نشریه پژوهش در نشخوارکنندگان
جلد پنجم، شماره سوم، ۱۳۹۶
http://ejrr.gau.ac.ir

امکان‌سنجی کاربرد آنتروپی نسبی در خوشه‌بندی تعدادی از ژن‌های مؤثر بر تولید شیر در گاو شیری

***هوشنگ دهقان‌زاده[۱]، سید ضیاءالدین میرحسینی[۲]، مصطفی قادری زفره‌یی[۳]، حسن توکلی[۴] و سعید اسماعیل خانیان[۵]**

[۱]دانشجوی دکتری ژنتیک و اصلاح‌نژاد و [۲]استاد گروه علوم دامی، [۴]استادیار گروه مهندسی برق، دانشگاه گیلان، [۳]استادیار گروه علوم دامی، دانشگاه یاسوج و [۵]دانشیار موسسه تحقیقات علوم دامی کشور، سازمان تحقیقات، آموزش و ترویج کشاورزی

تاریخ دریافت: ۹۶/۵/٤؛ تاریخ پذیرش: ۹۶/۸/۱۷

چکیده

سابقه و هدف: جدا از این‌که شیر نقش مهمی در تغذیه انسان ایفا می‌نماید، افزایش تولید شیر و یا تغییر در میزان ترکیبات آن بیشترین توجه اصلاحگران گاو شیری را به خود اختصاص داده است به همین علت پژوهش و بررسی ژن‌هایی که روی تولید و ترکیب شیر نقش مؤثری دارند، بسیار با اهمیت است. نظریه اطلاعات، شاخه‌ای از ریاضیات است که با مهندسی ارتباطات، زیست شناسی و پزشکی همپوشانی دارد. آنتروپی اندازه‌ای از عدم قطعیت در مجموعه اطلاعات است. شانون در مقاله مشهور خود در سال ۱۹٤۸ این مفهوم را معرفی کرده و نتایج آن را در تعدادی از مسائل پایه‌ای نظریه کدگذاری و انتقال داده‌ها مورد استفاده قرار داد که پایه نظریه اطلاعات جدید را تشکیل می‌دهد. از تئوری اطلاعات در تجزیه و تحلیل‌های ژنتیکی و بیوانفورماتیکی استفاده گردیده و می‌توان از آن جهت بسیاری از آنالیزهای مربوط به ساختارها و توالی‌های زیستی استفاده نمود.

مواد و روش‌ها: توالی DNA ۳۰ ژن مربوط به تولید پروتئین شیر به‌صورت جداگانه از پایگاه داده ژنوم NCBI استخراج و در فرمت FASTA ذخیره شد. در این پژوهش برای هر مجموعه ژن و اگزون‌های آن فراسنجه آنتروپی در مراتب یک الی چهار محاسبه شد. در این پژوهش برای هر مجموعه ژن و اگزون‌های آن فراسنجه آنتروپی در مراتب یک الی چهار محاسبه شد. در این راستا از زنجیره مارکف تا رتبه ۳ استفاده گردید. بر اساس آنتروپی نسبی حاصله برای ژن‌ها و اگزون‌ها، واگرایی کولبک- لیبلر برای ژن‌ها و اگزون‌ها تعریف و محاسبه گردید. سپس ماتریس واگرایی کولبک- لیبلر ژن‌ها و اگزون‌ها به‌عنوان ورودی ۷ روش معمول خوشه‌بندی Single، Complete،Average، Weighted، Centroid، Median و K-Means در نظر گرفته شد. برای تجمیع نتایج حاصل از خوشه‌بندی‌های مختلف، از الگوریتم AdaBoost استفاده گردید. در پایـــان جهت تأیید نتایج حاصل از AdaBoost و پیش‌بینی عملکرد ژن‌ها و ارتباط بین آن‌ها، با مراجعه به GeneMANIA prediction server نتایج بر اساس حاشیه‌نویسی ژنومی آن‌ها مورد بررسی و مقایسه قرار گرفت. همه محاسبات با استفاده از نرم‌افزار مهندسی متلب نسخه ۲۰۱۵ انجام گردید.

*مسئول مکاتبه: h_dehghanzadeh@yahoo.com

یافته‌ها: با بررسی نتایج در GeneMANIA prediction server، ارتباط متقابل و مسیرهای متابولیکی مشترک ژن‌ها براساس حاشیه‌نویسی ژنومی آن‌ها، روش خوشه‌بندی ارایه شده را روشی صحیح، منطقی و در عین حال سریع نشان داد. این روش علاوه‌بر این‌که زمانبر بودن حاصل از همتراز نمودن ژن‌ها را نداشته، محتوا و اندازه واقعی ژن‌ها را مورد بررسی قرار داده و نیاز به حافظه بالا برای پردازش فایل‌های همردیف توالی‌های با طول بزرگ را ندارد.

نتیجه‌گیری: نتایج نشان داد که روش پیشنهادی جهت خوشه‌بندی مجموعه‌ای از ژن‌ها به لحاظ زیستی بسیار جذاب به‌نظر می‌رسد. اعتقاد بر این است که روش ارائه شده می‌تواند با سایر روش‌ها از جهت خوشه‌بندی مجموعه‌ای از ژن‌ها رقابت نماید. روش یاد شده می‌تواند به‌عنوان یک روش پیش‌بینی عملکرد زیستی ژن‌هایی با داده‌های حاشیه‌نویسی ژنومی ضعیف نیز در نظر گرفته شود.

واژه‌های کلیدی: تئوری اطلاعات، واگرایی کولبک- لیبلر، گاو شیری، خوشه‌بندی ژن

مقدمه

طبق آخرین آمار رسمی وزارت جهاد کشاورزی، تعداد ۱۸۸۳۰ واحد صنعتی گاو داری با ظرفیت ۲۰٤۸٥٦۳ راس گاو شیرده در کشور مشغول فعالیت هستند. طی سال‌های ۱۳۸۳ تا ۱۳۸۷ تولید شیر دارای یک روند رو به رشد بوده است (۱٦). با وجود روند افزایشی تولید شیر در کشور اما هنوز سرانه مصرف شیر از حد استاندارد جهانی پایین تر است. سرانه مصرف شیر در کشور برای هرنفر برابر با ۹۵ کیلوگرم می‌باشد، در حالی‌که سرانه مصرف شیر در جهان برابر با ۱٦۹ کیلوگرم و در اروپا برابر با ۳۵۰ کیلوگرم در سال است (۱٦). با توجه به آمار و اطلاعات موجود می‌توان دریافت که اهداف اصلاح‌نژادی در ایران بایستی برای افزایش تولید شیر در کشور برنامه‌ریزی شود. لذا مطالعه و بررسی ژن‌هایی که روی تولید و ترکیب شیر نقش مؤثری دارند اهمیتی دو چندان می‌یابد (۱٦). به‌علاوه، در حوزه ژنتیک و اصلاح، اطلاع از ساختار ژنتیکی جمعیت‌ها می‌تواند کمک بزرگی برای برنامه‌ریزی برای طرح‌های اصلاح نژادی و از همه مهمتر، حفظ ذخایر ژنتیکی باشد. روش‌های مولکولی و استفاده از نشانگرهای مولکولی در این زمینه یکی از بهترین گزینه‌ها به حساب می‌آید، زیرا با توجه به اطلاعات زیادی که به‌دست می‌دهد می‌تواند نتایجی که از تجزیه و تحلیل رکوردها با روش‌های آماری به‌دست آمده است را تأیید و تکمیل نموده و حتی ممکن است که آن‌ها را رد کند (۲). به‌علاوه، استفاده از ژنتیک ملکولی فواید زیادی دارد که یکی از این فواید معنی‌دار تعیین ژنوتیپ افراد برای جایگاه خاصی است (۲۷، ۲۸). همچنین استفاده از نشانگرهای ژنتیکی در انتخاب و اصلاح نژاد حیوانات ممکن است به‌طور مهیجی پیشرفت ژنتیکی را تسریع کند (۱۲) و مطالعه ساختار ژنتیکی نژادها برای حفاظت از منابع ژنتیکی لازم و ضروری است (۲٦). حفاظت باید بر اساس دانش عمیقی از منابع ژنتیکی نژادها باشد، لذا تلاش برای شناسایی و تعیین خصوصیات ژنتیکی نژادها بسیار اهمیت دارد (۳٦، ٤۳).

بسیاری از صفات اقتصادی که دربرگیرنده صفات تولیدی هستند از جمله صفات ترکیبات شیر، تحت کنترل تعداد زیادی ژن قرار دارند که به دنبال آن تعیین چندشکلی ژن‌های کاندیدای مؤثر بر صفات تولیدی و شناسایی آلل‌ها و ژنوتیپ‌های مطلوب برای صفات موردنظر می‌تواند زمینه را برای انتخاب به کمک نشانگر فراهم کند (۱٤، ۱۵).

جدا از این‌که شیر نقش مهمی در تغذیه انسان ایفا می‌نماید، افزایش تولید شیر و یا تغییر در میزان ترکیبات آن بیشترین توجه اصلاحگران گاو شیری را به خود اختصاص داده است به همین علت پژوهش و بررسی ژن‌هایی که

روی تولید و ترکیب شیر نقش مؤثری دارند، بسیار با اهمیت است و می‌تواند گامی مهم در جهت شناسایی و توسعه انتخاب به کمک نشانگر[1] و تدوین برنامه‌های اصلاح‌نژادی جهت بهبود صفات تولیدی به شمار آید (۱، ۱۷ و ۳۷). در طی سالیان گذشته سامانه داده‌های زیستی با نرخ بسیار بالایی تولید و پایگاه‌های اطلاعاتی جهت ثبت و پذیرش و نگهداری توالی[2] ژن‌ها و پروتئین‌های مختلف جانداران ایجاد گردیدند. در اصلاح‌نژاد 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[3] بررسی و با تفسیر زیستی نتایج حاصله، رهیافت جدیدی برای افزایش تولید شیر و یا دستکاری‌های ژنی با اهداف متفاوت را ایجاد کرد.

امروزه ما شاهد ظهور ابزارها و الگوریتم‌های رایانشی[4] جدید جهت آزمایش و فرموله کردن فرضیه‌هایی همچون چگونگی سازماندهی و تکامل ژنوم و رویت یک فنوتیپ مشخص از یک ژنوم رمز نگاری شده هستیم. نظریه اطلاعات که شاخه‌ای از ریاضیات است و با مهندسی ارتباطات، زیست‌شناسی و پزشکی همپوشانی دارد نقش مهمی در این زمینه بازی می‌کند. این نظریه که در سال ۱۹٤۸ توسط کلود شانون ارائه شد، به کشف و بررسی قوانین ریاضی حاکم بر رفتار داده‌ها در مراحل انتقال، ذخیره و بازیابی داده‌ها می‌پردازد (۲٤). آنتروپی[5] شانون هسته اصلی نظریه اطلاعات می‌باشد و گاهی اوقات تحت عناوینی مثل اندازه عدم قطعیت یا میزان تصادفی بودن[6]، درهم ریختگی[7] و پیش‌بینی‌ناپذیری[8] 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 موقعیّتی کاملاً قابل پیش‌بینی است، هیچ اطلاعاتی در مورد آن وجود ندارد. این وضعیّت را استحکام (نگو آنتروپی[9]) می‌گویند (۳٤). واحد آنتروپی معمولا بیت[10] 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 (۱۰). از نظریه اطلاعات به‌عنوان ابزاری مهم و به چند صورت برای جستجوی الگوهایی در توالی‌های DNA (۳۸)، نقش آمینواسیدها در ساختار پروتئین‌ها در مخمر (۱۸)، تحلیل جایگاه‌های صفات کمی[11] و اپیستاسیس[12] (۳۳)، بررسی اطلاعات ژنوم جهانی[13] (۲۵)؛ تحلیل داده‌های ریزآرایه DNA (۱۳)، طبقه‌بندی ژن‌های درگیر در سرطان (۳۲)، مقایسه اندازه پیچیدگی برای آنالیز توالی‌های DNA (۱۰، ۲۳ و ۲۹)، بازساخت درختان فیلوژنتیکی بدون همردیف کردن بازها (۳۱)، پژوهش‌های تکاملی (۵)، تنوع ژنتیکی (۳۵)، مقایسه محتوای اطلاعات

1- Marker-assisted selection
2- Sequence
3- Information theory
4- Computational
5- Entropy
6- Randomness
7- Disorderliness
8- Unpredictability
9- Nego Entropy
10- Bit
11- QTL
12- Epistasis
13- Global genomic information

نواحی اینترون و اگزون ژن‌ها (٤٢) آنالیز جزایر CpG ژنوم (٣) و تحلیل زیر گونه‌های انگل کریپتوزپوریدیوم[1] (٣٠) استفاده شده است.

در نظریه احتمالات و نظریه اطلاعات، واگرایی کولبک- لیبلر[2]- یا به‌عبارتی آنتروپی نسبی[3]- یک معیار نامتقارن برای اندازه‌گیری تفاوت دو توزیع احتمالاتی Q و P می‌باشد. این واگرایی یک نمونه خاص از دسته وسیع‌تری از واگرایی‌ها به نام واگرایی اف[4] می‌باشد (٢٢). این واگرایی برای اولین بار توسط سولومون کولبک و ریچارد لیبلر (١٩٥١) به عنوان یک واگرایی جهت دار بین دو توزیع معرفی گردید (١٩). ماهیت و ساخت نظری این معیار کاربردهای زیادی را در عمل برای آن در حوزه‌های مختلف متصور ساخته است. مقاله حاضر کاربرد الگوریتمی متکی به واگرایی کولبک- لیبلر را نشان می‌دهد که نویسندگان مقاله برای اولین بار جهت خوشه‌بندی تعدادی از ژن‌های مؤثر روی تولید شیر ارایه کردند. الگوریتم یاد شده می‌تواند در گستره‌ای از خوشه‌بندی[5] ژن‌ها و حتی ژنوم‌ها بر اساس آنتروپی توالی DNA آن‌ها به‌کار رود. این الگوریتم از یک روش بی نیاز از همترازی[6] با استفاده از واگرایی کولبک- لیبلر که مبتنی بر آنتروپی ژن‌ها می‌باشد، جهت خوشه‌بندی ژن‌ها استفاده می‌کند. تازگی این الگوریتم این است که با استفاده از نظریه اطلاعات و آنتروپی نسبی، 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 دسته‌بندی کننده چندگانه[7] و یا سامانه‌های شورایی حل می‌کند. یکی از مطرح‌ترین این روش‌های شورایی آدابوست[8] می‌باشد (٧) که این پژوهش از آن بهره می‌برد. تاکنون، بر اساس دانش نویسندگان، هیچگونه پژوهشی ژن‌های مؤثر بر تولید شیر را با استفاده از نظریه اطلاعات خوشه‌بندی نکرده است. انتظار می‌رود که استخراج الگوهای ژنی حاصل از این خوشه‌بندی بتواند در کنکاش‌های زیستی، دارویی و اصلاح‌نژادی به‌کار رود.

مواد و روش‌ها

استخراج توالی‌های DNA ژن‌ها: گزارش گردیده است که در کل، حدود ٦٨٧٥ ژن وجود دارد که در تولید شیر مؤثر 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 (٢١). ژن‌های مورد بررسی در این مقاله از نتایج پژوهش لیمی و همکاران (٢٠٠٩) انتخاب گردیدند (٢١). در این مقاله ٣٠ ژن از تعداد ٣٨٨٩ ژن دسته ژن‌های پستانی مؤثر در تولید شیر مربوط به گاوهای تلیسه[9] (٢١) به‌طور تصادفی انتخاب و مورد بررسی و واکاوی قرار گرفتند. توالی و همچنین سایر اطلاعات ژن‌ها از جمله اندازه هر ژن، محتوای C-G، شماره دست‌یابی[10]، تعداد و طول هر اگزون[11] و جایگاه آن بر روی

1- Cryptosporidium
2- Kullback-Leibler divergence
3- Relative entropy
4- F-divergence
5- Clustering
6- Free-Alignment
7- Multiple classifier system
8- AdaBoost
9- Virgin mammary gene set
10- Accession number
11- Exone

کروموزوم از بانک ژنی NCBI[1] دریافت و سپس با پیکربندی فستا[2] ذخیره گردیدند (جداول ۱ و ۲ فایل ضمیمه). جهت آماده‌سازی اطلاعات استخراج شده از پایگاه داده به‌دلیل زیاد بودن حجم اطلاعات ژن‌ها و اگزون‌های مربوط به آن، نرم‌افزاری طراحی شد که به‌طور هوشمند، ویژگی‌های ژن‌ها را استخراج می‌کرد. لذا، در این نرم‌افزار با توجه به خواسته پژوهش، خروجی‌های مناسب به‌دست آمدند. برای ایجاد این این نرم‌افزار از زبان برنامه‌نویسی C# استفاده شد.

محاسبه مراتب[3] آنتروپی: 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[4]) از فرمول زیر استفاده شد:

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

که در آن p_i احتمال i^{th} نوکلئوتید از مجموعه {A, T, G, C} در زنجیره DNA می‌باشد. در این نوع از آنتروپی فرض شد که ظاهر شدن هر نوکلئوتید، مستقل از نوکلئوتید دیگر در رشته DNA می‌باشد و به نوع نوکلئوتید مجاورش بستگی ندارد.

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

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

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) 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 {A، T، G، C} در توالی DNA ژن است.

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

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

که i، j و k نشانگر آگاهی از وقوع ۳ نوکلئوتید قبلی است و $p_{i,j,k}(m)$ هم احتمال نوکلئوتید m به شرط وقوع نوکلئوتیدهای k ,i و j از مجموعه {A، T، G، C} در توالی DNA ژن است. همان‌طور که نشان داده شده در کل ٤ 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 توالی نیز محاسبه شد تا میزان تصادفی بودن توالی ژن مورد مقایسه قرار گیرد (جدول ۱). درضمن، اندیسی که در H ظاهر می‌شود نشان دهنده مرتبه آنتروپی موردنظر است.

1- http://www.ncbi.nlm.nih.gov/genbank/gene
2 -Fasta
3- Orders
4- Markov chain

اندازه‌گیری واگرایی کولبک- لیبلر: جهت محاسبه واگرایی کولبک- لیبلر از فرمول زیر استفاده شد:

$$D_{KL}(P(x)||Q(x)) = \sum_{i=1}^{n} P(x)\log_2\frac{P(x)}{Q(x)}$$

که n تعداد نوکلئوتید در یک رشته DNA و $P(x), Q(x) \neq 0$ D_{KL} یک معیار متقارن نیست و یک فاصله حقیقی نمی‌باشد، بنابراین:

$$D_{KL}(P(x)||Q(x)) = D_{KL}(Q(x)||P(x))$$

در این پژوهش D_{KL} به‌صورت زیر استفاده شد:

$$\frac{[D_{KL}(Q(x)||P(x)) + D_{KL}(P(x)||Q(x))]}{2}$$

این روش که برای سهولت استفاده در متن از آن به‌عنوان KL_H می‌شود بر پایه مقادیر آنتروپی ژن‌ها و اگزون‌ها می‌باشد. آنتروپی دو توالی ژنی مورد بررسی به‌طور مجزا محاسبه که مقدار عددی توالی اول به‌عنوان P و مقدار عددی توالی دوم به‌عنوان Q در فرمول جاگذاری و محاسبه گردید. برای هر ژن و اگزون‌های آن به‌طور مجزا آنتروپی‌های مراتب ۱ الی ٤ سپس آنتروپی نسبی آن‌ها با این روش محاسبه شد. در این روش یک ماتریس نامتقارن به اندازه تعداد ژن‌ها و یا اگزون‌های مورد بررسی ایجاد شد که بر این اساس ژن‌ها و اگزون‌هایی که بیشترین شباهت و بیشترین فاصله را از هم داشتند مشخص گردید. قادری و همکاران شکل ساده‌ای از این معیار را برای پیدا کردن فاصله بین کانتیگ‌های ژنوم اشرشیا کلی مؤثر بر ورم پستان به کار بردند (۹).

ترکیب نتایج حاصل از انواع روش‌های خوشه‌بندی: معیار به‌دست آمده فاصله کولبک- لیبلر در مجموعه ژن‌ها و اگزون‌ها، به‌عنوان ورودی ۷ روش معمول 'Single'[۱]، 'Complete'[۲]، 'Average'[۳]، 'Weighted'[٤]، 'Centroid'[٥]، 'Median'[٦] و 'KMeans' به‌کار رفتند و خوشه‌بندی ژن‌ها به‌دست آمدند. در این مقاله تنها نتایج خوشه‌بندی حاصل از روش Single آورده شده و بقیه در فایل ضمیمه قابل مشاهده می‌باشد. برای ترکیب نتایج خوشه‌بندی ایجاد شده حاصل از ۷ الگوریتم بالا و ترکیب نتایج خوشه‌بندی‌ها، از الگوریتم AdaBoost برای ترکیب نتایج خوشه‌بندی استفاده شد (۷). در پایان، جهت تأیید نتایج حاصل از AdaBoost و بررسی همخوانی نتیجه خوشه‌بندی ژن‌ها با داده‌های حاشیه‌نویسی ژنوم آن‌ها، از GeneMANIA prediction server[۷] استفاده شد (۳۱). همه محاسبات با استفاده از نرم‌افزار متلب[۸] (۲۰۱۵) انجام گردید. برای محاسبات KL_H از یک رایانه شخصی با پردازشگر intel Core i5 CPU 2.40 GHz و حافظه ٤ گیگا بایت استفاده شد.

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 (به‌ترتیب با طول ٦۰۱٦۷ و ۱٤٤٥) 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

1- Nearest distance (single linkage method)
2- Furthest distance (complete linkage method)
3- Unweighted pair group method average (UPGMA, group average)
4- Weighted pair group method average (WPGMA)
5- Unweighted pair group method centroid (UPGMC)
6- Weighted pair group method centroid (WPGMC)
7- http://www.genemania.org
8- Matlab engineering software

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 درهر رتبه در جدول ۱ آمده است.

جدول ۱- آنتروپی محاسبه شده مراتب مختلف و آنتروپی تصادفی متناظرشان در توالی DNA ژن های موثر در تولید پروتئین شیر گاو.

Table 1. Calculated entropy with different orders and their corresponding random entropies in cow's milk protein governing genes.

$H(x)_{VI}/RH(x)_{VI}$	$H(x)_{III}/RH(x)_{III}$	$H(x)_{II}/RH(x)_{II}$	$H(x)_{I}/RH(x)_{I}$	Gene symbol	No
7.7993/7.9920	5.8699/5.9980	3.9327/3.9994	1.9871/2.0000	EIF3L	1
7.7275/7.9677	5.8338/5.9929	3.9176/ 3.9986	1.9878/1.9991	DES	2
7.5078/ 7.9056	5.7134/5.9842	3.8513/ 3.9917	1.9617/1.9994	HPS6	3
7.6626/ 7.9928	5.7688/5.9979	3.8667/3.9996	1.9566/1.9999	FAM192A	4
7.7404/7.9375	5.8660/5.9701	3.9388/3.9973	1.9931/1.9999	COPS6	5
7.7365/7.8605	5.9045/5.9527	3.9649/3.9940	1.9994/1.9983	YWHAH	6
7.6681/7.9966	5.7703/5.9990	3.8665/3.9998	1.9551/2.0000	NSUN3	7
7.8161/ 7.9734	5.8955/5.9910	3.9504/ 3.9984	1.9913/1.9998	CALM1	8
7.7877/7.9926	5.8681/5.9975	3.9404/ 3.9996	1.9974/1.9999	CD34	9
7.7917/7.9900	5.8681/5.9974	3.9345/ 3.9995	1.9899/1.9998	TBC1D20	10
7.7743/ 7.9315	5.8759/5.9837	3.9364/ 3.9966	1.9872/1.9997	HTRA2	11
7.5880/7.9857	5.7152/5.9962	3.8293/3.9994	1.9332/1.9995	SLC35A3	12
7.7304/7.9881	5.8216/5.9971	3.9033/3.9994	1.9736/2.0000	CNOT8	13
7.6828/7.9848	5.7941/5.9971	3.8899/3.9990	1.9666/1.9999	DGCR8	14
7.7081/7.9961	5.7988/5.9993	3.8839/3.9998	1.9598/1.9999	SMIM14	15
7.7967/7.9817	5.8769/5.9946	3.9417/3.9984	1.9945/1.9998	MRPS11	16
7.7677/7.9596	5.8663/5.9891	3.9344/3.9982	1.9889/1.9991	CDK9	17
7.6230/7.9247	5.7672/5.9798	3.8711/ 3.9960	1.9585/1.9995	DALRD3	18
7.5244/7.9645	5.6846/5.9926	3.8173/3.9979	1.9390/1.9998	SPSB3	19
7.7540/7.9687	5.8516/5.9939	3.9284/3.9981	1.9925/1.9997	ZNF419	20
7.8150/7.9746	5.8876/5.9941	3.9435/3.9970	1.9863/2.0000	ZDHHC4	21
7.7756/7.9964	5.8552/5.9991	3.9296/3.9999	1.9949/2.0000	B4GALT1	22
7.6984/7.9656	5.8109/5.9895	3.9017/3.9971	1.9855/1.9997	GRWD1	23
7.6828/7.9954	5.7813/5.9990	3.8736/3.9998	1.9572/2.0000	ACTR2	24
7.6163/7.9745	5.7561/5.9927	3.8754/3.9990	1.9835/1.9996	S100A16	25
7.7089/7.9768	5.8093/5.9943	3.8946/3.9983	1.9683/1.9998	SNRPG	26
7.8252/7.9611	5.9043/5.9916	3.9570/3.9974	1.9946/1.9996	TIMM21	27
7.6539/7.9692	5.7820/5.9927	3.8845/3.9979	1.9769/1.9998	NR1H2	28
7.8366/7.9816	5.9041/5.9954	3.9576/3.9994	1.9991/2.0000	C1H21orf59	29
7.7327/7.9597	5.8295/5.9872	3.9038/3.9978	1.9668/1.9997	RPS3A	30

* سلول‌های با رنگ خاکستری و نارنجی به‌ترتیب بیشترین و کمترین مقادیر آنتروپی ژن‌ها را در رتبه مربوطه نشان می‌دهد.

در جدول ۲، نتایج آنتروپی کمینه و بیشینه در مراتب ۱ الی ٤ ژن‌های مربوطه، نشان داده شده است.

جدول ۲- نتایج آنتروپی کمینه و بیشینه در مراتب ۱ الی ٤ ژن‌های مربوطه.

Table 2. Results of maximum and minimum entropy orders of 1 to 4 in respected genes.

آنتروپی	کمترین مقدار آنتروپی Minimum entropy			بیشترین مقدار آنتروپی Maximum entropy		
	نام ژن Name of genes	طول ژن Lenght of genes	مقدار Value	نام ژن Name of genes	طول ژن Lenght of genes	مقدار Value
$H(x)_I$	SLC35A3	14901	1.9332	YWHAH	1445	1.9994
$H(x)_{II}$	SPSB3	5570	3.8173	YWHAH	1445	3.9649
$H(x)_{III}$	SPSB3	5570	5.6846	YWHAH	1445	5.9045
$H(x)_{IV}$	HPS6	2622	7.5078	TIMM21	5716	7.8252

با بررسی آنتروپی مرتبه چهارم ژن‌ها (که نتایج آن نسبت به مراتب دیگر قابل تامل‌تر می‌باشد)، مشاهده شد که آنتروپی ژن‌های C1H21orf59، TIMM21، ZDHHC4 و CALM1 به‌ترتیب با مقادیر ۷/۸۳٦٦، ۷/۸۲۵۲، ۷/۸۱۵۰ و ۷/۸۱٦۱ از دیگر ژن‌ها بیشتر می‌باشد. در جدول ۳، نتایج آنتروپی کمینه و بیشینه اگزون‌های ژن‌ها از مرتبه ۱ الی ٤ ارائه شده است.

جدول ۳- نتایج آنتروپی بیشینه و کمینه در مراتب ۱ تا ٤ اگزون‌های ژن‌ها.

Table 3. Results of different entropy orders of 1 to 4 over gene exones.

آنتروپی	کمترین مقدار آنتروپی Minimum entropy			بیشترین مقدار آنتروپی Maximum entropy		
	نام ژن Name of genes	طول ژن Lenght of genes	مقدار Value	نام ژن Name of genes	طول ژن Lenght of genes	مقدار Value
$H(x)_I$	Exon 1 gene SPSB3	34	1.6457	Exon 1 gene YWHAH	1445	1.9994
$H(x)_{II}$	Exon 1 gene SPSB3	34	2.9220	Exon 1 gene YWHAH	1445	3.9649
$H(x)_{III}$	Exon 1 gene ACTR2	10	2.7500	Exon 8 gene TBC1D20	2286	5.8458
$H(x)_{IV}$	Exon 1 gene ACTR2	10	2.8074	Exon 1 gene YWHAH	1445	7.7365

اگزون‌هایی که آنتروپی آن‌ها در رتبه چهارم بیشینه بود (یعنی مقدار آنتروپی نزدیک به ۸ بود) عبارت بودند از: اگزون ۱ ژن YWHAH، اگزون ۱ ژن DGCR8 و اگزون ۸ ژن TBC1D20.

به‌نظر می‌رسد یکی از دلایلی که اگزون ۱ ژن ACTR2 کمترین مقدار آنتروپی 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 به‌علت طول بالاتر نسبت به سایر اگزون‌ها مقدار آنتروپی بالاتری را به خود اختصاص داده است. البته این موضوع همیشه صدق نمی‌کند همان‌طور که در جدول ۳ مشاهده می‌شود اگزون ۸ ژن TBC1D20 که طول بیشتری نسبت به اگزون ۱ ژن YWHAH دارد، فقط در مرتبه سوم آنتروپی مقادیر بالاتری را به خود اختصاص داده است. اکثر ژن‌های مورد بررسی دارای اگزون‌هایی بودند که مقادیر بالا و پایین آنتروپی را در بر داشتند. ولی در این میان، همه اگزون‌های ژن‌های DES و FAM192A دارای آنتروپی پایین‌تری از آنتروپی سایر اگزون‌های ژن‌های مورد بررسی بودند. آنتروپی مراتب ۱ الی ٤ تمام ژن‌ها، اگزون‌های و توالی متناظر تصادفی آن‌ها در جدول ۳ فایل ضمیمه قابل دسترس می‌باشد.

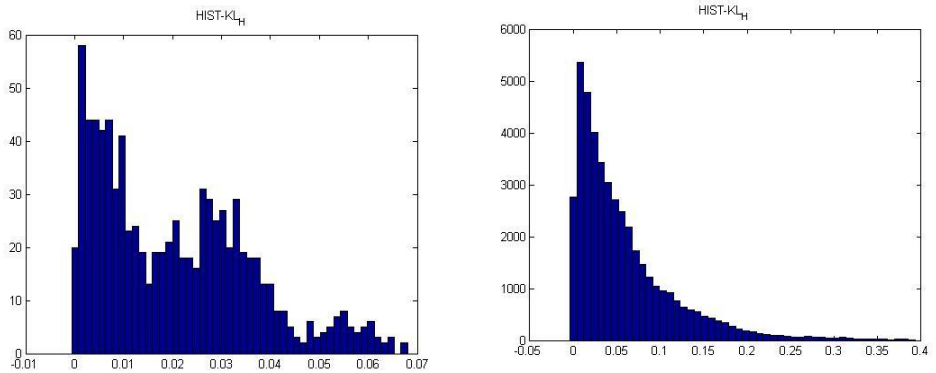
خوشه‌بندی ژن‌ها با استفاده از واگرایی کولبک- لیبلر: در ابتدا آنتروپی مراتب یک الی چهارم در ژن‌ها و اگزون‌ها محاسبه و سپس مقادیر کولبک- لیبلر برای هر مرتبه از آنتروپی به‌طور جداگانه محاسبه شد. این روش به ژن‌ها و اگزون‌ها این اجازه را می‌دهد که با طول حقیقی و متفاوت و محتوای واقعی خود نسبت به یکدیگر مورد ارزیابی قرار گیرند. نتایج این بخش در جدول ٤ ارائه شده است. سپس مرحله خوشه بندی ژن‌ها و اگزون‌ها بر پایه آنتروپی نسبی شان و بدون اعمال همترازی انجام شد. کلیه نتایج مربوط به خوشه‌بندی ژن‌ها و اگزون‌ها با ۷ الگوریتم یاد شده در بخش ضمیمه آورده شده است.

جدول ٤- نتایج کولبک لیبلر (آنتروپی نسبی) ژن‌ها و اگزون‌های بر اساس آنتروپی مراتب ۱ الی ٤.

Table 2. The results of KL_H (relative entropy) of genes and exons based upon entropy orders of 1 to 4.

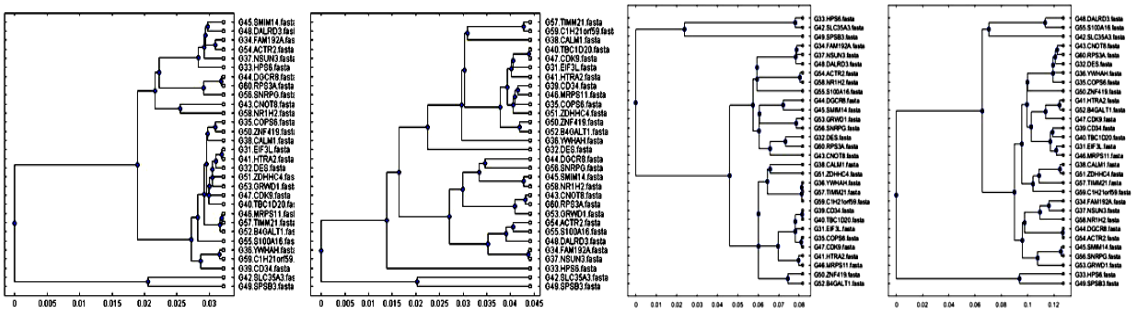
مقدار	نام اگزون		مقدار	نام ژن		KL_H	
1.4359e-006	Exon 7 gene CD34	Exon 9 gene COPS6	0.00011152	HTRA2	EIF3L	کمترین فاصله Min Distance	$H(x)_I$
0.38926	Exon 1 gene PSB3	Exon 1 gene YWHAH	0.067366	SLC35A3	YWHAH	بیشترین فاصله Max Distance	
5.19e-005	Exon 3 gene SNRPG	Exon 8 gene DES	8.8164e-005	TBC1D20)	CDK9	کمترین فاصله Min Distance	$H(x)_{II}$
1.4861	Exon 1 gene ACTR2	Exon 1 gene YWHAH	0.15033	SPSB3	YWHAH	بیشترین فاصله Max Distance	
3.1074e-005	Exon 9 gene EIF3L	Exon 3 gene CNOT8	2.8624e-005	TBC1D20	CD34	کمترین فاصله Min Distance	$H(x)_{III}$
4.5117	Exon 1 gene ACTR2	Exon 1 gene YWHAH	0.22413	SPSB3	YWHAH	بیشترین فاصله Max Distance	
1.2944e-015	Exon 4 gene TIMM21	Exon 7 gene EIF3L	1.7079e-005	ACTR2	DGCR8	کمترین فاصله Min Distance	$H(x)_{IV}$
7.8426	Exon 1 gene ACTR2	Exon 1 gene YWHAH	0.33595	HPS6	C1H21orf59	بیشترین فاصله Max Distance	

هیستوگرام فراوانی مقادیر کولبک- لیبلر (آنتروپی نسبی) ژن و اگزون‌ها بر اساس آنتروپی مراتب ۱ الی ٤ در جداول ضمیمه موجود می‌باشد. به‌عنوان نمونه در شکل ۱ هیستوگرام فراوانی مقادیر آنتروپی ژن‌ها و اگزوهای‌شان بر اساس مرتبه ۱ مشاهده می‌شود.



شکل ۱- هیستوگرام فراوانی مقادیر کولبک- لیبلر (آنتروپی نسبی) ژن‌ها (سمت چپ) و اگزون‌ها (سمت راست) بر اساس آنتروپی مرتبه ۱ (همان‌طور که مشاهده می‌گردد حدود ۲/۲ درصد از برآوردها در ژن‌ها (حدود ۲۰ مشاهده) و ٦/٦ درصد در اگزون‌ها (حدود ۱۸۰۰ مشاهده) کاملاً شبیه هم بوده و در مقادیر نزدیک صفر قرار گرفتند).

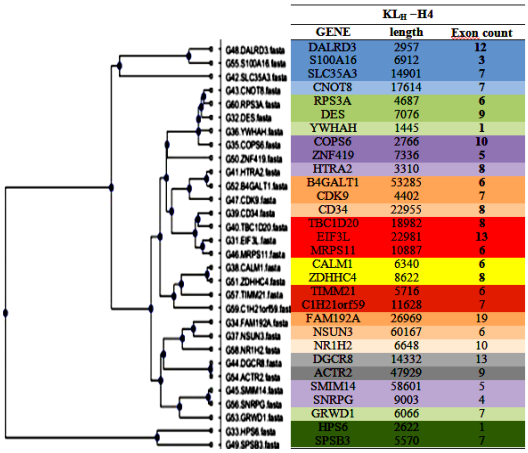
Figure 1. Frequency histogram of KL_H for genes (left) and exons (right) based on first order relative entropy. About 2.2% of data (20 data points) in genes and 6.6% of data (1800 data point) in exon were almost identically lumped around zero value.



شکل ۲- نتایج خوشه‌بندی کولبک- لیبلر مبتنی‌بر آنتروپی نسبی (KL_H) ژن‌ها از مرتبه ۱ (چپ) به مرتبه ٤ (راست) با استفاده از روش Single.

Figure 2. Results of kullback-leibler clustering based on relative entropy (KL_H) of genes from the first order (left) to the fourth order (right).

در کل و در مجموع تعداد ۲۸ خوشه ایجاد گردید (برای هر مرتبه از آنتروپی ۷ الگوریتم استفاده شد). با مقایسه و بررسی همه خوشه‌های ایجاد شده با روش Single مشاهده گردید که ژن‌ها به دو گروه اصلی تفکیک شدند که گروه دوم در همه به جز خوشه‌بندی که با مرتبه سوم آنتروپی شکل گرفته و شامل ۳ ژن است، از دو ژن تشکیل گردیده است. ژن SPSB3 در همه این چهار خوشه‌بندی در گروه دوم به‌طور ثابت قرار گرفته و ژن SLC35A3 که تا خوشه‌بندی با مرتبه سوم در کنار این ژن بود در مرتبه چهارم از این گروه جدا و به گروه اول و در کنار ژن‌های DALRD3 و S100A16 قرار گرفت. همچنین ژن HPS6 که در خوشه‌بندی ناشی از مرتبه اول آنتروپی به سمت خوشه دوم نزدیک شده بود، بعد از خوشه‌بندی شدن ژن در جای دیگر، SLC35A3 در مرتبه سوم آنتروپی، جایگزین این ژن گردید. گروه اول نیز خود به دو شاخه فرعی عمده خوشه‌بندی شدند. ژن GRWD1 و S100A16 در مرتبه دوم آنتروپی از شاخه فرعی اول به شاخه فرعی دوم گروه اول منتقل شدند. همچنین ژن DES در مرتبه سوم آنتروپی از خوشه خود جدا و به خوشه فرعی دوم اضافه شد. ژن‌های CNOT8، DALRD3 و RPS3A در مرتبه چهارم آنتروپی از شاخه فرعی خود در گروه اول جدا و به خوشه فرعی دیگر این گروه اضافه شدند. به‌طور کلی توپولوژی خوشه‌بندی‌ها با تغییر مرتبه آنتروپی از ۱ به ٤، تغییر یافت که عمده‌ترین تغییرات را در خوشه‌بندی‌های حاصل از آنتروپی مرتبه ٤ شاهد بودیم. مشاهده گردید که روش KL_H که مبتنی بر محاسبه آنتروپی مراتب متفاوت آنتروپی می‌باشد، مستقل از طول ژن بوده و کاملا با محتوای ژن و فراوانی نوکلئوتیدها در زنجیره در ارتباط است (شکل ۳).

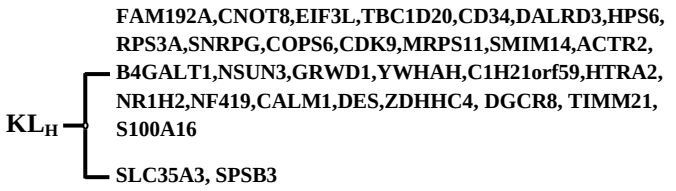


KL_H –H4		
GENE	length	Exon count
DALRD3	2957	12
S100A16	6912	3
SLC35A3	14901	7
CNOT8	17614	7
RPS3A	4687	6
DES	7076	9
YWHAH	1445	1
COPS6	2766	10
ZNF419	7336	5
HTRA2	3310	8
B4GALT1	53285	6
CDK9	4402	7
CD34	22955	8
TBC1D20	18982	8
EIF3L	22981	13
MRPS11	10887	6
CALM1	6340	6
ZDHHC4	8622	8
TIMM21	5716	6
C1H21orf59	11628	7
FAM192A	26969	19
NSUN3	60167	6
NR1H2	6648	10
DGCR8	14332	13
ACTR2	47929	9
SMIM14	58601	5
SNRPG	9003	4
GRWD1	6066	7
HPS6	2622	1
SPSB3	5570	7

شکل ۳- عدم وابستگی کولبک- لیبلر (KL_H) ناشی از آنتروپی مرتبه چهارم ژن‌ها به طول ژن‌ها در خوشه‌بندی به روش Single.

Figure 3. Lack of dependency of KL_H due to forth order of entropy on the length of genes in single clustering.

تجمیع نتایج حاصل از خوشه‌بندی: در این بخش از پژوهش بر اساس اطلاعات مفیدی که از ۷ روش معمول خوشه‌بندی ژن‌ها در روش محاسبه کولبک- لیبلر به‌دست آمد، نتایج تجمیع گردید در این راه از طبقه‌بند Adaboost استفاده شد. شکل ٤ خوشه‌بندی نهایی بر اساس نتایج حاصل از Adaboost را نشان می‌دهد. طبقه‌بند Adaboost ژن‌ها را به دو دسته عمده تقسیم نمود. با توجه به ارتولوگ بودن این ژن‌ها با انسان، ارتباط بین آن‌ها و عملکردشان با مراجعه به سایت GeneMANIA مورد بررسی و پیش‌بینی قرار گرفت تا صحت خوشه‌بندی روش‌ها بررسی و مقایسه گردد.



شکل ٤- نتایج حاصل از تجمیع نتایج خوشه‌بندی با استفاده از الگوریتم Adaboost روی ژن‌های مورد بررسی.

Figure 4. The aggregation of clustering results with the Adaboost algorithm over investigated genes.

بررسی دقت نتایج حاصل از دسته‌بند Adaboost: طبق نتایج حاصل از دسته‌بند Adaboost و بر اساس نتایج کولبک- لیبلر مبتنی‌بر آنتروپی نسبی (KL_H) ژن‌ها به دو گروه اصلی تقسیم شدند. با مراجعه به سایت GeneMANIA و در بررسی ژن‌های خوشه اول (۲۸ ژن) وخوشه دوم (۲ ژن)، شاهد ارتباط این ژن‌ها با تعدادی از ژن‌های دیگر بودیم که به‌ترتیب ۷۵/۷۳ درصد و ٦٧/٦٤ درصد بیان مشترک[1] با ژن‌های دیگر نشان دادند. با مشاهده ارتباط ژن‌های خوشه‌های اول و دوم مشاهده شد که تمام ژن‌ها در هر خوشه با همدیگر ارتباط متقابل داشته و مسیر متابولیکی مشترکی را دارا هستند. همچنین با بررسی عملکرد ژن‌ها مشاهده گردید که ژن‌های هر خوشه وظایف متفاوتی داشته پس مسیرهای متابولیکی آن‌ها هم متفاوت خواهد بود. بررسی‌ها نشان داد که ژن‌های دو خوشه مسیر متابولیک مشترک با یکدیگر نداشته و هیچ‌کدام از ژن‌های خوشه اول در ارتباط متقابل با ژن‌های خوشه دو نبودند و بالعکس.

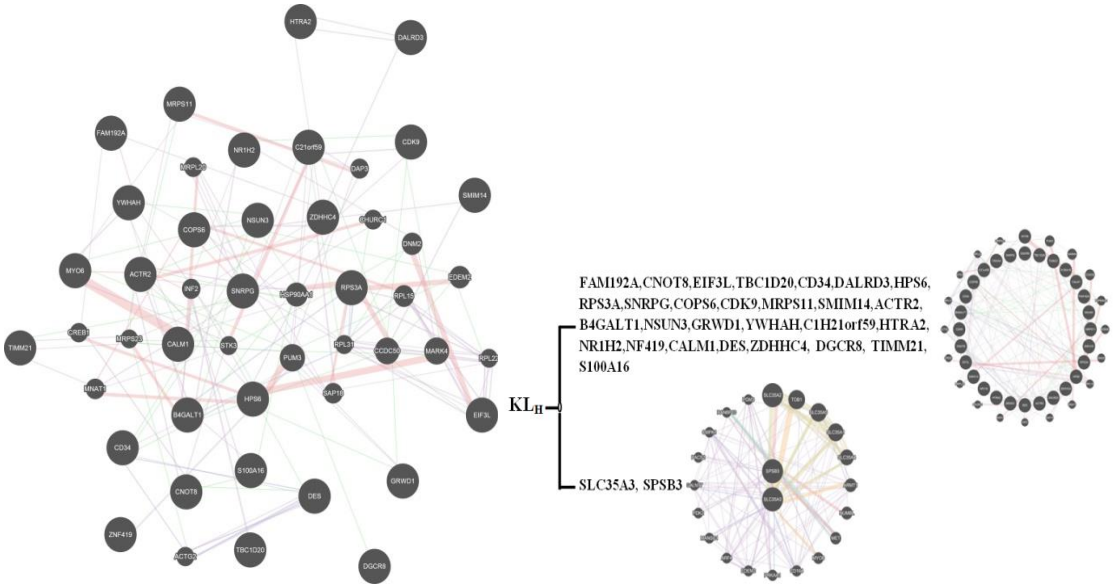
همچنین با بررسی عملکرد ژن‌های خوشه اول و دوم مشخص گردید که هیچ یک عملکرد یکسان و مشابه‌ای نداشتند و ژن‌های هر خوشه مسیرهای متابولیکی متفاوت و متمایزی را کددهی می‌نمایند و عملکرد مشترک و یکسانی در دو خوشه ژنی مشاهده نگردید که این خود به نوعی تأییدی بر قدرت و دقت روش ارایه شده می‌باشد.

جدول ٥- وظایف و عملکرد ژن‌های خوشه اول و دوم.

Table 5. Functions of the first and second gene clusters.

عملکرد ژن‌های خوشه اول	عملکرد ژن‌های خوشه دوم
function of first cluster genes	Function of second cluster genes
viral transcription	nucleotide transport
ribosome	carbohydrate derivative transport
nitric oxide metabolic process	organophosphate ester transport
ribosomal subunit	glycosylation
cytosolic part	macromolecule glycosylation
translational initiation	protein glycosylation
establishment of protein localization to membrane	nucleotide transmembrane transporter activity
regulation of nitric-oxide synthase activity	organophosphate ester transmembrane transporter activity
viral gene expression	phosphate transmembrane transporter activity
protein targeting	nucleobase-containing compound transmembrane transporter activity

1- Co-expression



شکل ۵- سمت چپ، نتایج دسته‌بندی Adaboost ژن‌ها مبتنی بر آنتروپی نسبی (KL_H)؛ سمت راست، ارتباط درونی بین ژن‌های خوشه اول.

Figure 5. Left panel, the results of the Adaboost genes grouping based on KL_H, right panel, the interrelationship among genes in the first cluster.

بررسی نتایج حاصل از عملکرد، ارتباط متقابل و مسیرهای متابولیکی مشترک ژن‌ها، روش خوشه‌بندی مبتنی‌بر آنتروپی (KLH) را روشی صحیح، منطقی و در عین حال سریع نشان داد. این روش علاوه‌بر این‌که معایب همتراز نمودن ژن‌ها را نداشته شکل ١٤ محتوا و شکل واقعی ژن را مورد بررسی قرار می‌دهد و نیاز به حافظه بالا برای توالی‌های با طول بزرگ ندارد و با توجه به این‌که محتوای اطلاعات درون توالی DNA از دست نمی‌رود، صحت خوشه‌بندی داده‌های توالی افزایش می‌یابد. یکی از مهمترین پیش پردازش‌ها به‌منظور بهبود عملکرد سامانه طبقه‌بندی، کاهش ابعاد فضای ویژگی می‌باشد. کاهش ابعاد فضای ویژگی باعث کاهش پیچیدگی فرایند طبقه‌بندی و در نتیجه کاهش وقوع خطا می‌شود. یکی از روش‌هایی که برای کاهش بعد فضای ویژگی معرفی شده است، استخراج ویژگی نامیده می‌شود. تئوری اطلاعات و معیارهای اندازه‌گیری مبتنی‌بر آن از جمله کولبک- لیبلر و اطلاعات متقابل ویژگی‌هایی مختلفی از ارتباط ژن‌ها در اختیار ما می‌گذارد که دارای حداکثر اطلاعات در مورد خروجی می‌باشد که باعث افزایش دقت طبقه‌بندی می‌گردد. مقایسه نتایج آنتروپی ژن‌ها و اگزون‌ها با معیار کولبک- لیبلر مبتنی‌بر آنتروپی نشان داد که با توجه به ارتباط مستقیم فرمول محاسبه کولبک- لیبلر به آنتروپی و حساسیت آن به مقادیر آن، ژن‌ها و اگزون‌هایی که کمترین و بیشترین مقادیر آنتروپی را در رتبه‌های ١ الی ٤ کسب نمودند دقیقا همان ژن‌ها و اگزون‌هایی بودند که بیشترین فاصله ژنی در کولبک- لیبلر مبتنی‌بر آنتروپی را دارا شدند. همچنین می‌توان نتیجه گرفت اگر آنتروپی دو قطعه DNA مشابه به هم باشند، آنگاه، فاصله کولبک- لیبلر آن‌ها صفر خواهد شد و انتظار می‌رود آن دسته از قطعات DNA که چنین خاصیتی را داشته باشند یا فاصله کولبک- لیبلر آن‌ها به صفر نزدیک باشد (خصوصاً در آنتروپی‌های مراتب بالا)، احتمالا یک نقش زیستی مشابه دارند. ماتریس فاصله ایجاد شده در روش‌های یادشده می‌تواند ورودی الگوریتم‌های نظارت نشده‌ایی مثل خوشه‌بندی سلسله مراتبی باشد. در آن صورت قطعه‌های DNA که خود را به‌صورت 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 (۸). نتایج حاصل از محاسبه فاصله کولبک- لیبلر روی قطعات DNA تا حدی نشان داد که ساخت و ترکیب DNA ژن‌ها، اگر به فضای دیگری نگاشت شود (مثل فضای آنتروپی)، 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 به‌طور نسبی وجود دارد. البته این در صورتی درست است که هیچ‌گونه اطلاعات زیستی دیگری موجود نباشد. پژوهش‌های مختلفی در این خصوص انجام گرفته است در پژوهشی، نظریه درختچه حیات برای تحلیل مسیرهای متابولیتی به‌کار گرفته شد. این پژوهش از اولین پژوهش‌های محسوب می‌شود که ترکیب داده‌های سطح DNA و مسیرهای متابولیکی را با استفاده از درختچه حیات انجام داد (۸). لی و همکاران (۲۰۰۹) از کولبک- لیبلر به‌عنوان روشی نو در بازسازی درخت فیلوژنتیک کورنوویروس و سارس ویروس‌ها استفاده کردند (۲۰). همچنین پژوهش‌هایی جهت استفاده هر چه بیشتر داده‌های متابولیکی برای درک بهتر ارتباط تکاملی گونه‌های مختلف (٤)، با توجه به انباشت داده‌های متابولیکی و با استفاده از نظریه گراف انجام شده است که نشان دهنده همخوانی درختچه فیلوژنی ایجاد شده با نتایج آزمایشگاهی بود (٤ و ۱۱). باید خاطر نشان کرد که در پژوهش صورت گرفته برخلاف تعدادی از پژوهش‌های انجام شده که داده‌های ورودی آن‌ها متعلق به گونه‌های مختلف بودند از داده ژنی یک گونه (گاو شیری) برای ایجاد درختچه تکاملی استفاده شد. با این وجود مشاهده گردید که بنیاد نظری ایجاد کننده درختچه تکاملی را می‌توان برای ارتباط متابولیکی نیز به کار برد.

نتیجه‌گیری

در روش ارائه شده ما از خوشه‌بندی به یک گروه‌بندی زیستی از ژن‌ها دست یافتیم. با توجه به استخراج ویژگی‌های حاصل شده از نتایج خوشه‌بندی، از این روش نو و بدیع می‌توان در خوشه‌بندی ژن‌های دیگر استفاده نمود. نتایج نهایی خوشه‌بندی ژن‌ها و بررسی عملکرد ژن‌های هر خوشه، روش ارائه شده در این پژوهش را برای خوشه‌بندی ژن‌ها مورد تأیید قرار داد. ما معتقدیم روش ارایه شده در این مقاله می‌تواند برای اختصاص دادن و پیشی بینی فعالیت زیستی برای آندسته از ژن‌هایی که حاشیه نویسی ژنومی قوی ندارند، کمک کننده باشد، چرا که فقط متکی به توالی DNA ژن‌ها بوده و اندازه و طول ژن‌ها اثری در ماهیت الگوریتم ارایه شده ندارد. بنابراین، خوشه‌بندی توام ژن‌هایی که حاشیه نویسی ژنومی قوی دارند با آن‌هایی که ندارند، می‌تواند ارزش افزوده تحلیل و زیستی به گروه دوم ژن‌ها (بدون حاشیه‌نویسی ژنومی) بدهد.

منابع

1. Buitenhuis, A.J., Sundekilde, U.K., Poulsen, N., Bertram, H.C., Larsen, L.B., and Sørensen, P. 2013. Estimation of Genetic Parameters and Detection of QTL for Metabolites in Danish Holstein Milk. Journal of Dairy Science. 14: 1-10.
2. Alinaghizadeh, H., Mohammad Abadi, MR., and Zakizadeh, S. 2010. Exon 2 of BMP15 gene polymorphismin Jabal Barez Red Goat. Journal of Agricultural Biotechnology. 2: 69-80.
3. Barazandeh, A., Mohammadabadi, MR., Ghaderi, M., and Nezamabadipour, H. 2016. Genome-wide analysis of CpG islands in some livestock genomes and their relationship with genomic features. Czech Journal of Animal Science. 61: 487-495.
4. Clemente, J.C., Satou, K., and Valiente, G. 2007. Phylogenetic reconstruction from non-genomic data. Bioinformatics. 23: 110–115.
5. Erill, I. 2012. Information Theory and biological sequences: Insights from an evolutionary prespective. Nova Science Publishers, Inc.
6. Freund, Y., and Schapire, R. 1996. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 55: 119.
7. Freund, Y., and Schapire, R. 1996. Experiments with a new boosting algoritm. Paper read at Proceeding of the Thirteenth Internatioanal Conference on Machine Learning.
8. Forst, C.V., and Schulten, K. 2001. Phylogenetic analysis of metabolic pathways. Journal of Molecular Evolution. 52: 471–489.
9. Ghaderi-Zefrehei, M., A. Bandi Dastjerdi, A., Bahreini Behzadi, M.R., F. Samadian, F., and Meamar, M. 2016. Investigation of Information Accumulation in Escherichia Coli's DNA Sequence Affecting Mastitis in Dairy Cow Using Information Theory. Journal of Ruminant Research. 4: 2016.
10. Gray, R.M. 2013. Entropy and Information Theory. First Edition. Springer-Verlag New York publisher.
11. Heymans, M., and Singh, A.K. 2003. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics. 19: 138–146.
12. Javanmard, A., Mohammadabadi, M.R., Zarrigabayi, G.E., Gharahedaghi, A.A., Nassiry, MR., Javadmansh, A., and Asadzadeh, N. 2008. Polymorphism within the intron region of the bovine leptin gene in Iranian Sarabi cattle (Iranian Bos Taurus). Russian Journal of Genetics. 44: 495-497.
13. Jiang, S.C., Tang, C., Zhang, L., and Zhang, A. 2014. A Maximum Entropy Approach to Classifying Gene Array Data Sets. Workshop on Data Mining for Genomics, First SIAM International Conference on Data Mining.
14. Kharrati Koopaei, H., Mohammad Abadi, MR., Ansari Mahyari, S., Tarang, AR., Potki, P., and Esmailizadeh, AK. 2012a. Effect of DGAT1 variants on milk composition traits in Iranian Holstein cattle population. Animal Science Papers and Reports. 30: 231-240.
15. Kharrati Koopaei, H., Mohammadabadi, M.R., Tarang, A., Kharrati Koopaei, M., and Esmailizadeh Koshkoiyeh, A. 2012b. Study of the association between the allelic variations in DGAT1 gene with mastitis in Iranian Holstein cattle. Modern Genetics Journal. 7: 101-104.
16. Kharrati Koopaei, H., Mohammadabadi, M.R., Ansari Mehyari, S., Esmailizadeh, A.K., Tarang, A., and Nikbakhti, M. 2011. Genetic variation of DGAT1 gene and its association with milk production in Iranian Holstein cattle breed population. Iranian Journal of Animal Science Research. 3: 185-192.
17. Khatib, H., Monson, R.L., Schutzkus, V., Kohl, D.M., Rosa, GJM., and Rutledge, J.J. 2008. Mutations in the STAT5A gene are associated with embryonic survival and milk composition in cattle. Journal of Dairy Science. 91: 784–793.
18. Kim, J., Kim, S., Lee, K., and Kwon, Y. 2009. Entropy analysis in yeast DNA. Chaos, Solitons and Fractals. 39: 1565–1571.

19. Kullback, S., and Leibler, R. 1951. On information and sufficiency. The Annals of Mathematical Statistics. 22: 79–86.
20. Lee, L. 2009. Used kullback-Liebler measure as a new method for the reconstruction of the phylogenetic tree of the Cornavirus and SARS viruses.
21. Lemay, D.G., Lynn, D.J., Martin, W.F., Neville, M.C., Casey, T.M., Rincon, G., Kriventseva, E.V., Barris, W.C., Hinrichs, A.S., Molenaar, A.J., Pollard, K.S., Maqbool, N.J., Singh, K., Murney, R., Zdobnov, E.M., Tellam, R.L., Medrano, J.F., German, J.B., and Rijnkels, M. 2009. The bovine lactation genome: insights into the evolution of mammalian milk. Genome Biology. 10: R43
22. Li, C., and Wang, J. 2005. Relative entropy of DNA andits application. Physica A. 347: 465–471.
23. Liou, C.Y., Tseng, S.H., Cheng, W.C., and Tsai, H.Y. 2013. Structural Complexity of DNA Sequence. Computational and Mathematical Methods in Medicine. 2013: 1-11.
24. Liu, B. 2007. Uncertainty Theory, 2nd ed., Springer-Verlag, Berlin.
25. Machado, J.T. 2012. Shannon entropy analysis of the genome code. Mathematical Problems in Engineering. 2012: 1-12.
26. Mohammad Abadi, M.R., Mohammadi, A. 2010a. Study of beta-lactoglobulin genotypes in native and Holstein cattle of Kerman province. Journal of Animal Productions. 12: 61-67.
27. Mohammadabadi, M.R., Nikbakhti, M., Mirzaee, H.R., Shandi, A., Saghi, D.A., Romanov, M.N., and Moiseyeva, I.G. 2010b. Genetic variability in three native Iranian chicken populations of the Khorasan province based on microsatellite markers. Russian Journal of Genetics. 46: 505-509.
28. Mousavizadeh, A., Mohammad Abadi, MR., Torabi, A., Nassiry, MR., Ghiasi, H., and Esmailizadeh, AK. 2009. Genetic polymorphism at the growth hormone locus in Iranian Talli goats by polymerase chain reaction-single strand conformation polymorphism (PCR-SSCP). Iranian Journal of Biotechnology. 7: 51-53.
29. Monge, R.E., and Crespo, J.L. 2014. Comparison of Complexity Measures for DNA Sequence Analysis.. International Work Conference on Bio-inspired Intelligence (IWOBI). Pp: 71-75.
30. Neagoe, I.M., Popescu, D., and Niculescu, V.I.R. 2014. Applications of entropic divergence measures for DNA segmentation into high variable regiones of cryposporidium spp. GP60 gene. Romanian Reports in Physics. 66: 1078–1087.
31. Pham, T.D., Crane, D.I., Tannock, D., and Beck, D. 2004. Kullback-Leibler Dissimilarity of Markov Models for Phylogenetic Tree Reconstruction. Proceeding of international Symposium on Inteligent Multimedia, Video and Speech Processing. October Pp: 20-22 HongKong.
32. Porto-DIaz, L., BolOn-Canedo, V., Alonso-Betanzos, A., and Fontenla-Rome, O. 2011. A study of performance on microarray data sets for a classifier based on information theoretic learning. Neural Networks. 24: 888–896.
33. Ruiz-Marin, M., Matilla-Garcia, M., Cordoba, J.A.G., Susillo-Gonzalez, J.L., Romo-Astorga, A., Gonzalez-Pérez, A., Ruiz, A., and Gayan, J. 2010. An entrpyetest for single-locus genetic association analysis. BMC Genetics. 11: 19.
34. Shannon, C.E. 1948. A mathematical theory of communication. Bell System Technical Journal. 27: 379–423 and 623–656.
35. Sherwin, B.W. 2010. Entropy and information approaches to genetic diversity and its expression: genomic geography. EntropyPp: 1765-1798. Shojaei, M., Mohammad Abadi, MR., Asadi Fozi, M., Dayani, O., Khezri, A., and Akhondi, M. 2010. Association of growth tra*i*t and Leptin gene polymorphism in Kermani sheep. Journal of Cell and Molecular Research. 2: 67-73.
36. Sundekilde, U.K., Larsen, L.B., and Bertram, H.C. 2013. NMR-Based Milk Metabolomics. Metabolites. 3: 204-222.

37. Tautz, D., Trick, M., Dover, G.A. 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature, 322: 652–656.
38. Vinga, S., Almeida, J. 2003. Alignment-free sequence comparison: review. Bioinformatics 19: 513–523.
39. Vinga, S. 2013. Information theory applications for biological sequence analysis. Briefings in bioinformatics. 15: 376-389.
40. Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C.T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G.D., and Morris, Q. 2010, The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research. 38: W214-W220.
41. Xie, X., Yu, Y., Liu, G., Yuan, Z., and Song, J. 2010. Complexity and Entropy Analysis of DNA Methyltransferase. J Data Mining in Genom Proteomics. Volume 1, Issue 2, 1000105.
42. Zamani, P., Akhondi, M., Mohammadabadi, MR., Saki, A.A., Ershadi, A., Banabazi, M.H., and Abdolmohammadi, AR. 2013. Genetic variation of Mehraban sheep using two intersimple sequence repeat (ISSR) markers. African Journal of Biotechnology. 10: 1812-1817.
43. Zhang, J.L., Zan, L.S., Fang, P., Zhang, F., Shen, GL., and Tian, WQ. 2008. Genetic variation of PRLR gene and association with milk performance traits in dairy cattle. Canadian Journal of Animal Science. 88: 33-39.

Gorgan University of Agricultural Sciences and Natural Resources
J. of Ruminant Research, Vol. 5(3), 2017
http://ejrr.gau.ac.ir

Possibility of application of relative entropy in clustering of some milk governing genes in dairy cattle

[*]H. Dehghanzadeh[1], S.Z. Mirhoseini[2], M. Ghaderi Zefrehei[3], H. Tavakoli[4] and S. Esmaeilkhaniyan[5]

[1]Ph.D. Student of Genetic and [2]Professor, Dept., of Animal Sciences, [4]Assistance Prof., Dept., of Electrical Engineering, University of Guilan, Rasht, Iran, [3]Assistant Prof., Dept., of Animal Sciences, University of Yasouj, Yasouj, Iran, [5]Associate Prof., Dept., of Animal Science Research Institute, Agricultural Research, Education and Extension Organization, Karaj, Iran

Received: 07/26/2017; Accepted: 11/08/2017

Abstract

Background and objectives: Apart from the fact that milk plays an important role in human nutrition, increasing milk production or changing its composition has attracted the attention of animal breeders, therefore, it is crucial to study and evaluate the genes underpinning milk production and its composition. Information theory is a branch of mathematics that overlaps with communications, biology, and medical engineering. Entropy is a measure of uncertainty in the set of information. In his famous article in 1948, Shannon introduced this concept and used its results in a number of basic issues of coding and data transferring theory, which forms the basis of new information theory. Information theory is used in genetic and bioinformatics analyses and can be used for many analyses related to the structures and sequences. Bio-computational grouping of genes facilitates genetic analysis, sequencing and structural-based analyses.

Materials and methods: DNA sequence of 30 genes involved with milk protein production were extracted *ad hoc* from NCBI genome database and stored in FASTA format. In this study, for each gene and its exons sets, the entropy was calculated in orders one to four. In this way, the Markov chain up to order three was used. Based on the relative entropy of genes and exons, kullback-Leibler divergence was calculated. After obtaining the kullback-Leibler distance for genes and exons sets, the results were entered as input into seven clustering algorithms: Single, Complete, Average, Weighted, Centroid, Median, and K-Means. 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. All calculations were performed using the MATLAB Engineering Software (2015).

Results: By investigating the results of genes metabolic pathways based on their gene annotations, it was turned out that proposed clustering method, yielded correct, logical, and fast results. This method at the same that hadn't had the disadvantages of aligning allowed the genes with actual length and content to be considered and also didn't require high memory for large-length sequences.

Conclusion: It can be concluded that the performance of the proposed method could be used with other competitive gene clustering methods to group biologically relevant set of genes. Also, the proposed method can be seen as a predictive method for those genes bearing up weak genomic annotations.

Keywords: Information theory, Dairy cattle, Kullback-Leibler divergence, Gene clustering

*Corresponding author: h_dehghanzadeh@yahoo.com