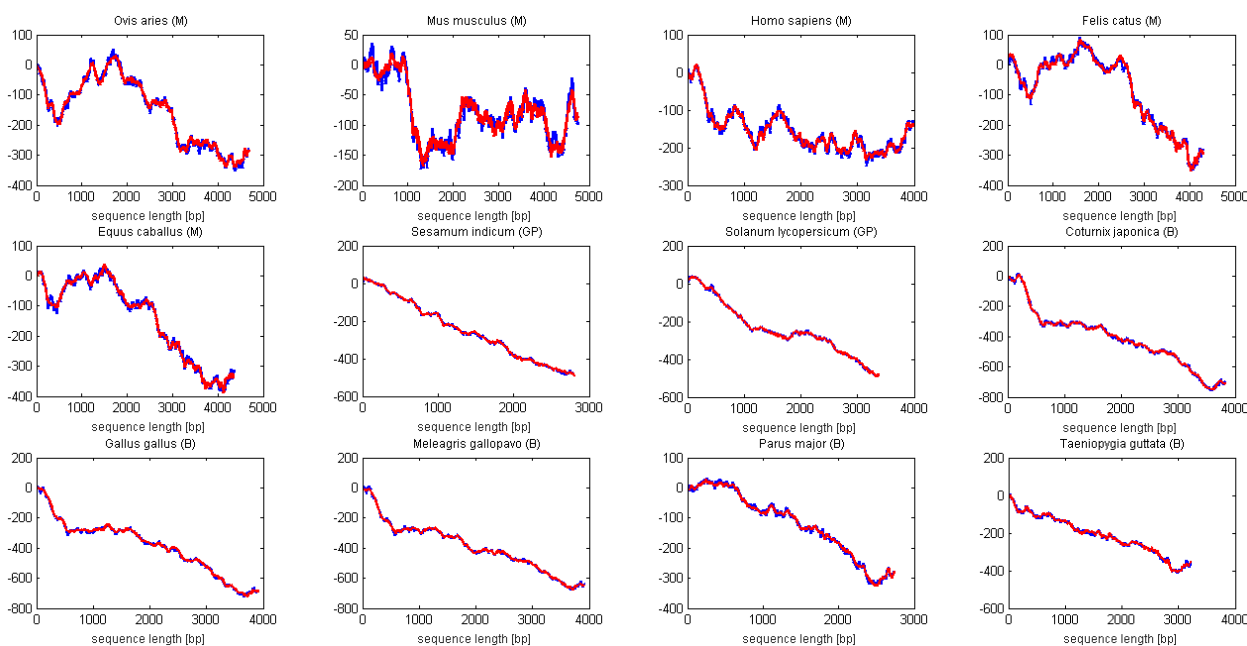
به روش فاز انباره ای است. هم ترازوی دو به دو توالی ها درکی از شباهت و خویشاوندی دو توالی تحت مطالعه را فراهم می کند و نیز پایه جستجوی تشابهات در پایگاه داده ها و هم ترازوی های چندگانه است [۱۸]. ما از الگوریتم DTW که الگوریتمی برای هم ترازوی سیگنال ها است و به مثابه هم ترازوی کلی عمل می کند استفاده کردیم. الگوریتم فوق به سیگنال های فشرده شده مرحله قبل اعمال شد.

جدول ۲: مقایسه زمان لازم برای فشرده سازی - بازسازی با دو روش مبتنی بر CS و DWT

DWT	CS	
۷۵	۷۵	نرخ فشرده سازی (%)
۱۶	۰/۷	زمان اجرای برنامه (ثانیه)

### ۴-۳ هم ترازوی سیگنال ها

مقایسه توالی ها در قلب آنالیزهای بیوانفورماتیک قرار دارد. لازمه این امر هم ترازوی دو به دو دنباله های DNA و یا سیگنال های استخراج شده



شکل ۷: سیگنال های اصلی (آبی) و بازسازی شده از روی فشرده های با CR = 75% (قرمز) متعلق به GAPDH ارگانسیم ها

M: پستانداران، GP: گیاهان زمینی، B: پرندگان

جدول ۳ ماتریس فاصله نرمالیزه شده ارگانسیم ها به روش سیگنال مینا و با استفاده از رابطه ۱۵ را نشان می دهد. همچنین جهت مقایسه، ماتریس فاصله بین دنباله های 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 ماتریس ها در مرحله بعد بمنظور رسم درختچه های فیلوژنتیکی مورد استفاده قرار خواهند گرفت. زمان لازم برای هم ترازوی سیگنال های فشرده شده و نیز دنباله های کاراکتری در شکل ۸ آورده شده است. همان طور که از شکل مشاهده می شود زمان صرف شده برای هم ترازوی به روش سیگنالی و از روی سیگنال های فشرده شده

از آن جایی که جفت سیگنال ها بعد از هم ترازوی کلی، دارای طول یکسان  $n$  می شوند، فاصله آن ها از هم به کمک رابطه زیر محاسبه می شود:

$$d = \sqrt{\sum_{i=1}^n [x(i) - y(i)]^2} \quad (16)$$

که  $x(n)$  و  $y(n)$  سیگنال های فشرده شده هم تراز شده هستند و  $d$  فاصله دو سیگنال از هم است [۷]. نتایج حاصل از هم ترازوی سیگنال ها به صورت ماتریسی تحت عنوان ماتریس فاصله ثبت می گردد. با کمک این ماتریس، امکان تحلیل خوشه ای دنباله ها فراهم می شود که در بخش بعدی مفصل تر توضیح داده خواهد شد.

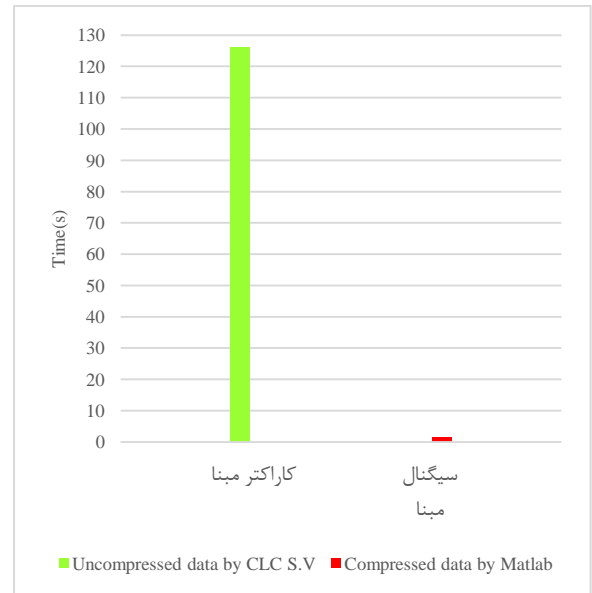
	Felis	Equus	Ovis	Homo	Mus	Gallus	Meleagris	Coturnix	Parus	Taeniopygia	Sesamum	Solanum
Felis	.											
Equus	۰/۲۰	.										
Ovis	۰/۲۷	۰/۲۷	.									
Homo	۰/۲۸	۰/۲۹	۰/۲۳	.								
Mus	۰/۲۴	۰/۲۴	۰/۴۲	۰/۲۸	.							
Gallus	۰/۶۶	۰/۶۷	۰/۷۰	۰/۶۴	۰/۶۹	.						
Meleagris	۰/۶۸	۰/۶۹	۰/۷۱	۰/۶۵	۰/۷۱	۰/۰۹	.					
Coturnix	۰/۶۹	۰/۷۱	۰/۷۲	۰/۶۸	۰/۷۴	۰/۱۲	۰/۱۳	.				
Parus	۰/۵۵	۰/۵۸	۰/۵۷	۰/۵۲	۰/۵۳	۰/۲۹	۰/۲۹	۰/۳۰	.			
Taeniopygia	۰/۶۲	۰/۶۳	۰/۶۴	۰/۵۹	۰/۶۱	۰/۳۰	۰/۳۰	۰/۳۲	۰/۰۸	.		
Sesamum	۱/۱۱	۱/۱۲	۱/۱۶	۱/۰۷	۱/۱۲	۱/۰۲	۰/۹۸	۱/۰۲	۰/۹۸	۱/۰۷	.	
Solanum	۱/۱۲	۱/۱۲	۱/۲۳	۱/۱۷	۱/۱۵	۱/۱۴	۱/۰۹	۱/۱۲	۱/۱۶	۱/۱۷	۰/۴۷	.

### ۳-۳ خوشه‌بندی ارگانسیم‌ها

فیلوژنتیک، مطالعه تاریخ تکاملی ارگانسیم‌های زنده با استفاده از دیاگرام‌های مشابه درخت است تا شجره‌نامه این ارگانسیم‌ها 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 ارگانسیم‌های موجود فراهم کنند. زیرا وقتی ارگانسیم 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 ارگانسیم‌ها 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<sup>[۲۲]</sup>) برای دستیابی به بهترین فاصله خوشه‌ای استفاده شده است. در شکل ۹ خوشه‌ها به‌طور مناسبی از هم جدا شده‌اند و ارگانسیم‌های مربوط به یک فرمانرو در کنار هم قرار گرفته‌اند. لذا، روش پیشنهادی ما توانسته ارگانسیم‌های مختلف را با استفاده از یک ژن خانه‌دار به‌طور مناسبی خوشه بندی نماید. همان‌طور که مشاهده می‌شود، ارگانسیم‌های هر فرمانرو در مجاورت هم قرار گرفته‌اند. به‌علاوه، فرمانرو پستانداران و پرندگان دارای ریشه مشترک هستند.

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



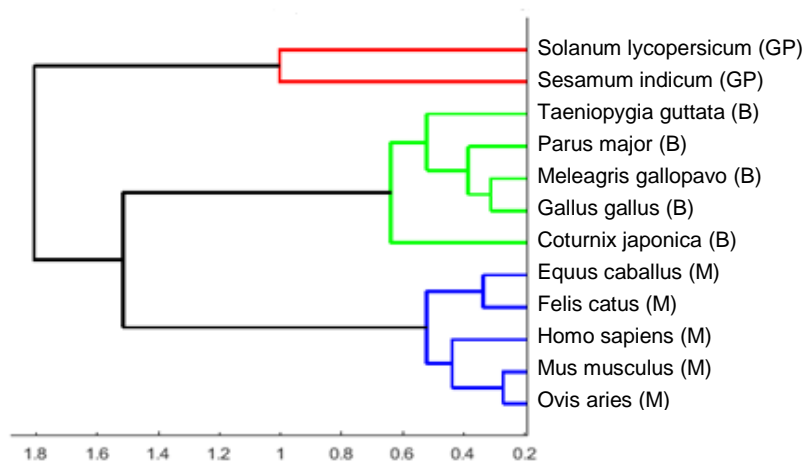
شکل ۸: زمان صرف شده برای هم‌ترازی GAPDH ارگانسیم‌ها به تفکیک روش‌ها و برنامه‌ها [محور عمودی زمان برحسب ثانیه و محور افقی روش‌ها و برنامه‌ها]

جدول ۳: فاصله اقلیدسی سیگنال‌های فشرده GAPDH ارگانسیم‌ها [پستانداران (ارغوانی)، گیاهان زمینی (سبز)، پرندگان (آبی)]

	Ovis	Mus	Homo	Felis	Equus	Sesamum	Solanum	Coturnix	Gallus	Meleagris	Parus	Taeniopygia
Ovis	.											
Mus	۰/۲۸	.										
Homo	۰/۲۶	۰/۲۷	.									
Felis	۰/۲۶	۰/۲۲	۰/۲۸	.								
Equus	۰/۵۶	۰/۴۲	۰/۴۶	۰/۳۵	.							
Sesamum	۰/۸۳	۰/۷۲	۰/۶۱	۰/۷۲	۰/۷۲	.						
Solanum	۰/۶۳	۰/۵۵	۰/۵۰	۰/۵۶	۰/۶۰	۰/۶۹	.					
Coturnix	۰/۹۰	۰/۸۵	۰/۷۶	۰/۸۷	۰/۹۲	۰/۹۶	۱	.				
Gallus	۰/۸۲	۰/۸۰	۰/۶۸	۰/۸۳	۰/۸۷	۰/۹۶	۰/۸۸	۰/۷۴	.			
Meleagris	۰/۹۲	۰/۸۱	۰/۷۹	۰/۸۵	۰/۸۸	۰/۹۸	۰/۸۷	۰/۹۵	۰/۸۰	.		
Parus	۰/۶۰	۰/۴۶	۰/۴۲	۰/۴۶	۰/۵۴	۰/۴۱	۰/۵۰	۰/۸۹	۰/۸۱	۰/۸۵	.	
Taeniopygia	۰/۶۵	۰/۵۰	۰/۵۰	۰/۵۲	۰/۶۰	۰/۵۲	۰/۵۹	۰/۹۱	۰/۸۷	۰/۸۲	۰/۴۱	.

جدول ۴: فاصله کاراکتری GAPDH ارگانسیم‌ها با نرم‌افزار CLC





شکل ۹: درختچه فیلوژنتیکی حاصل از سیگنال های فشرده شده GAPDH ارگانیزم ها با  $CR=75\%$  به وسیله نرم افزار متلب، [M: پستانداران، GP: گیاهان زمینی، B: پرندگان]

- [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 comparison," *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 شگری و محمدحسین کهایی، «حسگری فشرده تصاویر ابرطیفی با دسته بندی طیفی و بازسازی با تنظیم کننده 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] محمد مهدی محدث و محمدحسین کهایی، «ساخت ماتریس های نمونه برداری یقینی براساس توابع هش»، *مجله مهندسی برق دانشگاه تبریز*، دوره ۴۶، شماره ۴، صفحه ۳۰۷-۳۱۵، زمستان ۱۳۹۵.

[۱۸] رضانی، مجیدی، مهدی زاده، ملکی و پوربرخورداری، *اصول بیوانفورماتیک*، زنجان، جامعه نگر، ۱۳۹۳.

#### ۴- نتیجه

جوامع علمی و علی الخصوص غربی سرمایه گذاری ها و پژوهش های زیادی بر روی ژنوم، رمزگشایی و تحلیل آن به منظور استفاده در کارکردهای مختلف پزشکی، زیستی، نظامی و... انجام می دهند. تاکنون اکثر این مطالعات با روش های کاراکتر مینا بوده است. به نظر می آید با وجود طیف عظیمی از الگوریتم های متنوع پردازش سیگنال و مزایایی که برای بیان سیگنالی ژن ذکر گردید، می توان از ظرفیت های این حوزه برای مقابله با چالش های 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 می گردد تحلیل خوشه ای به روش سیگنالی و از روی سیگنال های فشرده شده با  $CR = 75\%$  هم پای تحلیل خوشه ای به روش کاراکتری عمل کرده است. در حالی که زمان خیلی کمتری را صرف کرده است، و دوماً با بیان سیگنالی، می توان ویژگی های توالی ها را در مقیاس بزرگ دید. سوماً با فشرده سازی هزینه پردازش، ذخیره سازی و انتقال کاهش می یابد.

#### 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 Electro/Information Technology*, pp. 307–310, 2009.

- <sup>16</sup> Block sparse bayesian learning
- <sup>17</sup> Discrete cosine transform
- <sup>18</sup> Maximum a posterior
- <sup>19</sup> Percentage root mean square difference
- <sup>20</sup> Normalized mean square error
- <sup>21</sup> Compression ratio
- <sup>22</sup> Unweighted pair group method with arithmetic mean

#### زیر نویس ها

- <sup>1</sup> Deoxyribonucleic acid
- <sup>2</sup> Sequencing
- <sup>3</sup> Alignment
- <sup>4</sup> Genomic signal processing
- <sup>5</sup> Discrete wavelet transform
- <sup>6</sup> Compressed sensing
- <sup>7</sup> CLC sequence viewer
- <sup>8</sup> Glyceraldehyde 3-phosphate dehydrogenase
- <sup>9</sup> <https://www.ncbi.nlm.nih.gov/gene/?term=GAPDH>
- <sup>10</sup> Housekeeping gene
- <sup>11</sup> Master gene`
- <sup>12</sup> Cumulated phase
- <sup>13</sup> Adenine, Cytosine, Guanine, Thymine
- <sup>14</sup> Sensing matrix
- <sup>15</sup> Underdetermined problem