



خوشه‌بندی تعدادی از ژن‌های موثر در تولید شیر با استفاده از تئوری اطلاعات و اطلاعات متقابل

هوشنگ دهقان‌زاده^۱، سید ضیاء‌الدین میرحسینی^۲، مصطفی قادری زفره‌بی^۳، حسن توکلی^۴ و سعید اسماعیل‌خانین^۵

۱- استادیار بخش تحقیقات علوم دامی، مرکز تحقیقات کشاورزی و منابع طبیعی استان گیلان، سازمان تحقیقات، آموزش و ترویج کشاورزی، رشت، ایران،
(نویسنده مسوول: H_dehghanzadeh@yahoo.com)

۲- استاد گروه علوم دامی، دانشکده کشاورزی، دانشگاه گیلان، رشت، ایران

۳- استادیار گروه علوم دامی، دانشکده کشاورزی، دانشگاه یاسوج، یاسوج، ایران

۴- استادیار گروه مهندسی برق، دانشکده فنی، دانشگاه گیلان، رشت، ایران

۵- دانشیار موسسه تحقیقات علوم دامی کشور، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج، ایران

تاریخ دریافت: ۹۶/۷/۱۴ تاریخ پذیرش: ۹۷/۶/۳۱

صفحه: ۱۱۷ تا ۱۳۲

چکیده

نظریه اطلاعات، شاخه‌ای از ریاضیات است. از تئوری اطلاعات در تجزیه و تحلیل‌های ژنتیکی و بیوانفورماتیکی استفاده گردیده و می‌توان از آن در آنالیزهای مربوط به ساختارها و توالی‌های زیستی نیز استفاده نمود. در این پژوهش بعد از استخراج توالی DNA ژن و آگزون‌های موثر بر تولید شیر در گاو شیری، فراسنجه آنتروپی در مراتب یک الی چهار برای هر ژن و آگزون‌های هر ژن محاسبه شد. برای استخراج تشابه میان ژن‌ها از یکدیگر، از اطلاعات متقابل بین ژن‌ها استفاده شد. نتایج با استفاده از هفت روش معمول خوشه‌بندی شدند. با توجه به تعدد نتایج، جهت افزایش دقت و تجمیع نتایج حاصل، از الگوریتم آدابوست استفاده گردید. در پایان جهت تایید نتایج حاصل از آدابوست و پیش‌بینی عملکرد ژن‌ها و ارتباط بین آن‌ها، با مراجعه به تارگاه GeneMANIA نتایج بر اساس حاشیه‌نویسی ژنومی آن‌ها مورد بررسی و مقایسه قرار گرفت. تجمیع نتایج هر خوشه‌بندی که با الگوریتم آدابوست انجام شد و خود نوعی درخت ژنی را تداعی می‌کند، نشان داد که روش پیشنهادی برای خوشه‌بندی مجموعه‌ای از ژن‌ها، از نظر زیستی جواب معقولی را حاصل می‌کند چرا که با نتایج حاشیه‌نویسی ژنومی ژن‌های حاصل در تارگاه GeneMANIA مطابقت داشت. اعتقاد بر این است که روش ارائه شده برای ایجاد درخت ژنی با سایر روش‌های متکی به توالی DNA برای خوشه‌بندی مجموعه‌ای از ژن‌ها، می‌تواند رقابت نماید و لذا می‌تواند در گروه‌بندی ژن‌های سایر گونه‌ها نیز به کار رود.

واژه‌های کلیدی: آنتروپی، اطلاعات متقابل، تئوری اطلاعات، خوشه‌بندی ژن، گاو شیری

مقدمه

مطالعه روی ژن‌های موثر در تولید شیر می‌تواند گامی مهم برای شناسایی و توسعه انتخاب به کمک نشانگر و تدوین برنامه‌های اصلاح نژادی برای بهبود این صفات در صنعت تولید شیر به شمار آید (۴،۱۷،۳۳،۴۵،۵۶). در طی سالیان گذشته سامانه داده‌های بیولوژیکی با نرخ بسیار بالایی تولید و پایگاه‌های اطلاعاتی جهت ثبت و پذیرش و نگهداری توالی ژن‌ها و پروتئین‌های مختلف جانداران ایجاد گردید. در اصلاح نژاد دام با در دست بودن ژن‌های مرتبط با صفات تولیدی مثل تولید شیر، این امکان وجود دارد که میزان اطلاعات ذخیره شده در بخش‌های مختلف آنها را با استفاده از نظریه اطلاعات^۱ بررسی و با تفسیر زیستی نتایج حاصله، رهیافت جدیدی را برای افزایش تولید شیر و یا دستکاری‌های ژنی با اهداف متفاوت ایجاد کرد.

امروزه جهان شاهد ظهور ابزارها و الگوریتم‌های رایانشی^۲ جدید جهت آزمایش و فرموله کردن فرضیه‌هایی همچون چگونگی سازماندهی و تکامل ژنوم و رویت یک فنوتیپ مشخص از یک ژنوم رمز نگاری شده است. نظریه اطلاعات، شاخه‌ای از ریاضیات است که با مهندسی ارتباطات، زیست‌شناسی و پزشکی هم‌پوشانی دارد این نظریه در سال ۱۹۴۸ توسط کلود شانون ارائه شد که به کشف و بررسی قوانین ریاضی حاکم بر رفتار داده‌ها در مراحل انتقال، ذخیره و بازیابی اختصاص دارد. از نظریه اطلاعات در تجزیه و تحلیل‌های ژنتیکی و بیوانفورماتیکی استفاده گردیده و می‌توان از آن

جهت بسیاری از تحلیل‌های مربوط به ساختارها و توالی‌های زیستی استفاده نمود (۲۹).

آنتروپی^۳ شانون هسته اصلی نظریه اطلاعات است و گاهی اوقات تحت عناوینی مثل اندازه عدم قطعیت یا میزان تصادفی بودن^۴، درهم ریختگی و پیش‌بینی‌ناپذیری^۵ شناخته می‌شود. اطلاعات، مقیاس عدم اطمینان یا آنتروپی در یک موقعیت است، هرچه عدم قطعیت (آنتروپی) یک سامانه بیشتر باشد، اطلاعات آن نیز بیشتر خواهد بود. وقتی موقعیتی کاملاً قابل پیش‌بینی است، هیچ اطلاعاتی در مورد آن وجود ندارد. این وضعیت را استحکام (نگو آنتروپی^۶) می‌گویند (۴۱). واحد آنتروپی بیت^۷ است و آنتروپی یک سامانه با میزان اطلاعات موجود در آن مرتبط است. سامانه با نظم بیشتر می‌تواند با بیت‌های کمتری از اطلاعات توصیف شود، در حالیکه سامانه‌ای با نظم کمتر برای توصیف شدن به بیت‌های بیشتری از اطلاعات نیازمند است (۱۶).

گاتلین در سال ۱۹۷۲ یک نسخه کلاسیک در کاربرد نظریه اطلاعات برای سامانه‌های زنده ارائه داد: حیات یک سامانه پردازش اطلاعات می‌باشد که از طریق تکامل، توانایی ذخیره و پردازش اطلاعات لازم برای بازساخت خود را بدست می‌آورد. گاتلین در واقع حیات را به عنوان یک فرآیند اطلاعاتی تعریف کرد. او راه را برای استفاده از یک چارچوب تحلیلی قدرتمند برای سیستم‌های بیولوژیکی بوسیله نظریه خود هموار ساخت (۱۲).

1- Information theory
6- Nego Entropy

2- Computational
7- Bit

3- Entropy

4- Randomness

5- Unpredictability

(۴۴،۴۳)، روش دستوری دینامیک^۹ و روش مدل مارکف^{۱۰} (۴۳،۳۷) و روش‌های فراکتال^{۱۱} (۵۴،۵۳).

روابط فیلوژنتیک میان ژن‌ها و موجودات با یک درخت فیلوژنتیک توصیف می‌شود. ساخت درخت فیلوژنی از توالی‌های DNA دارای دو فاز اصلی است: اولین فاز ساخت ماتریس فاصله حاصل از اندازه‌گیری جفتی دو توالی DNA با استفاده از همترازی چندگانه یا همترازی آزاد^{۱۲} و فاز دوم ساخت درخت فیلوژنی از ماتریس فاصله با استفاده از روش‌های ساخت درخت، که اغلب الگوریتم‌های معمول برای مقایسات بیولوژیکی توالی‌ها بر اساس همترازی توالی‌ها است. در عین حالی که همترازی چندگانه یک نقش اساسی در مقایسه توالی‌ها بازی می‌کند و به صورت معمول برای خوشه‌بندی توالی‌های DNA و پروتئین استفاده می‌شود (۵۱) اما پیچیدگی محاسباتی و نیاز به حافظه بالا برای توالی‌های با طول بزرگ دارد (۲۳،۱۰). به‌علاوه اغلب روش‌های همترازی چندگانه به مینیمم کردن تعداد الحاق و حذف شکاف‌ها^{۱۳} در توالی‌های DNA منجر گردیده است بنابراین روش‌های همترازی چندگانه در توالی‌هایی که شامل نواحی همولوگوس ضعیف یا جهش باشد که اکثر بخش‌های توالی‌های ژنومیکی درگیر آن هستند ممکن است یک ناهمترازی ایجاد نماید. برای رفع این مشکلات در همترازی‌های چندگانه، تحقیقات قابل توجهی روی همترازی‌های آزاد صورت پذیرفت. بلازیدل (۲) برای اولین بار یک روش همترازی آزاد بر اساس فراوانی کلمات K-mer توالی‌های DNA ارائه داد. این روش بطور وسیعی اینک در آنالیز ژنومیک به عنوان یک روش همترازی آزاد به کار می‌رود (۴۸،۴۰،۲۱،۸).

هرچند روش‌های معمول همترازی آزاد ممکن است مشکلاتی که همترازی چندگانه ایجاد می‌نماید را حل نماید اما اغلب این روش‌ها نیاز به توان محاسباتی و فضای حافظه بالا دارد و یک نکته مهم این است که در اغلب این روش‌ها محتوای اطلاعات درون توالی DNA از دست می‌رود و صحت خوشه بندی داده‌های توالی کاهش می‌یابد (۶).

در این مقاله یک روش جدید برای خوشه‌بندی ژن‌ها و ژنوم‌ها ارائه شده است. یک روش همترازی آزاد با استفاده از اطلاعات متقابل بر اساس فاصله توالی‌های DNA جهت خوشه‌بندی آنها ارائه شد که این فاصله بر اساس بردار ماتریسی از توالی‌های DNA بود. تازگی این روش این بود که با استفاده از تئوری اطلاعات و اطلاعات متقابل توالی‌های با طول متفاوت توانستند بدون همترازی مورد مقایسه قرار گیرند.

اطلاعات متقابل می‌تواند اطلاعات مشترک بین دو توالی را محاسبه و توصیف نماید و بدین ترتیب می‌توان از آن به عنوان معیاری جهت اندازه‌گیری تشابه و تفاوت دو توالی استفاده کرد (۵۵). که این بسیار سودمند می‌باشد زیرا اطلاعات کدشده توالی‌های بیولوژیک و وقوع اتفاقات تکاملی اشتراکات دو توالی با یک چد مشترک را تفکیک و جدا می‌سازد که منتج به از دست رفت اطلاعات مشترکشان می‌گردد (۹).

از نظریه اطلاعات به عنوان ابزاری مهم و به چند صورت برای جستجوی الگوهایی در توالی‌های DNA (۴۵)، نقش آمینو اسیدها در ساختار پروتئین‌ها در مخمر (۲۵)، تحلیل جایگاه‌های صفات کمی^۱ و اپیستاسیس^۲ (۳۹)، بررسی اطلاعات ژنوم جهانی^۳ (۳۰)، تحلیل داده‌های ریزآرایه DNA (۲۰)، طبقه‌بندی ژن‌های درگیر در سرطان (۳۶)، مقایسه اندازه پیچیدگی برای آنالیز توالی‌های DNA (۳۱،۲۸،۱۶)، بازساخت درختان فیلوژنتیکی بدون هم‌مدیف کردن بازها (۳۵)، پژوهش‌های تکاملی (۱۲)، تنوع ژنتیکی (۴۲)، مقایسه محتوای اطلاعات نواحی اینترون و اگزون ژن‌ها (۵۲) و تحلیل زیر گونه‌های انگل کریپتوزپوریوم^۴ (۳۲) استفاده شده است.

پس از کامل شدن برخی از پروژه‌های تعیین توالی ژنوم، علاقه‌مندی به روشن شدن و یافتن اینکه چطور سیستم‌های زنده در همه سطوح اطلاعات بیولوژیکی عمل می‌کنند، بوجود آمد و به دنبال آن به محققین امکان بازسازی شبکه‌های متابولیکی در مقیاس بزرگ ژنی با استفاده از روش‌های توپولوژیک داده شد. یکی دیگر از آنالیزهای توپولوژیک در شبکه‌های بیولوژیک، آنالیز کلاسترینگ^۵ می‌باشد که نشان‌دهنده دسته‌بندی شبکه به قسمت‌های مختلف (کلاستر) است. در چند دهه اخیر چندین روش برای خوشه‌بندی ژن‌ها و پروتئین‌ها پیشنهاد شد اغلب این روش‌ها بر اساس همترازی ژن‌ها بود که با استفاده از سیستم‌های امتیازدهی بدست می‌آمدند. تعدادی از فیلوژنی‌ها بوسیله روش‌های معمول و با همترازی توالی‌ها ساخته می‌شود اما با توجه به اینکه بسیاری از توالی‌ها بزرگ هستند و روش‌های استاندارد بر اساس مقایسه هر نوکلئوتید به نوکلئوتید متناظر توالی دیگر استوار است در یک رنج بالا کارایی را پایین و کمی مشکل و غیر ممکن می‌سازد (۵۷). این روش‌ها خوشه‌بندی دقیقی از توالی‌های بیولوژیکی فراهم می‌نماید و در این راستا چندین الگوریتم توسعه و به طور موفقیت‌آمیزی به کار گرفته شدند (۲۶،۲۲،۱۰).

همترازی‌های چندگانه توالی‌ها، یک روش مشهور برای طبقه‌بندی توالی‌های DNA می‌باشد هرچند که با محدودیت‌های اساسی در پیچیدگی محاسباتی مواجه می‌باشد. روش‌های همترازی آزاد همچون روش K-mer هم در چند دهه گذشته برای مقایسات و طبقه‌بندی توالی‌های DNA متداول گشته که کارآمدتر از روش‌های همترازی چند گانه هستند (۶). هر چند در اغلب روش‌های همترازی آزاد ممکن است اطلاعات ساختاری و عملکردی توالی‌های DNA از دست رفته باشد زیرا همه آنها بر اساس استخراج ویژگی عمل می‌نمایند. همچنین ممکن است به طور کامل تفاوت‌های واقعی میان زنجیره‌های DNA منعکس نشود لذا روش‌های همترازی آزاد با حفظ اطلاعات برای مقایسات با دقت بیشتر برای طبقه‌بندی DNA مورد نیاز می‌باشد. تاکنون روش‌های مختلف و جدیدی برای بازسازی درخت فیلوژنی بدون همترازی توالی‌ها پیشنهاد گردیده است، مثل آنالیز بر اساس مولفه‌ها^۷ (۱۱)، روش تجزیه مقادیر منفرد^۸

| | | | | |
|---------------------------------------|-------------|---------------------------------------|-------------------------|---------------|
| 1- Quantitative trait locus (QTL) | 2-Epistasis | 3- Global genomic information | 4- Cryptosporidium | 5- Clustering |
| 6- Multiple sequence alignment (MSA) | | 7- Principal component analysis (PCA) | | |
| 8- Singular value decomposition (SVD) | | 9- Dynamical language method | 10- Markov model method | |
| 11- Fractal methods | | 12- Alignment-free methods | 13- GAPS | |

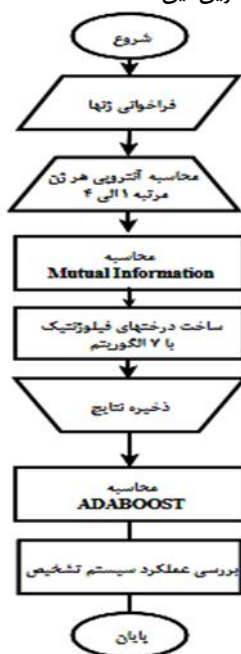
روش‌ها آدابوست^۲ می‌باشد که فروند و شاپیر (۱۳،۱۴) آن را معرفی نمودند که از آن زمان توجه زیادی را در حوزه هوش مصنوعی و یادگیری ماشینی به خود جلب کرده است.

مواد و روش‌ها

استخراج توالی‌های DNA ژن‌ها

گزارش شده است که در کل، حدود ۶۸۷۵ ژن وجود دارد که روی تولید شیر در پستانداران موثر هستند. بعضی از این ژن‌ها فقط در غده پستانی بیان می‌شوند و بعضی دیگر در بافت‌های دیگری مثل کبد، کلیه، ماهیچه‌ها و غیره نیز بیان می‌شوند (۲۷). ژن‌های مورد بررسی در این مقاله از نتایج پژوهش لمی و همکاران (۲۷) انتخاب گردیدند. در مقاله یاد شده، ۳۰ ژن از دسته ژن‌های پستانی موثر در تولید شیر مربوط به گاوهای تلیسه^۳ به صورت تصادفی انتخاب و مورد بررسی و واکاوی قرار گرفتند. توالی و همچنین سایر اطلاعات ژن‌ها از جمله اندازه هر ژن، محتوای گوانین-سیتوزین^۴، شماره دست‌یابی^۵، تعداد و طول هر اگزون^۶ و جایگاه آن بر روی کروموزوم از بانک ژنی NCBI^۷ دریافت و سپس با پیکربندی فستا^۸ ذخیره گردیدند. (جدول ۱ و ۲ فایل ضمیمه).

اطلاعات متقابل ارائه شده مبتنی بر آنتروپی می‌باشد و محاسبه اطلاعات متقابل نیاز به دانستن آنتروپی منابع مورد مقایسه می‌باشد. در روش ارائه شده اطلاعات متقابل در رتبه‌های ۱ تا ۴ آنتروپی برای هر توالی DNA محاسبه گردید و در هر مرتبه از محاسبه آنتروپی ماتریس تشابه-فاصله آن ایجاد و سپس جهت افزایش صحت خوشه‌بندی در هر رتبه از آنتروپی، آنها را به طور جداگانه با ۷ روش معمول در خوشه‌بندی سلسله مراتبی آنالیز نمودیم. تعدد نتایج خوشه‌بندی حاصل از روش‌های یاد شده باعث شد که با این سوال مواجه که بهترین نتیجه حاصل کدام است و تفسیر نتایج بر اساس کدام دسته‌بندی باید صورت پذیرد برای حل این مشکل از روش ترکیب نتایج استفاده شد. استفاده از ترکیب نتایج چند دسته‌بندی، یکی از روش‌های افزایش کارایی و صحت سیستم‌های بازشناسی الگو است که در سال‌های اخیر محققین زیادی به آن پرداخته‌اند. لذا برای بهبود صحت دسته‌بندی پس از استفاده از دسته‌بندی‌های مختلف که برای نتایج هر بخش از محاسبات اطلاعات متقابل از آنها استفاده گردید از ترکیب نتایج خروجی به عنوان بهترین نتیجه جهت تفسیر بیولوژیک ژن‌ها استفاده شد. این روش اغلب تحت عنوان سیستم‌های طبقه‌بند چندگانه^۱ و یا سیستم‌های شورایی خوانده می‌شود. یکی از مطرح‌ترین این



شکل ۱- روند اجرای پژوهش
Figure 1. Workflow of this research

محاسبه مراتب^۹ آنتروپی

نظریه اطلاعات کلاسیک بر روی تابع زیر و خواص و تعبیرهای آن که تابع آنتروپی یا تابع شانون نامیده می‌شود بنا شده است. ژنوم می‌تواند به‌عنوان یک رشته طویل و متشکل از یک الفبا () با ۴ سمبل (A, T, C, G) در نظر گرفته شود که هر کدام از آنها ممکن است در یک موقعیت معین از

جهت آماده‌سازی اطلاعات استخراج شده از پایگاه داده به دلیل زیاد بودن حجم اطلاعات ژن‌ها و اگزون‌های مربوط به آن، نرم‌افزاری طراحی شد که به طور هوشمند، ویژگی‌های ژن‌ها را استخراج کرد. لذا در این نرم‌افزار با توجه به خواسته پژوهش، خروجی‌های مناسب بدست آمدند. برای ایجاد این نرم‌افزار از زبان برنامه‌نویسی C# استفاده شد.

1- Multiple classifier system

2- AdaBoost

3- Virgin mammary gene set

4- C-G content

5- Accession number

6- Exone

7- <http://www.ncbi.nlm.nih.gov/genbank/gene>

8- Fasta

9- Orders

توالی نیز محاسبه شد تا میزان تصادفی بودن توالی ژن مورد مقایسه قرار گیرد (جدول ۱). در ضمن، اندیسی که در H ظاهر می‌شود نشان‌دهنده مرتبه آنتروپی مورد نظر است.

محاسبه اطلاعات متقابل ژن‌ها و اگزون‌های آنها

هرگاه دو متغیر تصادفی (X, Y) داشته باشیم که لزوماً از هم مستقل نباشند تابع آنتروپی یا اطلاعات به طور طبیعی به شکل زیر تعریف می‌شود: (رابطه ۱)

$$H(X, Y) := - \sum_{x,y} p(x, y) \log_2 p(x, y)$$

درحالی‌که دو متغیر تصادفی مستقل باشند یعنی $p(x; y) = p(x)q(y)$ از رابطه (۱) می‌توان نتیجه گرفت: (رابطه ۲)

دو متغیر تصادفی $(X, Y)_i$ توزیع آن‌ها با تابع $P(x, y)$ مشخص می‌شود $X \perp Y \Rightarrow H(X, Y) = H(X) + H(Y)$ فرض کنید که مقدار یکی از متغیرهای تصادفی مثل Y را می‌دانیم و این مقدار برابر است با y . در این صورت توزیع متغیر تصادفی X عوض خواهد شد و تبدیل خواهد شد به توزیع $P(X|y)$ که در آن y یک پارامتر است و X مقادیر متغیر را به خود می‌گیرد. می‌دانیم که:

$$H(X|y) := - \sum_x P(x|y) \log_2 P(x|y)$$

اگر بخواهیم بدانیم که به طور متوسط دانستن یک مقدار از Y چه مقدار اطلاعات در X باقی می‌گذارد باید روی $H(X|y_i)$ متوسط بگیریم بنابراین خواهیم داشت: رابطه (۳)

$$= - \sum_{x,y} P(x, y) \log_2 P(x|y) = - \sum_{x,y} P(x, y) \log_2 \frac{P(x, y)}{P(y)}$$

$$= H(X, Y) - H(Y).$$

به همان دلیلی که تابع $H(X)$ مثبت است تابع $H(X|y_i)$ و در نتیجه تابع $H(X|Y)$ نیز مثبت خواهند بود. $H(X|Y)$ را اطلاعات X مشروط به Y می‌خوانیم و این کمیت بیان‌کننده میزان اطلاعات باقیمانده در X است هرگاه ما مقادیر Y را دانسته باشیم. باید توجه داشت که این تابع متقارن نیست یعنی $H(X|Y) \neq H(Y|X)$

از رابطه (۳) به این نتیجه می‌رسیم: رابطه (۴)

$$H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$$

اگر دو متغیر تصادفی X, Y مستقل باشند آنگاه دانستن Y هیچ تاثیری در اطلاعات باقیمانده در X نخواهد داشت و در نتیجه:

$$H(X|Y) = H(X)$$

و بنابر رابطه (۴):

$$H(X, Y) = H(X) + H(Y)$$

بالعکس هرگاه X و Y کاملاً به هم وابسته باشند انتظار داریم که دانستن Y برای دانستن X نیز کفایت کند یعنی هیچ

رشته DNA جای گرفته باشند که می‌شود احتمال هر کدام از این کاراکترها را $f(A)=f(T)$ و $f(C)=f(G)$ بوسیله وقوعشان در توالی خطی ژنوم برآورد کرد. از این رو با شمارش تعداد A, C, T, G موجود در ژنوم و ظهور همه آنها در موقعیت‌هایی در ژنوم و نرمال کردن آنها به فراوانی‌های نسبی، آنتروپی ژنوم را می‌توان نتیجه گرفت (۳۸، ۱۲).

در این پژوهش برای هر ژن و اگزون هر ژن، فراسنجه آنتروپی در مراتب یک الی چهار محاسبه شد. در این راستا از زنجیره مارکف تا درجه ۳ استفاده شد. برای محاسبه آنتروپی مرتبه اول (مرتبه صفر زنجیره مارکف^(۱)) از فرمول زیر استفاده شد:

$$H(x)_I = - \sum_{i=1}^n p_i \log_2 p_i$$

که i و j نشانگر آگاهی از وقوع دو نوکلئوتید قبلی است و $p_{i,j}(k)$ احتمال وقوع نوکلئوتید k به شرط وقوع نوکلئوتیدهای i و j از مجموعه $\{C, G, T, A\}$ در توالی DNA ژن است.

آنتروپی مرتبه دوم (مرتبه یک زنجیره مارکف) با استفاده از فرمول زیر محاسبه شد:

$$H(x)_{II} = - \sum_{i=1}^n p_i \sum_{j=1}^n p_i(j) \log_2 p_i(j)$$

که i نشانگر وقوع نوکلئوتید قبلی و $p_i(j)$ هم احتمال وقوع نوکلئوتید j به شرط وقوع نوکلئوتید i از مجموعه $\{A, T, G, C\}$ زنجیره DNA است. آنتروپی مرتبه سوم (مرتبه دو زنجیره مارکف) نیز با استفاده از فرمول زیر محاسبه شد:

$$H(x)_{III} = - \sum_{i=1}^n p_i \sum_{j=1}^n p_i(j) \sum_{k=1}^n p_{i,j}(k) \log_2 p_{i,j}(k)$$

که i و j نشانگر آگاهی از وقوع دو نوکلئوتید قبلی است و $p_{i,j}(k)$ احتمال وقوع نوکلئوتید k به شرط وقوع نوکلئوتیدهای i و j از مجموعه $\{C, G, T, A\}$ در توالی DNA ژن است.

آنتروپی مرتبه چهارم (مرتبه سوم زنجیره مارکف) با استفاده از فرمول زیر محاسبه گردید:

$$H(x)_{IV} = - \sum_{i=1}^n p_i \sum_{j=1}^n p_i(j) \sum_{k=1}^n p_{i,j}(k) \sum_{m=1}^n p_{i,j,k}(m) \log_2 p_{i,j,k}(m)$$

که i و j و k : نشانگر آگاهی از وقوع ۳ نوکلئوتید قبلی است و $p_{i,j,k}(m)$ هم احتمال وقوع نوکلئوتید m به شرط وقوع نوکلئوتیدهای i, j, k از مجموعه $\{C, G, T, A\}$ در توالی DNA ژن است. همان‌طور که نشان داده شده در کل ۴ مرتبه آنتروپی محاسبه شد. محاسبه آنتروپی هم برای طول کل ژن‌ها و هم اگزون‌ها انجام شد. برای هر مرتبه از آنتروپی یک توالی تصادفی متناظر (RH) با فرض تصادفی بودن

می‌باشد. برای ترکیب نتایج ۲۸ خوشه ایجاد شده حاصل از ۷ الگوریتم بالا و ترکیب نتایج خوشه‌بندی‌ها، از الگوریتم آدابوست^۷ برای ترکیب نتایج خوشه‌بندی استفاده شد و با ترکیب این دسته‌بندی‌ها کارایی دسته‌بندی افزایش داده شد (۱۳، ۱۴). در پایان، جهت تایید نتایج حاصل از آدابوست و بررسی همخوانی نتیجه خوشه‌بندی ژن‌ها با داده‌های حاشیه نویسی ژنوم آن‌ها، از *server GeneMANIA prediction*^۸ استفاده شد و نتایج مورد بررسی قرار گرفت (۵۰). برای محاسبات از امکانات مرکز ملی ابررایانش شیخ بهایی دانشگاه اصفهان^۹ استفاده شد. همه محاسبات با کمک نرم‌افزار مهندسی متلب (۲۰۱۵) انجام گرفت.

نتایج و بحث

اطلاعات ۳۰ ژن مورد پژوهش در جدول ۱ و ۲ فایل ضمیمه قابل مشاهده می‌باشد. بررسی مشخصات ژن‌ها نشان داد، دو ژن *NOP2* و *YWHAH* (به ترتیب با طول ۶۰۱۶۷ و ۱۴۴۵) از نظر اندازه، بزرگترین و کوچکترین ژن‌های مورد بررسی در این پژوهش بودند. ژن‌های مورد بررسی در کل دارای ۲۱۱ اگزون بودند، اگزون شماره ۱ ژن *HSP6* و اگزون شماره ۱ ژن *ACTR2* (به ترتیب با طول‌های ۲۶۲۲ و ۱۰) بزرگترین و کوچکترین اگزون‌های مورد بررسی در این پژوهش بودند. همچنین ژن‌های *EIF3L* و *DGCR8* با ۱۳ اگزون و ژن‌های *HPS6* و *YWHAH* با ۱ اگزون بیشترین و کمترین تعداد اگزون را در این بررسی دارا بودند. مقادیر آنتروپی و آنتروپی تصادفی رشته متناظر کلیه ژن‌ها در هر رتبه در جدول ۱ آمده است.

اطلاعی در X باقی نگذارد یعنی $H(X|Y)=0$ که با توجه به رابطه ۴ به این معناست که: $H(X,Y) = H(Y)$. اطلاعات متقابل در دو متغیر تصادفی X و Y به شکل زیر تعریف می‌شود: رابطه (۵)

$$I(X : Y) := H(X) + H(Y) - H(X, Y)$$

این کمیت نسبت به دو متغیر تصادفی X و Y متقارن است. با توجه به رابطه (۴) می‌توان آن را به شکل زیر بازنویسی کرد: رابطه (۶)

$$I(X : Y) := H(X) - H(X|Y)$$

قبل از آنکه مقدار Y را بدانیم، اطلاعات موجود در X با $H(X)$ سنجیده می‌شود. با دانستن Y این اطلاعات به $H(X|Y)$ تقلیل پیدا می‌کند. بنابراین تفاوت این دو میزان اطلاعی است که Y درباره X حمل می‌کند. $I(X : Y)$ یک کمیت نامنفی است.

بر این اساس اطلاعات متقابل برای تمامی ژن‌ها و اگزون‌های مورد بررسی به طور جداگانه بر اساس آنتروپی تا مرتبه ۳ زنجیره مارکف محاسبه شد.

ساخت درخت فیلوژنی و ترکیب نتایج حاصل از خوشه‌بندی

معیار بدست آمده فاصله اطلاعات متقابل در مجموعه ژن‌ها و اگزون‌ها، به‌عنوان ورودی ۷ روش معمول خوشه‌بندی *'Single'*، *'Complete'*، *'Average'*، *'Weighted'*، *'Centroid'*، *'Median'* و *'KMeans'* به کار رفتند و خوشه‌بندی ژن‌ها یا درخت‌های ژنی بدست آمدند. در این مقاله تنها خوشه حاصل از روش *UPGMA* و *Nearest distance* آورده شده و بقیه در فایل ضمیمه قابل مشاهده

1- Nearest distance (single linkage method)
3- Unweighted pair group method average (UPGMA, group average)
5- Unweighted pair group method centroid (UPGMC)
7- AdaBoost
9- Sheikh Bahaei National High Performance Computing Center (SBNHPCC) Isfahan University

2- furthest distance (complete linkage method)
4- Weighted pair group method average (WPGMA)
6- Weighted pair group method centroid (WPGMC)
8- <http://www.genemania.org>

جدول ۱- آنتروپی محاسبه شده مراتب مختلف و آنتروپی تصادفی متناظرشان در توالی DNA ژن‌های موثر در تولید شیر گاو
Table 1. Calculated different of entropy orders and their corresponding random entropies in cow's milk governing genes

| شماره | نماد ژن | آنتروپی/رتبه ۱ تصادفی $H(x)_I/RH(x)_I$ | آنتروپی/رتبه ۲ تصادفی $H(x)_{II}/RH(x)_{II}$ | آنتروپی/رتبه ۳ تصادفی $H(x)_{III}/RH(x)_{III}$ | آنتروپی/رتبه ۴ تصادفی $H(x)_{IV}/RH(x)_{IV}$ |
|-------|------------|--|--|--|--|
| ۱ | EIF3L | 2.0000/1.9871 | 3.9994/3.9327 | 5.9980/5.8699 | 7.9920/7.7993 |
| ۲ | DES | 1.9991/1.9878 | 3.9986/3.9176 | 5.9929/5.8338 | 7.9677/7.7275 |
| ۳ | HPS6 | 1.9994/1.9617 | 3.9917/3.8513 | 5.9842/5.7134 | 7.9056/7.5078 |
| ۴ | FAM192A | 1.9999/1.9566 | 3.9996/3.8667 | 5.9979/5.7688 | 7.9928/7.6626 |
| ۵ | COPS6 | 1.9999/1.9931 | 3.9973/3.9388 | 5.9701/5.8660 | 7.9375/7.7404 |
| ۶ | YWHAH | 1.9983/1.9994 | 3.9940/3.9649 | 5.9527/5.9045 | 7.8605/7.7365 |
| ۷ | NSUN3 | 2.0000/1.9551 | 3.9998/3.8665 | 5.9990/5.7703 | 7.9966/7.6681 |
| ۸ | CALM1 | 1.9998/1.9913 | 3.9984/3.9504 | 5.9910/5.8955 | 7.9734/7.8161 |
| ۹ | CD34 | 1.9999/1.9974 | 3.9996/3.9404 | 5.9975/5.8681 | 7.9926/7.7877 |
| ۱۰ | TBC1D20 | 1.9998/1.9899 | 3.9995/3.9345 | 5.9974/5.8681 | 7.9900/7.7917 |
| ۱۱ | HTRA2 | 1.9997/1.9872 | 3.9966/3.9364 | 5.8759/5.9837 | 7.9315/7.7743 |
| ۱۲ | SLC35A3 | 1.9332/1.9995 | 3.8293/3.9994 | 5.7152/5.9962 | 7.5880/7.9857 |
| ۱۳ | CNOT8 | 1.9736/2.0000 | 3.9033/3.9994 | 5.8216/5.9971 | 7.7304/7.9881 |
| ۱۴ | DGCR8 | 1.9666/1.9999 | 3.8899/3.9990 | 5.7941/5.9971 | 7.6828/7.9848 |
| ۱۵ | SMIM14 | 1.9598/1.9999 | 3.8839/3.9998 | 5.7988/5.9993 | 7.7081/7.9961 |
| ۱۶ | MRPS11 | 1.9945/1.9998 | 3.9417/3.9984 | 5.8769/5.9946 | 7.7967/7.9817 |
| ۱۷ | CDK9 | 1.9889/1.9991 | 3.9344/3.9982 | 5.8663/5.9891 | 7.7677/7.9596 |
| ۱۸ | DALRD3 | 1.9585/1.9995 | 3.8711/3.9960 | 5.7672/5.9798 | 7.6230/7.9247 |
| ۱۹ | SPSB3 | 1.9390/1.9998 | 3.8173/3.9979 | 5.6846/5.9926 | 7.5244/7.9645 |
| ۲۰ | ZNF419 | 1.9925/1.9997 | 3.9284/3.9981 | 5.8516/5.9939 | 7.7540/7.9687 |
| ۲۱ | ZDHC4 | 1.9863/2.0000 | 3.9435/3.9970 | 5.8876/5.9941 | 7.8150/7.9746 |
| ۲۲ | B4GALT1 | 1.9949/2.0000 | 3.9296/3.9999 | 5.8552/5.9991 | 7.7756/7.9964 |
| ۲۳ | GRWD1 | 1.9855/1.9997 | 3.9017/3.9971 | 5.8109/5.9895 | 7.6984/7.9656 |
| ۲۴ | ACTR2 | 1.9572/2.0000 | 3.8736/3.9998 | 5.7813/5.9990 | 7.6828/7.9954 |
| ۲۵ | SI00A16 | 1.9835/1.9996 | 3.8754/3.9990 | 5.7561/5.9927 | 7.6163/7.9745 |
| ۲۶ | SNRPG | 1.9683/1.9998 | 3.8946/3.9983 | 5.8093/5.9943 | 7.7089/7.9768 |
| ۲۷ | TIMM21 | 1.9946/1.9996 | 3.9570/3.9974 | 5.9043/5.9916 | 7.8252/7.9611 |
| ۲۸ | NR1H2 | 1.9769/1.9998 | 3.8845/3.9979 | 5.7820/5.9927 | 7.6539/7.9692 |
| ۲۹ | C1H21orf59 | 1.9991/2.0000 | 3.9576/3.9994 | 5.9041/5.9954 | 7.8366/7.9816 |
| ۳۰ | RPS3A | 1.9668/1.9997 | 3.9038/3.9978 | 5.8295/5.9872 | 7.7327/7.9597 |

*: سلول‌های با رنگ خاکستری و صورتی به ترتیب بیشترین و کمترین مقادیر آنتروپی ژن‌ها را در رتبه مربوطه نشان می‌دهد.

جدول ۲- نتایج آنتروپی کمینه و بیشینه در مراتب ۱ الی ۴ ژن‌های مربوطه
Table 2. Results of maximum and minimum entropy orders of 1 to 4 in respected genes

| آنتروپی | نام ژن | مقدار | نام ژن | بیشترین مقدار آنتروپی | مقدار |
|--------------|---------|-------|------------|-----------------------|--------|
| $H(x)_I$ | SLC35A3 | ۱۴۹۰۱ | YWHAH | ۱۴۴۵ | ۱/۹۹۹۴ |
| $H(x)_{II}$ | SPSB3 | ۵۵۷۰ | YWHAH | ۱۴۴۵ | ۳/۹۶۴۹ |
| $H(x)_{III}$ | SPSB3 | ۵۵۷۰ | YWHAH | ۱۴۴۵ | ۵/۹۰۴۵ |
| $H(x)_{IV}$ | HPS6 | ۲۶۲۲ | C1H21orf59 | ۱۱۶۲۸ | ۷/۸۳۶۶ |

جدول ۳- نتایج آنتروپی بیشینه و کمینه در مراتب ۱ تا ۴ اگزون‌ها
Table 3. Results of different entropy orders of 1 to 4 over exons

| آنتروپی | نام اگزون | طول اگزون | مقدار | نام اگزون | طول اگزون | بیشترین مقدار آنتروپی |
|--------------|-------------------|-----------|--------|---------------------|-----------|-----------------------|
| $H(x)_I$ | Exon 1 gene SPSB3 | ۳۴ | ۱/۶۴۵۷ | Exon 1 gene YWHAH | ۱۴۴۵ | ۱/۹۹۹۴ |
| $H(x)_{II}$ | Exon 1 gene SPSB3 | ۳۴ | ۲/۹۲۲۰ | Exon 1 gene YWHAH | ۱۴۴۵ | ۳/۹۶۴۹ |
| $H(x)_{III}$ | Exon 1 gene ACTR2 | ۱۰ | ۲/۷۵۰۰ | Exon 8 gene TBC1D20 | ۲۲۸۶ | ۵/۸۴۵۸ |
| $H(x)_{IV}$ | Exon 1 gene ACTR2 | ۱۰ | ۲/۸۰۷۴ | Exon 1 gene YWHAH | ۱۴۴۵ | ۷/۸۳۶۶ |

نظر تشابه قطعات ژن هم وجوه مشترک زیادی داشتند چون در محاسبه آنروپی رتبه‌های ۳ و ۴ که مبتنی بر فراوانی جملات ۳ و ۴ حرفی (نوکلئوتیدی) می‌باشد با اطمینان ژن‌های با آنروپی نزدیک به هم که فاصله ژنی آن‌ها از هم نیز بسیار کم بوده و در یک خوشه قرار گرفتند شباهت ساختاری زیادی نسبت به همدیگر داشته و فراوانی این جملات مشترک و یکسان در آن‌ها به مراتب زیادتر از ژن‌های قرار گرفته در خوشه‌های دیگر است. طی تحقیقی نشان داده شد که آنروپی توالی‌های DNA نزدیک به بیشینه است (۱۸). همچنین طی مطالعه‌ای بر پایه‌های هندسی آنروپی، الگوریتم جدیدی برای تعیین ساختمان دوم پروتئین ارائه شد و نشان داده شد که قطعات دارای آنروپی پایین‌تر، نظم ساختاری بیشتری دارند (۱۷). با مقایسه پراکندگی مقدار آنروپی ژن‌ها نتیجه می‌گیریم که هرچقدر آنروپی توالی کل ژن به مقدار بیشینه خود نزدیک شود، آن ژن از اهمیت بیشتری برخوردار است. با توجه به اینکه اگزون‌های با مقادیر آنروپی بالا به‌خصوص در رتبه ۴ و نزدیک به مقدار حداکثر آنروپی، قطعاتی هستند که دارای عدم قطعیت بالا می‌باشند، لذا اگزون‌های آن نیز از آنروپی بیشتری برخوردارند و دلیل آن این است که اگزون‌ها قسمت معنی‌دار ژن بوده و نقش اصلی را بر عهده دارند و در بردارنده اطلاعات غالب توالی هستند و انتظار می‌رود که این نواحی تغییرات بیشتری را در ژن متحمل گردند که می‌توان این قطعات را جهت مقایسات ژنوتایی ژن‌های افراد یک جمعیت با استفاده از نشانگرهای مولکولی پیشنهاد نمود. هر چند تنوع ژنتیکی این قطعات در یک جمعیت باید در واقعیت مورد بررسی قرار گیرد که در صورت تایید این فرض می‌توان قطعات زیادی از ژن‌های بی‌شماری که از لحاظ اصلاح دام جزو ژن‌های تاثیرگذار در بروز صفات اقتصادی می‌باشند را با استفاده از تئوری اطلاعات مشخص و برای بررسی ارتباط ژنوتایی آنها با صفات تولیدی در گونه‌های مختلف موجودات معرفی نمود. با توجه به نتایج بالا پیش‌بینی می‌گردد اگزون‌هایی که در رتبه ۴ و نزدیک به حداکثر مقدار آنروپی می‌باشند مناسب جهت بررسی ژنوتایی با استفاده از نشانگرهای مولکولی باشند.

نتایج حاصل از محاسبه اطلاعات متقابل ژن‌ها و اگزون‌ها

پس از محاسبه آنروپی مراتب ۱ الی ۴ در ژن‌ها و اگزون‌ها، مقادیر اطلاعات متقابل برای هر مرتبه از آنروپی به طور جداگانه محاسبه شد. این روش به ژن‌ها و اگزون‌ها این اجازه را می‌دهد که با طول حقیقی و متفاوت و محتوای واقعی خود نسبت به یکدیگر مورد ارزیابی قرار گیرند. نتایج این بخش در جدول ۴ ارائه شده است. اطلاعات متقابل دسته ژنی و اگزون‌های آن، به طور مجزا و با آنروپی‌های مراتب ۱ الی ۴ محاسبه شد. در این روش یک ماتریس نامتقارن به اندازه تعداد ژن‌ها و یا اگزون‌های مورد بررسی در هر دسته ایجاد می‌گردد که بر این اساس ژن‌ها و اگزون‌هایی که بیشترین و کمترین اطلاعات متقابل را دارند مشخص شد.

در جدول ۲ و ۳ به ترتیب نتایج آنروپی کمینه و بیشینه در مراتب ۱ الی ۴ ژن‌ها و اگزون‌های مربوطه، نشان داده شده است. با بررسی آنروپی، مرتبه چهارم ژن‌ها (که نتایج آن، نسبت به مراتب دیگر قابل تامل تر می‌باشد)، مشاهده شد که آنروپی ژن‌های *C1H21orf59*، *TIMM21*، *CALM1* و *ZDHHC4* به ترتیب با مقادیر $7/8366$ ، $7/8252$ ، $7/8161$ و $7/8150$ از دیگر ژن‌ها بیشتر می‌باشد.

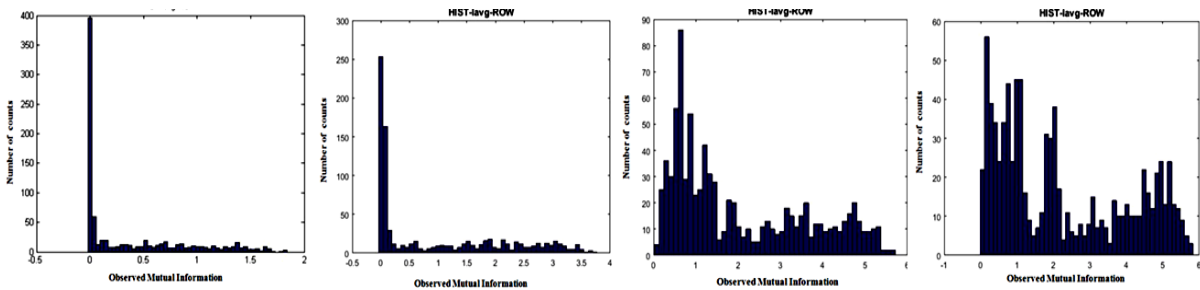
اگزون‌هایی که آنروپی آن‌ها در رتبه چهارم بیشینه بود (یعنی مقدار آنروپی نزدیک به ۸ بود) عبارت بودند از: اگزون ۱ ژن *YWHAH*، اگزون ۱ ژن *DGCR8* و اگزون ۸ ژن *TBCID20*. به نظر می‌رسد یکی از دلایلی که اگزون ۱ ژن *ACTR2* کمترین مقدار آنروپی در مراتب سوم و چهارم را در میان سایر اگزون‌های مورد بررسی از خود نشان داده است، طول کوتاه این اگزون باشد، در مقابل اگزون ۱ ژن *YWHAH* به علت طول بالاتر نسبت به سایر اگزون‌ها مقدار آنروپی بالاتری را به خود اختصاص داده است. البته این موضوع همیشه صدق نمی‌کند همانطور که در جدول ۳ مشاهده می‌شود اگزون ۸ ژن *TBCID20* که طول بیشتری نسبت به اگزون ۱ ژن *YWHAH* دارد، فقط در مرتبه سوم آنروپی مقادیر بالاتری را به خود اختصاص داده است. اکثر ژن‌های مورد بررسی دارای اگزون‌هایی بودند که مقادیر بالا و پایین آنروپی را در بر داشتند. ولی در این میان، همه اگزون‌های ژن‌های *DES* و *FAM192A* دارای آنروپی پایین تری از آنروپی سایر اگزون‌های ژن‌های مورد بررسی بودند. آنروپی مراتب ۱ الی ۴ تمام ژن‌ها، اگزون‌ها و توالی متناظر تصادفی آن‌ها در جدول ۳ فایبل ضمیمه قابل دسترس می‌باشد.

در مجموع نتایج نشان داد که مقادیر آنروپی مراتب ۱ و ۲ بسیاری از قطعه‌های DNA تفاوت چندانی با هم نداشتند و نزدیک بودن آنروپی‌ها به مقادیر حداکثر (به ترتیب ۲ و ۴) نشان می‌دهد که قطعات DNA به احتمال زیاد، دارای اطلاعات زیادی هستند. همچنین الگوهای مشابه آنروپی‌ها، نشان می‌دهد که قطعه‌های DNA دارای اطلاعات ناشناخته زیادی هستند که شاید با عملکردهای مستقل آن‌ها در ارتباط باشند. در آنروپی‌های مراتب ۳ و ۴ مشاهده شد که مقدار آنروپی محاسبه شده از حداکثر آنروپی ژن‌ها (به ترتیب ۶ و ۸) نسبت به مراتب ۱ و ۲ بیشتر فاصله می‌گیرد و کمی از حداکثر دور می‌گردند. در واقع هرچه آنروپی یک قطعه DNA بالاتر باشد، احتمال آنکه در آن بخش، یک الگوی خاص از DNA مثل نشانگر وجود داشته باشد نیز بیشتر است که پیشنهاد می‌گردد برای بررسی تنوع این قطعات در ژنوم برای آنها آغازگر طراحی و مورد بررسی قرار گیرند. با توجه به اینکه در مراتب بالاتر آنروپی، ژن‌های با آنروپی نزدیک به هم جدا از محتوا، از لحاظ توالی‌های داخل ژن نیز ممکن است نسبت به هم شبیه باشند لذا نتایج آنروپی مراتب بالاتر برای ما مهم‌تر بوده و به واقعیت نزدیک‌تر می‌باشد. چون با توجه به آنروپی مراتب ۳ و ۴ می‌توان ادعا نمود که ژن‌هایی که در کنار هم در یک خوشه قرار گرفتند جدا از محتوای اطلاعاتی نزدیک به هم، از

جدول ۴ - نتایج اطلاعات متقابل ژن‌ها و اگزون‌های بر اساس آنتروپی مراتب ۱ الی ۴

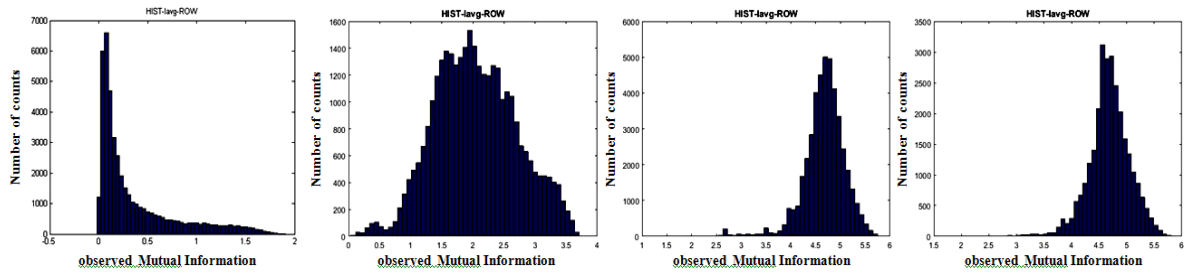
Table 4. The results of mutual information of genes and exons based on entropy orders of 1 to 4

| مقدار | نام اگزون | مقدار | نام ژن | اطلاعات متقابل |
|----------|---------------------------|----------------------------|----------|---------------------------|
| ۰/۰۸۰۹۴۶ | Exon 4 gene <i>DGCR8</i> | Exon 5 gene <i>EIF3L</i> | ۰/۰۲۷۵۵۷ | SMIM14 <i>YWHAH</i> |
| ۱/۸۸۵۲ | Exon 2 gene <i>CALM1</i> | Exon 1 gene <i>HPS6</i> | ۱/۸۲۱۷ | <i>YWHAH</i> <i>NSUN3</i> |
| ۰/۱۱۲۳۳ | Exon 5 gene <i>ZNF419</i> | Exon 1 gene <i>YWHAH</i> | ۰/۰۷۳۳۳۸ | SMIM14 <i>C1H21orf59</i> |
| ۳/۸۸۹۳ | Exon 3 gene <i>CNOT8</i> | Exon 1 gene <i>ACTR2</i> | ۳/۷۳۳۴ | <i>YWHAH</i> <i>NSUN3</i> |
| ۱/۳۹ | Exon 1 gene <i>HPS6</i> | Exon 8 gene <i>TBC1D20</i> | ۰/۰۷۲۲۰۹ | SMIM14 <i>B4GALT1</i> |
| ۵/۸۸۸۹ | Exon 1 gene <i>YWHAG</i> | Exon 1 gene <i>ACTR2</i> | ۵/۶۵۹۸ | <i>YWHAH</i> <i>NSUN3</i> |
| ۰/۸۷۵۶ | Exon 1 gene <i>HPS6</i> | Exon 1 gene <i>ACTR2</i> | ۰/۰۸۵۸ | SMIM14 <i>ACTR2</i> |
| ۷/۶۴۵۲ | Exon 1 gene <i>YWHAG</i> | Exon 3 gene <i>ACTR2</i> | ۷/۶۸۶۲ | <i>YWHAH</i> <i>NSUN3</i> |



شکل ۲- هیستوگرام فراوانی مقادیر اطلاعات متقابل ژن‌ها بر اساس آنتروپی مرتبه ۱ (چپ) تا مرتبه ۴ (راست)

Figure 2. Histogram of MI of genes due to first order entropy (left) and it goes to right which is due to fourth order entropy



شکل ۳- هیستوگرام فراوانی مقادیر اطلاعات متقابل اگزون‌ها بر اساس آنتروپی مرتبه ۱ (چپ) تا مرتبه ۴ (راست)

Figure 3. Histogram of MI of exons due to first order entropy (left) and it goes to right which is due to fourth order entropy

اصلاح نژاد داشته باشد. همچنین اساس تشخیص جایگاه‌های صفات کمی بر پایه عدم تعادل پیوستگی استوار است. بنابراین هر چقدر اطلاعات متقابل بین نوکلئوتیدها بالاتر باشد، میزان عدم تعادل پیوستگی نیز بالاتر خواهد بود. اگر آنتروپی دو قطعه DNA مشابه به هم باشند، آنگاه، اطلاعات متقابل آنها حداکثر خواهد شد و انتظار می‌رود آن دسته از قطعات DNA که چنین خاصیتی را داشته باشند یا اطلاعات متقابل آنها به حداکثر نزدیک باشد (خصوصاً در رتبه‌های بالا آنتروپی)، احتمالاً یک نقش زیستی مشابه دارند. همانطور که بیان شد، در اصلاح نژاد، شناسایی نواحی نشانگر بسیار مهم می‌باشد. اگر این نواحی اثرات عمده روی صفات تولیدی داشته باشند، پس از بررسی ارتباط ژنوتیپ‌های آن ژن با صفت تولیدی می‌توان نتایج مهمی را در این خصوص استخراج نمود. تئوری اطلاعات و معیارهای اندازه‌گیری مبتنی بر آن امکان دسته‌بندی ژن‌ها را داده که این دسته‌بندی در یک موجود در تحلیل مسیرهای متابولیکی مشترک ژن‌ها و در گونه‌های مختلف، در تحلیل مسیر تکاملی ژن‌ها روشن‌کننده راه است. بر اساس نتایج کسب شده از اجرای روند پژوهش، تئوری اطلاعات توانست در خصوص غربال ژن‌هایی که محتوای اطلاعات مشترک بالایی دارند کمک نماید. در این خصوص ژن‌ها و یا قطعاتی از ژن که از آنتروپی بالایی برخوردار بودند توانستند در شناسایی و مکان‌یابی نواحی که نوکلئوتیدهای آن از قطعیت کمتری برخوردار بوده و لذا تنوع بالایی را در افراد نشان می‌دهند کمک نمایند.

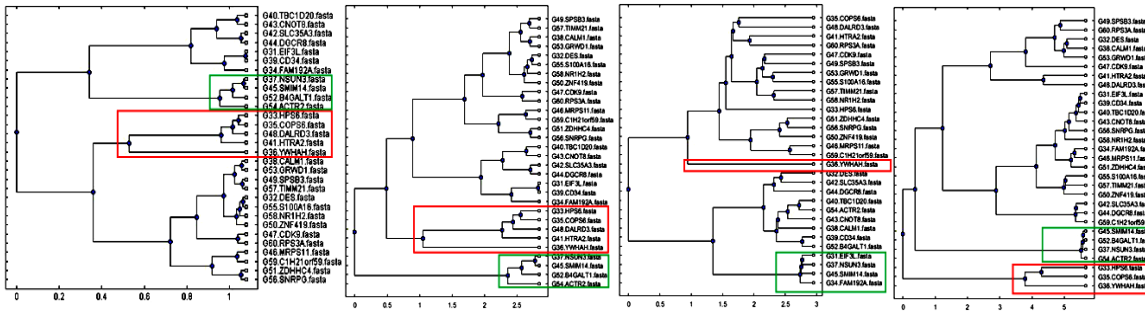
خوشه‌بندی ژن‌ها با استفاده از اطلاعات متقابل

پس از محاسبه اطلاعات متقابل، ژن‌ها در هر مرتبه از آنتروپی و بدون اعمال هم‌ترازی با ۷ روش ذکر شده، خوشه‌بندی و مورد بررسی قرار گرفتند. کلیه نتایج مربوط به خوشه‌بندی ژن‌ها و آگزون‌ها با ۷ الگوریتم یاد شده در بخش ضمیمه آورده شده است. شکل ۴ و ۵ نتایج خوشه‌بندی مراتب ۱ الی ۴ را با روش‌های UPGMA و Single را نشان می‌دهد.

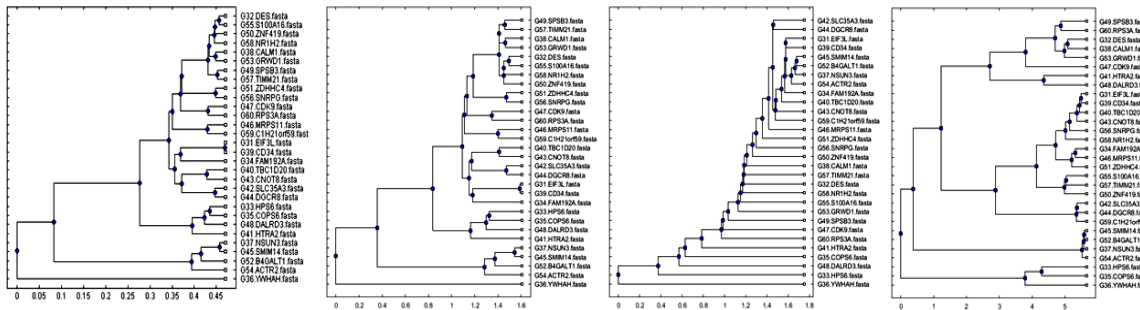
نتایج اطلاعات متقابل ژن‌ها و آگزون‌ها در جدول ۴ نشان داده شده است. در شکل ۲ و ۳ محور x مقادیر اطلاعات متقابل و محور y تعداد مقایسات را نشان می‌دهد. در این تحقیق ۳۰ ژن و ۲۱۱ آگزون بررسی شد بنابراین ماتریسی به ابعاد ۳۰×۳۰ برای ژن‌ها (شکل ۲) و ۲۱۱×۲۱۱ برای آگزون‌ها (شکل ۳) ایجاد شد. برای روشن شدن موضوع، به طور مثال در هیستوگرام سمت راست، شکل ۲ اولین میله تعداد مقایساتی که در آن مقادیر اطلاعات متقابل نزدیک به صفر بودند (حدود ۲۳ مقایسه) را نشان می‌دهد.

به طور کلی در مقایسه دو ژن با همدیگر هرچه اطلاعات متقابل آنها کمتر و نزدیک به صفر باشند یعنی دو ژن مستقل از همدیگر و هر چقدر به مقادیر حداکثر خود (مراتب مختلف) نزدیک‌تر باشند یعنی دو ژن شبیه‌تر می‌باشند.

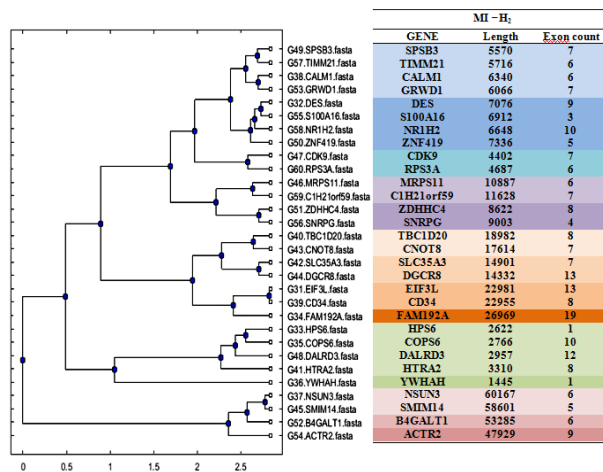
تحقیقات بسیاری با بررسی اطلاعات متقابل ژن‌ها انجام پذیرفته است. باینوالد و شاریو (۱) از اطلاعات متقابل وضعیت‌های یک توالی همدریف، به عنوان یک ویژگی برای پیش بینی ساختمان دوم RNA استفاده کردند. تومویچ و اکی (۴۷) اطلاعات متقابل را برای آنالیز نقاط اتصال فاکتور رونویسی برای شناسایی موقعیت‌هایی با همبستگی بالا به کار بردند. باسلج و همکاران (۵) شبکه‌هایی از اطلاعات متقابل بالا در ساختار جایگاه‌های کاتالیتیک را نشان دادند که برای پیش‌بینی این جایگاه‌ها از اطلاعات متقابل استفاده گردید. همچنین برونل و همکاران (۳) یک آزمون "معنی‌دار بودن آماری اطلاعات متقابل" برای مطالعات همبستگی ژنتیکی ابداع کردند. در واقع منظور از اطلاعات متقابل، میزان اشتراک اطلاعات دو منبع است. برای مثال اگر اطلاعات متقابل یک قطعه DNA با خودش را در نظر بگیریم معلوم است که اطلاعات یک قطعه DNA کاملاً با اطلاعات خودش برابر است. در این صورت اطلاعات متقابل یک قطعه DNA با خودش حداکثر مقدار اطلاعات متقابل را دارد. از لحاظ عددی مقدار اطلاعات یک قطعه DNA در مرتبه ۱ آنتروپی بین صفر و ۲ است. مقدار اطلاعات متقابل برابر صفر را می‌توان از لحاظ ژنتیک جمعیت و کمی به مفهوم نبود تعادل پیوستگی تلقی کرد. بنابراین به‌عنوان یک ایده جدید، محاسبه اطلاعات متقابل می‌تواند ارزش بسیار بالایی در



شکل ۴- نتایج خوشه‌بندی اطلاعات متقابل (MI) ژن‌ها از مرتبه ۱ (چپ) به مرتبه ۴ (راست) با استفاده از روش UPGMA
 Figure 4. The results of MI clustering of genes from first order (left) to fourth order (right) using UPGMA method



شکل ۵- نتایج خوشه‌بندی اطلاعات متقابل (MI) ژن‌ها از مرتبه ۱ (چپ) به مرتبه ۴ (راست) با استفاده از روش Single
 Figure 5. The results of MI clustering of genes from first order (left) to fourth order (right) using the single method



شکل ۶- مقایسه طول ژن‌ها بر اساس خوشه‌بندی اطلاعات متقابل آنتروپی مرتبه ۲
 Figure 6. Comparison of genes length based on MI clustering of second order entropy

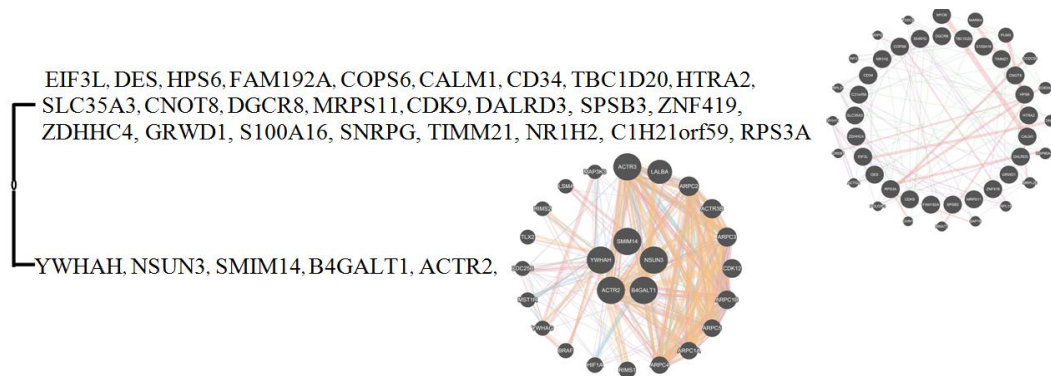
نزدیک به هم دارای قطعات مشابه بیشتری نسبت به ژن‌های سایر خوشه‌ها بوده و اطلاعات متقابل این ژن‌ها نسبت به ژن‌های خوشه‌های دیگر بیشتر می‌باشد.

یکی از مهم‌ترین پیش‌پردازش‌ها به‌منظور بهبود عملکرد سیستم طبقه‌بندی، کاهش بعد فضای ویژگی می‌باشد. کاهش بعد فضای ویژگی باعث کاهش پیچیدگی فرآیند طبقه‌بندی و در نتیجه کاهش وقوع خطا می‌شود. و یکی از روش‌هایی که برای کاهش بعد فضای ویژگی معرفی شده است، استخراج ویژگی نامید دارد. تئوری اطلاعات و معیارهای اندازه‌گیری مبتنی بر آن از جمله اطلاعات متقابل ویژگی‌های مختلفی از ارتباط ژن‌ها در اختیار ما می‌گذارد که دارای حداکثر اطلاعات در مورد کلاس خروجی می‌باشد که باعث افزایش دقت طبقه‌بندی می‌گردد.

در این بخش از پژوهش بر اساس اطلاعات مفیدی که از خوشه‌بندی ژن‌ها به دست آمد، از طبقه‌بند آدابوست استفاده شد. تجمیع و ترکیب نتایج با استفاده از طبقه‌بند آدابوست، ژن‌های مورد بررسی را در دو خوشه خوشه‌بندی نمود که خوشه اول شامل ۲۵ ژن و خوشه دوم شامل ۵ ژن بود (شکل ۷).

با تغییر مرتبه آنتروپی از ۱ به ۴ توپولوژی خوشه‌ها دچار تغییراتی شد. با توجه به شکل شماره ۴ (کادر قرمز) مشاهده شد که ژن *YWHAH* که در آنتروپی مراتب ۱ تا ۳ در یک خوشه کنار ژن‌های *DALRD3*، *COPS6*، *HPS6* و *HTRA2* قرار گرفته بود در مرتبه ۴ آنتروپی در یک خوشه دیگر تنها در کنار ژن‌های *COPS6* و *HPS6* قرار گرفت. همچنین ژن‌های *NSUN3* و *SMIM14* (کادر سبز) در همه مراتب آنتروپی در کنار هم و در یک خوشه قرار داشتند که نشان داد که اطلاعات متقابل مشترک زیادی بین این ژن‌ها وجود دارد.

نکته قابل توجه این که ژن‌های یک خوشه از نظر طول و تعداد اگزون نزدیک به هم بودند (شکل ۴). به طور مثال اگر به شکل ۵ که خوشه‌بندی ژن‌ها بر اساس اطلاعات متقابل مرتبه ۲ آنتروپی را نشان می‌دهد توجه شود کاملاً مشخص است که ژن‌های هر خوشه از نظر طول بسیار نزدیک به هم می‌باشند. در واقع می‌توان نتیجه گرفت که اطلاعات متقابل ژن‌ها به نوعی با طول ژن ارتباط مستقیم دارد. همچنین انتظار می‌رود که ژن‌هایی که در یک خوشه قرار گرفتند (به‌ویژه در مرتبه ۴ آنتروپی) که محاسبات بر اساس فراوانی جملات چهار حرفی صورت پذیرفت علاوه بر محتوای ژنی



شکل ۷- نتایج حاصل از ترکیب طبقه‌بند آدابوست روی ژن‌های مورد بررسی
Figure 7. The result of combining different clustering results AdaBoost algorithm on studied gene

ارتباط این ژن‌ها با ۲۵ ژن دیگر بودیم که در مجموع ۱۳/۵٪ هم بیانی نشان دادند و هیچ کدام از ۲۵ ژن مرتبط با ژن‌های خوشه دوم، ژن‌هایی نبودند که در خوشه اول حضور داشتند (شکل ۸). در جدول ۵ عملکرد ژن‌های خوشه‌های اول و دوم با سایر ژن‌های مرتبط با آن‌ها نشان داده شده است.

همانطور که مشاهده شد ژن‌های خوشه‌های اول و دوم هر کدام وظایف متفاوتی داشتند. وظایف ژن‌های خوشه اول بیشتر بر روی ریبوزوم و تنظیمات آن در سلول و وظایف ژن‌های خوشه دوم بیشتر بر روی فایگوسیتوزیس و سایتواسکتون بود. همچنین

بررسی ارتباط ژن‌ها و عملکردشان بر اساس نتایج حاصل از آدابوست

با توجه به ارتولوگ بودن ژن‌های مورد بررسی با انسان، ارتباط بین آن‌ها و عملکردشان در انسان با مراجعه به تارگه *GeneMANIA* مورد بررسی و پیش‌بینی قرار گرفت. در بررسی ۲۵ ژن خوشه اول در تارگه *GeneMANIA* شاهد ارتباط این ژن‌ها با ۴۵ ژن دیگر بودیم که در مجموع ۸۲/۲۲٪ هم بیانی^۱ نشان دادند. هیچ‌کدام از ۴۵ ژن مرتبط با ژن‌های خوشه اول، ژن‌هایی نبودند که در خوشه دوم حضور داشتند. در بررسی ۵ ژن خوشه دوم در *GeneMANIA* شاهد

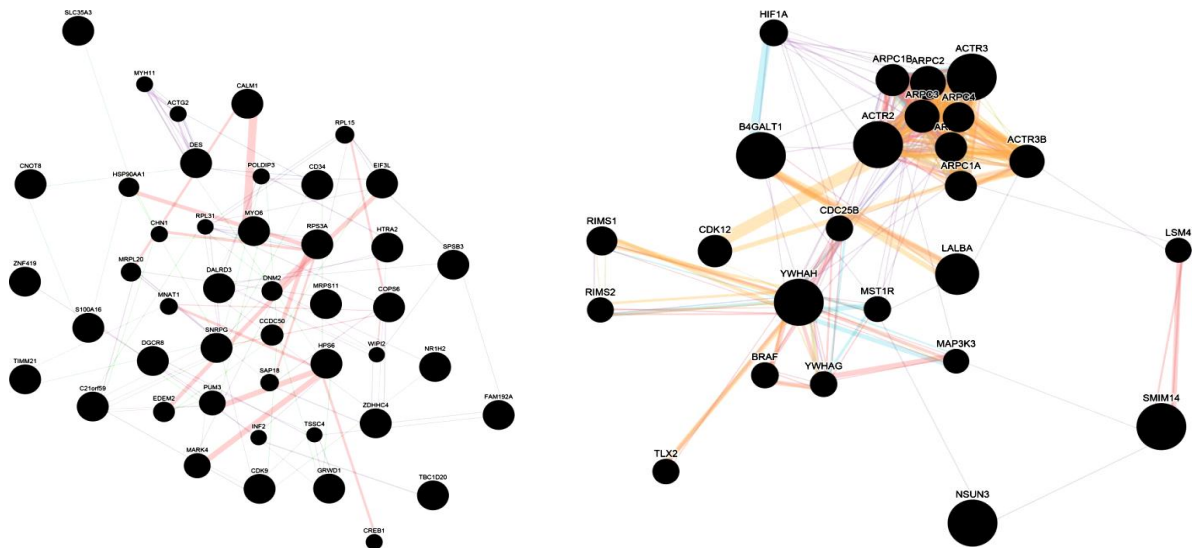
اطلاعات زیستی دیگری موجود نباشد. پژوهش‌های مختلفی در این خصوص انجام گرفته است. در پژوهشی، نظریه درختچه حیات برای تحلیل مسیرهای متابولیتی به کار گرفته شد که از اولین پژوهش‌هایی محسوب می‌شود که در آن ترکیب داده‌های سطح DNA و مسیرهای متابولیکی با استفاده از درختچه حیات انجام شد (۱۵). همچنین پژوهش‌هایی جهت استفاده هر چه بیشتر داده‌های متابولیکی برای درک بهتر ارتباط تکاملی گونه‌های مختلف (۷)، با توجه به انباشت داده‌های متابولیکی و با استفاده از نظریه گراف انجام شده است که نشان‌دهنده همخوانی درختچه فیلوژنی ایجاد شده با نتایج آزمایشگاهی بود (۱۹،۷).

هیچ ژن مشترک و مرتبط یکسانی در دو خانواده ژنی مشاهده نشد که این خود به نوعی تاییدی بر کلاسه‌بندی با روش ارائه شده بود.

نتایج حاصل از محاسبه اطلاعات متقابل روی قطعات DNA تا حدی نشان داد که ساخت و ترکیب DNA ژن‌ها، اگر به فضای دیگری نگاشت شود (مثل فضای آنترپی)، می‌تواند مشابهت‌های عملکردی آن‌ها را آشکار کند. آن دسته از توالی‌های DNA که ساختار یکسانی دارند، احتمالاً یک نوع پروتئین را کد می‌کنند و در نتیجه نقش عملکردی زیستی یکسانی خواهند داشت بنابراین امکان استخراج شبکه متابولیتی بین توالی‌های زیستی یک سازواره به طور نسبی وجود دارد. البته این در صورتی درست است که هیچ‌گونه

جدول ۵- وظایف و عملکرد ژن‌های خوشه اول و دوم

| عملکرد ژن‌های خوشه اول | عملکرد ژن‌های خوشه دوم |
|---|---|
| viral transcriptionER to Golgi vesicle-mediated transport | Fc-gamma receptor signaling pathway involved in phagocytosis |
| ribosome | immune response-regulating cell surface receptor signaling pathway involved in phagocytosis |
| regulation of nitric-oxide synthase activity | immune response-regulating cell surface receptor signaling pathway involved in phagocytosis |
| regulation of monooxygenase activity | Fc-gamma receptor signaling pathway |
| ribosomal subunit | Fc receptor mediated stimulatory signaling pathway |
| cytosolic part | Phagocytosis |
| translational initiation | actin cytoskeleton |
| regulation of oxidoreductase activity | immune response-activating cell surface receptor signaling pathway |
| cellular response to heat | Fc receptor signaling pathway |
| protein targeting | immune response-regulating cell surface receptor signaling pathway |
| | structural constituent of cytoskeleton |



شکل ۸- ارتباط ژن‌های خوشه اول شامل ۲۵ ژن (چپ) و خوشه دوم شامل ۵ ژن (راست) با سایر ژن‌های عملکردی مشابه
Figure 8. The relationship between the first cluster genes consists of 25 genes (left) and the second cluster genes including 5 genes (right) with other similar functional genes

ژن‌های هر خوشه، روش ارائه شده در این پژوهش را برای خوشه‌بندی ژن‌ها مورد تایید قرار داد.

الگوریتم یاد شده توانست در گستره‌ای از خوشه‌بندی ژن‌ها و حتی ژنوم‌ها بر اساس آن‌تروپی توالی DNA آن‌ها به کار رود. این الگوریتم از یک روش بی‌نیاز از هم‌ترازی با استفاده از اطلاعات متقابل، جهت خوشه‌بندی ژن‌ها استفاده کرد. تازگی این الگوریتم این است که با استفاده از نظریه اطلاعات، آن‌تروپی و اطلاعات متقابل توالی‌های با طول متفاوت را می‌تواند پشتیبانی و خوشه‌بندی کند. الگوریتم یاد شده تعدد نتایج خوشه‌بندی حاصل از روش‌های یاد شده را با استفاده از سامانه‌های طبقه‌بندی‌کننده چندگانه^۲ و یا سامانه‌های شورایی حل می‌کند که این پژوهش از یکی از مطرح‌ترین این روش‌ها بنام آدابوست بهره برد (۱۳،۱۴).

اعتقاد بر این است که روش ارائه شده در این مقاله می‌تواند جهت تخصیص و پیش‌بینی فعالیت زیستی آن دسته از ژن‌هایی که حاشیه‌نویسی ژنومی قوی ندارند، کمک کند، چرا که فقط متکی به توالی DNA ژن‌ها بوده و اندازه و طول ژن‌ها اثری در ماهیت الگوریتم ارائه شده ندارد. بنابراین، خوشه‌بندی توأم ژن‌هایی که حاشیه‌نویسی ژنومی قوی دارند با آن‌هایی که ندارند، می‌تواند ارزش افزوده تحلیل و زیستی به گروه دوم ژن‌ها (بدون حاشیه‌نویسی ژنومی) بدهد.

تشکر و قدردانی

نویسندگان مراتب تقدیر و تشکر صمیمانه خود را از مهندس کامیار شیوعی، دکتر سعید انصاری مهبیاری و همچنین دست‌اندرکاران مرکز ابرایانش ملی شیخ بهایی به جهت استفاده از امکانات پردازی آن مرکز اعلام می‌نمایند. این مرکز تحت حمایت معاونت علمی و فن‌آوری ریاست جمهوری و دانشگاه صنعتی اصفهان می‌باشد.

در پژوهش حاضر سعی شد که نتایج حاصل از محاسبه اطلاعات متقابل روی قطعات DNA با درختچه حاصل روی توالی قطعات DNA مقایسه شود.

پنر و همکاران (۳۴) روشی ساده و قوی برای محاسبه اطلاعات متقابل از هم‌ترازی‌های دوگانه کلی^۱ و اندازه‌گیری تشابه آنها و بازساخت درخت فیلوژنتیک ارائه دادند. یو و همکاران (۵۵) از اطلاعات متقابل به عنوان معیاری جهت اندازه‌گیری فاصله برای فیلوژنی تعدادی از مهره‌داران با استفاده از ژنوم میتوکندری آنها استفاده نمودند که نتایج حاصل تقریباً به توپولوژی موجود ارائه شده فیلوژنی مهره‌داران نزدیک بود. باید خاطر نشان کرد که در پژوهش صورت گرفته برخلاف تعدادی از مطالعات انجام شده که داده‌های ورودی آنها متعلق به گونه‌های مختلف بودند از داده ژنی یک گونه (گاو شیری) برای ایجاد درختچه تکاملی استفاده شد. بنابراین مفروضات این روش با ورودی‌های این مطالعه می‌تواند در تضاد باشد چرا که توالی‌های DNA در این پژوهش مربوط به یک ارگانسیم بود. با این وجود مشاهده شد که بنیاد نظری ایجادکننده درختچه تکاملی را می‌توان برای ارتباط متابولیکی نیز به کار برد. در واقع ماتریس فاصله ایجاد شده در روش‌های یاد شده می‌تواند ورودی الگوریتم‌های نظارت نشده مثل خوشه‌بندی سلسله‌مراتبی باشد. در آن صورت قطعات DNA که خود را به صورت خوشه نشان می‌دهند به راحتی قابل تشخیص خواهند بود. بر این اساس احتمالاً قطعاتی که در داخل یک خوشه قرار می‌گیرند در یک مسیر زیستی مشترک فعالیت دارند. در روش ارائه شده از خوشه‌بندی به یک گروه‌بندی زیستی از ژن‌ها دست یافته شد. با توجه به استخراج ویژگی‌های حاصل از نتایج خوشه‌بندی، از این روش نو و بدیع می‌توان در خوشه‌بندی ژن‌های دیگر استفاده نمود. نتایج نهایی خوشه‌بندی و بررسی عملکرد

منابع

1. Bindewald, E. and B.A. Shapiro. 2006. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers, *RNA* (2006), 12: 342-352. Published by Cold Spring Harbor Laboratory Press. Copyright 2006 RNA Society.
2. Blaisdell, B.E. 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceeding of National Academy of Sciences*, 83(14): 5155-5159.
3. Brunell, H., J.J. Gallardo-Chacon, A. Buil, M. Montserrat Vallverdu, J.M. Soria, P. Caminal and A. Perera. 2010. MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. *BIOINFORMATICS* 26(15): 1811-1818, DOI:10.1093/bioinformatics/btq273.
4. Buitenhuis, A.J., U.K. Sundekilde, N. Poulsen, H.C. Bertram, L.B. Larsen and P. Sørensen. 2013. Estimation of genetic parameters and detection of qtl for metabolites in Danish Holstein milk. *Journal of Dairy Science*, 14(79): 1-10.
5. Buslje, C.M., E. Teppa, T.D. Dome'nico, J.M. Delfino and M. Nielsen. 2010. Networks of high mutual information define the structural proximity of catalytic sites: Implications for Catalytic Residue Identification. *PLoS Computational Biology*, Volume 6(11).
6. Changchuan, Y., Y. Chen and S.T. Yau. 2014. A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. *Journal of Theoretical Biology*, 359: 18-28.
7. Clemente, J.C. K. Satou and G. Valiente. 2007. Phylogenetic reconstruction from non-genomic data. *Bioinformatics*, 23: 110-115.
8. Comin, M. and D. Verzotto. 2012. Alignment-free phylogeny of whole genomes using underlying subwords. *Algorithms for Molecular Biology*, 7(1).
9. Dawy, Z., J. Hagenauer, P. Hanus and J.C. Mueller. 2005. Mutual Information Based Distance Measures for Classification and Content Recognition with Applications to Genetics. 0-7803-8938-7/05/\$20.00 (C) 2005 IEEE.
10. Edgar, R.C. and S. Batzoglou. 2006. Multiple sequence alignment. *Curr. Opin. Struct. Biol*, 16(3): 368-373.

11. Edwards, S.V., B. Fertil, A. Giron and P.J. Deschavanne. 2002. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Systematic. Biology*, 51: 599-613.
12. Erill, I. 2012. *Information Theory and biological sequences: Insights from an evolutionary prespective*. 2012 Nova Science Publishers, Inc.
13. Freund, Y. and R. Schapire. 1996. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55: 119. CiteSeerX 10.1.1.32.8918, DOI: 10.1006/jcss.1997.1504.
14. Freund, Y. and R. Schapire. 1996. Experiments with a new boosting algorithm. Paper read at Proceeding of the Thirteenth Internatioanal Conference on Machine Learning.
15. Forst, C.V. and K. Schulten. 2001. Phylogenetic analysis of metabolic pathways. *Journal Molecular Evolution*, 52: 471-489.
16. Gray, R.M. 2013. *Entropy and Information Theory*. First Edition. Springer-Verlag New York publisher.
17. Habibi, M., H.Pezeshk, C. Eslahchi and M. Sadegi. 2007. Allocation of protein secondary structure using entropy. Iran's fifth largest biotechnology conference. Tehran, Iran. pp: 33-39 (In Persian).
18. Herzel, H., W. Ebelling and A.O. Schmitt. 1994. Entropies of biosequences: The role of repeats. *Physical Review Letters*, 50: 5061-5071.
19. Heymans, M. and A.K. Singh. 2003. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19(1): 138-146.
20. Jiang, S., C. Tang, L. Zhang and A. Zhang. 2014. A Maximum entropy approach to classifying gene array data sets. Workshop on Data Mining for Genomics, First SIAM International Conference on Data Mining.
21. Jun, S.R., G.E. Sims, G.A. Wu and S.H. Kim. 2010. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: analignment-free method with optimal featurere solution. *Proceedings of the National Academy of Sciences*, 107(1): 133-138.
22. Katoh, K., K. Misawa, K.I. Kuma and T. Miyata. 2002. Mafft: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14): 3059-3066.
23. Kemena, C. and C. Notredame. 2009. up coming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(19): 2455-2465.
24. Khatib, H., R.L. Monson, V. Schutzkus, D.M. Kohl, G.J.M. Rosa and J.J.Rutledge. 2008. Mutations in the STAT5A gene are associated with embryonic survival and milk composition in cattle. *Journal of Dairy Science*, 91: 784-793.
25. Kim, J., S. Kim, K. Lee and Y. Kwon. 2009. Entropy analysis in yeast DNA. *Chaos, Solitons and Fractals* 39: 1565-1571.
26. Larkin, M.A., G. Blackshields, N. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm and R. Lopez. 2007. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21): 2947-2948.
27. Lemay, D.G., D.J. Lynn, W.F. Martin, M.C. Neville, T.M. Casey, G. Rincon, E.V. Kriventseva, W.C. Barris, A.S. Hinrichs, A.J. Molenaar, K.S. Pollard, N.J. Maqbool, K. Singh, R. Murney, E.M. Zdobnov, R.L. Tellam, J.F. Medrano, J.B. German and M. Rijnkels. 2009. The bovine lactation genome: insights into the evolution of mammalian milk. *Genome Biology*, 10:R43 (DOI: 10.1186/gb-2009-10-4-r43).
28. Liou, C.Y., S.H. Tseng, W.C. Cheng and H.Y. Tsai. 2013. Structural complexity of DNA sequence. *Computational and mathematical methods in medicine*, Volume 2013, Article ID 628036, 11 pp.
29. Liu, B. 2007. *Uncertainty Theory*, 2nd ed., Springer-Verlag, Berlin.
30. Machado, J.T. 2012. Shannon Entropy Analysis of the Genome Code. *Hindawi Publishing Corporation Mathematical Problems in Engineering* Volume 2012, Article ID 132625, 12 pages DOI: 10.1155/2012/132625.
31. Monge, R.E. and J.L. Crespo. 2014. Comparison of complexity measures for DNA sequence analysis. 2014 International Work Conference on Bio-inspired Intelligence (IWOB).
32. Neagoe, I.M., D. Popescu and V.I.R. Niculescu. 2014. Applications of entropic divergence measures for DNA segmentation into high variable regiones of cryosporidium spp. GP60 gene. *Romanian Reports in Physics*, 66(4): 1078-1087.
33. Ogorevc, J., T. Kunej, A. Razpet and P. Dovc. 2009. Database of cattle candidate genes and genetic markers for milk production and mastitis. *Animal Genetics*, 40: 832-851.
34. Penner, O., P. Grassberger and M. Paczuski. 2011. Sequence Alignment, Mutual Information, and Dissimilarity Measures for Constructing Phylogenies. *PLOS ONE*, 6(1): e14373. DOI: 10.1371/journal.pone.0014373.
35. Pham, T.D., D.I. Crane, D. Tannock and D. Beck. 2004, Kullback-Leibler dissimilarity of markov models for phylogenetic tree reconstruction. *Proceeding of 2004 international Symposium on Inteligent Multimedia, Video and Speech Processing*. October 20-22, 2004 HongKong.
36. Porto-Díaz, L., V. BolOn-Canedo, A. Alonso-Betanzos and O. Fontenla-Rome. 2011. A study of performance on microarray data sets for a classifier based on information theoretic learning. *Neural Networks* 24: 888-896.
37. Qi, J., B.Wang and B. Hao. 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *Journal Molecular and Evolution*, 58: 1-11.
38. Reddy, Y.V. and A. Sebastian. 2009. Parameters for estimation of entropy to study price manipulation in stock markets", *Research publication university of Dehli*.
39. Ruiz-Marin, M., M. Matilla-Garcia, J.A.G. Cordoba, J.L. Susillo-Gonzalez, A. Romo-Astorga, A. Gonzalez-Pérez, A. Ruiz and J. Gayan. 2010. An entrpynetest for single-locus genetic association analysis. *BMC Genetics*, 11: 19.
40. Sims, G.E., S.R. Jun, G.A. Wu and S.H. Kim. 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*. 106(8): 2677-2682.
41. Shannon, C. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379-423 and 623-656.
42. Sherwin, B.W. 2010. Entropy and information approaches to genetic diversity and its expression: genomic geography. *Entropy*, 12: 1765-1798; DOI: 10.3390/e12071765.

43. Stuart G.W, K. Moffet and S. Baker. 2002. Integrated gene species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, 18: 100-108.
44. Stuart, G.W., K. Moffet and J.J. Leader. 2002. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Molecular Biology and Evolution.*, 19: 554-562.
45. Sundekilde, U.K., L.B. Larsen and H.C. Bertram. 2013. NMR-Based Milk Metabolomics. *Metabolites*, 3:204-222.
46. Tautz, D. and M. Trick, G.A. Dover. 1986. Cryptic simplicity in DNA is a major source of genetic variation. *Nature*, 322: 652-656.
47. Tomovic, A. and E.J Oakeley. 2007. Position dependencies in transcription factor binding sites. *Bioinformatics*, 23(8): 933-941 DOI: 10.1093/bioinformatics/btm055.
48. Vinga, S. and J. Almeida. 2003. Alignment-free sequence comparison: review. *Bioinformatics*, 19(4): 513-523.
49. Vinga, S. 2013. Information theory applications for biological sequence analysis. *Briefings in Bioinformatics*, 15(3): 376-389, DOI: 10.1093/bib/bbt068.
50. Warde-Farley, D., S.L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C.T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G.D. Bader and Q. Morris. 2010. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 2010, Vol. 38, Web Server issue DOI:10.1093/nar/gkq537.
51. Warnow, T. 2013. Large-scale multiple sequence alignment and phylogeny estimation. In: *Models and Algorithms for Genome Evolution*. Springer, 85-146pp.
52. Xie, X., Y. Yu, G. Liu, Z. Yuan and J. Song. 2010. Complexity and Entropy Analysis of DNA Methyltransferase. *J Data Mining in Genom Proteomics*, 1(2): 1000105.
53. Yu, Z.G., V. Anh and K.S. Lau. 2003. Multifractal and correlation analysis of protein sequences from complete genome, *Physical Review E*, 68: 021913.
54. Yu, Z.G, V.V. Anh and L.Q. Zhou. 2005. Fractal and dynamical language methods to construct phylogenetic tree based on protein sequences from complete genomes, in L. Wang, K. Chen and Y.S. Ong (Eds): *ICNC 2005, Lecture Notes in Computer Science*, 3612: 337-347, Springer-Verlag Berlin Heidelberg.
55. Yu, Z.G., L.Q. Zhou, V. Anh and K.H. Chu. 2007. Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from whole genome without sequence alignment, *Journal of Molecular Evolution*, 60: 538-545.
56. Zhang, J.L., L.S. Zan, P. Fang, F. Zhang, G.L. Shen and W.Q. Tian. 2008. Genetic variation of PRLR gene and association with milk performance traits in dairy cattle. *Canadian Journal of Animal Science*, 88: 33-39.
57. Zhou, L.Q., Z.G. Yu, V. Anh, P.R. Nie, F.F. Liao and Y.J. Chen. 2007. Log-correlation distance and Fourier transformation with Kullback-Leibler divergence distance for construction of vertebrate phylogeny using complete mitochondrial genomes. In *Proceedings of the 3rd International Conference on Natural Computation (ICNC2007)*, Haikou, China, August, 2007: 304-308.

Clustering of a Number of Genes Affecting in Milk Production using Information Theory and Mutual Information

Hoshang Dehghanzadeh¹, Seyed Zeaoddin Mirhoseini², Mostafa Ghaderi-Zefrehei³, Hasan Tavakoli⁴ and Saeed Esmaeilkhaniyan⁵

1- Assistant Professor, Department of Animal Science Research, Gilan Agricultural and Natural Resources Research and Education Center, AREEO, Rasht, Iran, (Corresponding author: H_dehghanzadeh@yahoo.com)

2- Professor of Genetics, Department of Animal Sciences, Faculty of Agricultural Sciences, University of Gilan, Rasht, Iran

3- Assistant Professor, Department of Animal sciences, Faculty of Agriculture, University of Yasouj, Yasouj, Iran

4- Assistance Professor, Department of Electrical Engineering, Faculty of Electrical Engineering, University of Gilan, Rasht, Iran

5- Associate Professor, Department of Animal Science Research Institute, Agricultural Research, Education and Extension Organization (AREEO), Karaj, Iran

Received: October 6, 2017 Accepted: September 22, 2018

Abstract

Information theory is a branch of mathematics. Information theory is used in genetic and bioinformatics analyses and can be used for many analyses related to the biological structures and sequences. Bio-computational grouping of genes facilitates genetic analysis, sequencing and structural-based analyses. In this study, after retrieving gene and exon DNA sequences affecting milk yield in dairy cattle, the entropy in orders one to four for each gene and eta exons was calculated. In order to extract gene distances, mutual information method was calculated. The results of mutual information of DNA and exon sequences were entered as input into 7 general clustering algorithms. In order to aggregate the results of clustering, AdaBoost algorithm was used. Finally, the results of AdaBoost algorithm were investigated by GeneMANIA prediction server to explore the results from gene annotation point of view. Integrated result of each clustering algorithm due to AdaBoost algorithm, which implied as gene tree, indicated that proposed method biologically grouped set of genes as it was proved by their gene annotation using GeneMANIA. We believe that the proposed method might be used with other DNA based clustering competitive methods and therefore, it can be used to group set of genes in other species.

Keywords: Dairy cattle, Entropy, Gene clustering, Information theory, Mutual information