تحقیقات تولیدات دامی

سال ششم/شماره سوم/پاییز ۱۳۹۶ (۲۸-۱۱)

Animal Production Research

دانشگاه گیلان

ایجاد درخت ژنی با استفاده از واگرایی کولبک- لیبلر روی ژن‌های موثر بر تولید شیر در گاو شیری

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

۱- دانشجوی دکتری گروه علوم دامی، دانشکده علوم کشاورزی، دانشگاه گیلان

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

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

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

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

(تاریخ دریافت: ۹۶/۰۶/۰۵ - تاریخ پذیرش: ۹۶/۰۹/۱۶)

چکیده

نظریه اطلاعات، شاخه‌ای از ریاضیات است که با مهندسی ارتباطات، زیست‌شناسی و پزشکی هم‌پوشانی دارد. هدف از بررسی حاضر ارائه روشی جهت خوشه‌بندی تعدادی از ژن‌های موثر روی تولید شیر در گاو شیری با استفاده از الگوریتمی متکی بر واگرایی کولبک - لیبلر بود. در این پژوهش بعد از استخراج توالی DNA ژن و اگزون‌های موثر بر تولید شیر در گاو شیری، فراسنجه آنتروپی در مراتب یک تا چهار برای هر ژن و اگزون‌های هر ژن محاسبه شد. جهت استخراج فاصله میان ژن‌ها از یکدیگر، از واگرایی کولبک - لیبلر در سه روش مختلف استفاده شد. روش‌های اول و دوم مبتنی بر همترازی ولی روش سوم غیر مبتنی بر همترازی و بر پایه آنتروپی نسبی ژن‌ها بود. نتایج هر سه روش واگرایی کولبک - لیبلر روی توالی DNA ژن‌ها و اگزون‌ها با استفاده از هفت روش معمول Single، Complete، Average، Weighted، Centroid، Median و KMeans خوشه‌بندی شدند. تجمیع نتایج هر خوشه‌بندی که با الگوریتم AdaBoost انجام شد، و خود نوعی درخت ژنی را تداعی کرد، نشان داد که روش سوم، خوشه‌بندی معقولی را از نظر زیستی برای مجموعه‌ای از ژن‌ها حاصل نمود چرا که با نتایج حاشیه-نویسی ژنومی ژن‌های حاصل از GeneMANIA مطابقت داشت. این اعتقاد وجود دارد که روش ارائه شده برای ایجاد درخت ژنی می‌تواند با سایر روش‌های متکی بر توالی DNA ژن‌ها جهت خوشه‌بندی مجموعه‌ای از ژن‌ها، رقابت نماید و لذا می‌تواند در گروه‌بندی ژن‌های سایر گونه‌ها نیز بکار رود.

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

* نویسندهٔ مسئول: mirhosin@guilan.ac.ir

مقدمه

شیر کامل‌ترین غذایی است که می‌تواند مورد استفاده انسان قرار گیرد. پژوهش و بررسی ژن‌هایی که روی تولید و ترکیب شیر نقش موثری دارند، بسیار با اهمیت بوده و می‌تواند گامی مهم در جهت شناسایی و توسعه انتخاب به کمک نشانگر[1] و تدوین برنامه‌های اصلاح نژادی جهت بهبود صفات تولیدی به شمار آید (Buitenhuis *et al.*, 2013; Khatib *et al.*, 2008; Sundekilde *et al.*, 2013; Zhang *et al.*, 2013). در اصلاح نژاد دام با در دست بودن ژن‌های مرتبط با صفات تولیدی همچون تولید شیر، این امکان وجود دارد که میزان اطلاعات ذخیره شده در بخش‌های مختلف آن با استفاده از نظریه اطلاعات[2] بررسی و با تفسیر زیستی نتایج حاصله، رهیافت جدیدی برای افزایش تولید شیر و یا دستکاری‌های ژنی با اهداف متفاوت ایجاد کرد.

امروزه شاهد ظهور ابزارها و الگوریتم‌های محاسباتی[3] جدید جهت آزمایش و فرموله کردن فرضیه‌هایی همچون چگونگی سازماندهی، تکامل ژنوم و رویت یک فنوتیپ مشخص از یک ژنوم رمزنگاری شده هستیم. نظریه اطلاعات که شاخه‌ای از ریاضیات است و با مهندسی ارتباطات، زیست‌شناسی و پزشکی همپوشانی دارد، نقش مهمی را در این زمینه بازی می‌کند. این نظریه به کشف و بررسی قوانین ریاضی حاکم بر رفتار داده‌ها در مراحل انتقال، ذخیره و بازیابی داده‌ها می‌پردازد (Liu, 2007). آنتروپی[4] شانون هسته اصلی نظریه اطلاعات است و گاهی اوقات تحت عناوینی مثل اندازه عدم قطعیت یا میزان تصادفی بودن[5]، درهم ریختگی[6] و پیش‌بینی‌ناپذیری[7] شناخته می‌شود. اطلاعات، مقیاس عدم اطمینان یا آنتروپی در یک موقعیت است و هر چه 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[8]) گویند (Shannon, 1948). واحد آنتروپی، بیت[9] 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 (Gray, 2013). حیات نیز به عنوان یک سامانه پردازش اطلاعات می‌تواند از طریق تکامل، توانایی ذخیره و پردازش اطلاعات لازم برای بازساخت خود را بدست آورد (Erill, 2012). از نظریه اطلاعات به عنوان ابزاری مهم و به چند صورت در جستجوی الگوهایی در توالی‌های DNA (Tautz *et al.*, 1986)، نقش آمینواسیدها در ساختار پروتئین‌ها در مخمر (Kim *et al.*, 2009)، تحلیل جایگاه‌های صفات کمّی[10] و اپیستازی[11] (Ruiz-Marin *et al.*, 2010)، بررسی اطلاعات ژنوم جهانی[12] (Machado, 2012)، تحلیل داده‌های ریزآرایه DNA (Jiang *et al.*, 2014)، طبقه‌بندی ژن‌های درگیر در سرطان (Porto-Diaz *et al.*, 2011)، مقایسه اندازه پیچیدگی برای تجزیه توالی‌های DNA (Gray, 2013; Liou *et al.*, 2013; Monge and Crespo, 2014)، بازساخت درختان فیلوژنتیکی بدون همتراز کردن بازها (Pham *et al.*, 2004)، پژوهش‌های تکاملی (Erill, 2012)، تنوع ژنتیکی (Sherwin, 2010)، مقایسه محتوای اطلاعات نواحی اینترون و اگزون ژن‌ها (Xie *et al.*, 2010) و تحلیل زیر گونه‌های انگل کریپتوزپوریدیوم[13] (Neagoe, 2014) استفاده شده است. در نظریه احتمالات و نظریه اطلاعات، واگرایی کولبک - لیبلر[14]- یا به عبارتی آنتروپی نسبی[15]- یک معیار نامتقارن برای اندازه‌گیری تفاوت دو توزیع احتمالاتی Q و P است (Li and Wang, 2005). ماهیت و ساخت نظری این معیار کاربردهای زیادی را در عمل در حوزه‌های مختلف امکان‌پذیر ساخته است. در چند دهه اخیر چندین روش برای خوشه‌بندی ژن‌ها- که درخت ژنی نیز نامیده می‌شوند- و پروتئین‌ها پیشنهاد شده که اغلب این روش‌ها بر اساس همترازی ژن‌ها بوده است. اما در یک دامنه بالا با توجه به بزرگی بسیاری از توالی‌ها و استواری روش‌های استاندارد بر اساس مقایسه هر

1. Marker-assisted selection
2. Information theory
3. Computational
4. Entropy
5. Randomness
6. Disorderliness
7. Unpredictability
8. Nego Entropy
9. Bit
10. QTL
11. Epistasis
12. Global genomic information
13. Cryptosporidium
14. Kullback-Leibler divergence
15. Relative entropy

نوکلئوتید به نوکلئوتید متناظر توالی دیگر، کارآیی پایین و اندکی مشکل و غیر ممکن می‌شود (Changchuan *et al.*, 2014; Zhou *et al.*, 2007). تاکنون روش‌های مختلف و جدیدی برای بازسازی درخت فیلوژنی بدون همترازی توالی‌ها پیشنهاد شده است مثل تجزیه بر اساس مولفه‌ها[1] (Edwards *et al.*, 2002)، روش تجزیه مقادیر منفرد[2] (Stuart *et al.*, 2002a; Stuart *et al.*, 2002b)، روش دستوری دینامیک[3]، روش مدل مارکف[4] (Qi *et al.*, 2004; Stuart *et al.*, 2002a) و روش‌های فراکتال[5] (Yu *et al.*, 2003; Yu *et al.*, 2005).

مقاله حاضر کاربرد الگوریتمی متکی به واگرایی کولبک-لیبلر را نشان می‌دهد که نویسندگان مقاله برای اولین بار جهت خوشه‌بندی تعدادی از ژن‌های موثر روی تولید شیر ارایه کردند. تاکنون، بر اساس دانش نویسندگان، هیچگونه پژوهشی ژن‌های موثر بر تولید شیر را با استفاده از نظریه اطلاعات خوشه‌بندی نکرده است. انتظار می‌رود که استخراج الگوها و درخت‌های ژنی حاصله از این خوشه-بندی[6] بتواند در کنکاش‌های زیستی، دارویی و اصلاح نژادی بکار رود.

مواد و روش‌ها

استخراج توالی DNA ژن‌ها: بر اساس گزارشات، حدود ۶۸۷۵ ژن وجود دارد که روی تولید شیر در پستانداران موثر هستند. بعضی ازاین ژن‌ها فقط در غده پستانی بیان شده و بعضی دیگر در بافت‌های دیگری مثل کبد، کلیه، ماهیچه‌ها و غیره نیز بیان می‌شوند (Lemay *et al.*, 2009). ژن‌های مورد بررسی در این مقاله از نتایج پژوهش Lemay *et al.* (2009) انتخاب شدند. در مقاله یاد شده، ۳۰ ژن از دسته ژن‌های پستانی موثر در تولید شیر مربوط به گاوهای تلیسه[7] (این اسم از مقاله یاد شده گرفته شد و معنای واقعی کلمه نیست) 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[8]، شماره دست‌یابی[9]، تعداد و طول هر اگزون[10] و جایگاه آن روی کروموزوم از بانک ژنی *NCBI*[11] دریافت و سپس با پیکربندی فستا[12] ذخیره شدند (جداول ۱ و ۲ فایل ضمیمه). 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.

روند پژوهش: شکل ۱ مراحل و روند انجام کار در این پژوهش را نشان می‌دهد. هر کدام از این بلوک‌ها در نرم-افزار مهندسی متلب (*MATLAB*)[13] کدنویسی شده و نتایج آن در قالب نمودار و جداولی در این مقاله ارائه شد.

محاسبه مراتب[14] 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 سه استفاده شد. برای محاسبه آنتروپی مرتبه اول (مرتبه صفر زنجیره مارکف[15]) 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 است. در این نوع از آنتروپی فرض شد که ظاهر شدن هر نوکلئوتید، مستقل از نوکلئوتید دیگر در رشته DNA بوده و به نوع نوکلئوتید مجاورش بستگی ندارد.

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

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

1. Principal Component Analysis (PCA)
2. Singular Value Decomposition (SVD)
3. Dynamical language method
4. Markov model method
5. Fractal methods
6 .Clustering
7. Virgin mammary gene set
8. C-G content
9. Accession number
10. Exone
11. http://www.ncbi.nlm.nih.gov/genbank/gene
12. Fasta
13. Matlab engineering software
14. Orders
15. Markov chain

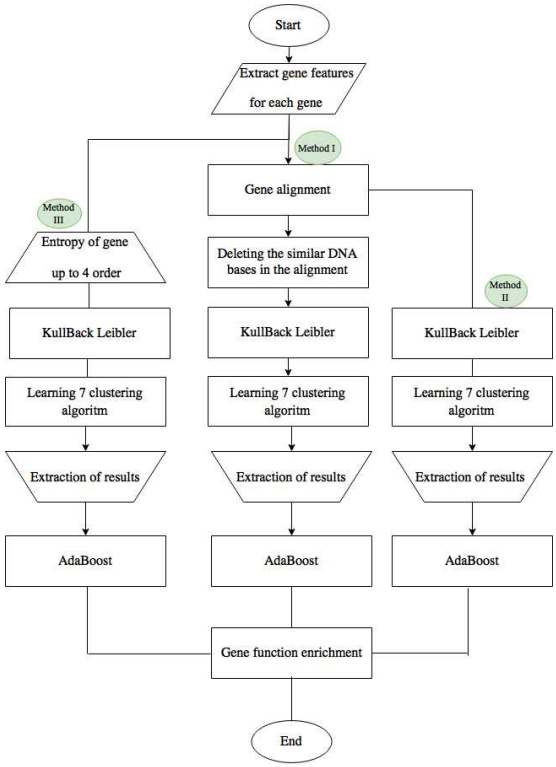


Fig. 1. The process of conducting this research

شکل ۱- روند انجام این پژوهش

کـه i نشانگـر وقـوع نوکلئـوتید قبلی و $p_i(j)$ هم احتمال وقـوع نوکلئـوتید j به شـرط وقـوع نوکلئـوتید i از مجمـوعه {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 سه 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 شد در کل چهار مرتبه آنتروپی محاسبه شد. آنتروپی هم برای طول کل ژن‌ها و هم اگزون‌ها محاسبه 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). در ضمن، اندیسی که در H ظاهر شد نشان‌دهنده مرتبه آنتروپی مورد نظر است.

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 $P(x), Q(x) \neq 0$ است. D_{KL} یک ماتریس متقارن نیست و در واقع یک فاصله حقیقی نمی‌باشد، بنابراین:

$$D_{KL}(P(x)||Q(x)) \neq D_{KL}(Q(x)||P(x))$$

لذا در این پژوهش D_{KL} به صورت زیر استفاده شد:

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

روش‌های اول و دوم مبتنی بر همترازی ولی روش سوم غیر مبتنی بر همترازی و بر پایه آنتروپی ژن‌ها بود. در روش اول و دوم جهت همترازسازی ژن‌ها و اگزون‌ها

هنگام محاسبه واگرایی کولبک- لیبلر، از نرم‌افزار *Bioedit* نسخه ۷,۲,۵ استفاده شد. در روش اول که برای سهولت اصطلاحاً به عنوان KL_B در متن از آن یاد شد در مقایسه دو توالی هم اندازه، ابتدا نوکلئوتیدهای متناظر مشابه و هم‌جایگاه حذف و سپس بر اساس فراوانی هر یک از نوکلئوتیدهای متناظر و متفاوت در دو توالی، در هر بار P (x) و Q (x) بدست آمده در فرمول جاگذاری و در انتها با جمع این اعداد، واگرایی کولبک- لیبلر مورد محاسبه قرار گرفت. این روش با توجه به ویژگی‌های خود، تفاوت-های دو توالی را مورد تاکید، بررسی و محاسبه قرار داد. در روش دوم که برای سهولت اصطلاحاً به عنوان KL_A در متن از آن یاد شد کاملاً شبیه روش اول بود و تنها تفاوت آن با روش قبلی این بود که نوکلئوتیدهای متناظر مشابه حذف نشده و در محاسبات دخالت داده شدند. این روش با توجه به ویژگی‌های خود شباهت‌های دو توالی را مورد تاکید، بررسی و محاسبه قرار داد. روش سوم هم که برای سهولت استفاده در متن از آن به عنوان KL_H یاد شد بر پایه مقادیر آنتروپی نسبی ژن‌ها و اگزون‌ها محاسبه شد. نحوه محاسبه بدین صورت بود که آنتروپی دو توالی ژنی مورد بررسی به طور مجزا محاسبه و مقدار عددی توالی اول به عنوان P و مقدار عددی توالی دوم به عنوان Q در نظر گرفته و در فرمول جاگذاری شدند. برای هر ژن و اگزون‌های آن به طور مجزا، آنتروپی‌های مراتب یک تا چهار و سپس آنتروپی نسبی آنها محاسبه شدند. در هر سه 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 شد. شکل ساده‌ای از این معیار برای پیدا کردن فاصله بین کانتیگ‌های ژنوم اشرشیاکلی موثر بر ورم پستان در تحقیق دیگری به کار برده شده است (Ghaderi-Zefrehei *et al.*, 2016).

ترکیب نتایج حاصل از انواع روش‌های خوشه‌بندی: معیار بدست آمده فاصله کولبک- لیبلر در مجموعه ژن‌ها و اگزون‌ها، به عنوان ورودی هفت روش معمول خوشه‌بندی *'Single'*[1]، *'Complete'*[2]، *'Average'*[3]، *'Weighted'*[4]، *'Centroid'*[5]، *'Median'*[6] و *'KMeans'* بکار رفت و درخت‌های ژنی بدست آمدند. در این مقاله تنها نتایج خوشه‌بندی حاصل از روش *Single* ارائه شده و بقیه در فایل ضمیمه قابل مشاهده است. برای ترکیب نتایج خوشه‌بندی، از الگوریتم آدابوست[7] (*AdaBoost*) استفاده شد (Freund and Schapire, 1996). در پایان، جهت تایید نتایج حاصل از *AdaBoost* و بررسی همخوانی نتایج خوشه‌بندی ژن‌ها با داده‌های حاشیه‌نویسی ژنوم آن‌ها، از *GeneMANIA prediction server*[8] استفاده شد (Warde-Farley *et al.*, 2010). همه محاسبات با استفاده از نرم-افزار مهندسی متلب (*MATLAB*) (2015) انجام شد.

نتایج و بحث

اطلاعات ۳۰ ژن مورد پژوهش در جداول ۱ و ۲ فایل ضمیمه قابل مشاهده است. 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 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 و۱۰) 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 دارا بودند. مقادیر آنتروپی و آنتروپی تصادفی رشته متناظر کلیه ژن‌ها در هر رتبه در جدول ۱ و نتایج آنتروپی کمینه و بیشینه ژن‌ها و اگزون‌ها در مراتب یک تا چهار در جداول ۲ و ۳ نشان داده شده است. اگزون‌هایی که آنتروپی آن‌ها در مرتبه چهارم بیشینه بود (یعنی مقادیر آنتروپی نزدیک به هشت بود) عبارت بودند از: اگزون ۱ ژن *YWHAH*، اگزون ۱ ژن *DGCR8* و اگزون ۸ ژن *TBC1D20*. به نظر 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 داد، طول کوتاه این اگزون بود، در مقابل، اگزون ۱ ژن *YWHAH* به علت طول بالاتر نسبت به سایر اگزون‌ها، مقدار آنتروپی بالاتری را به خود

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. AdaBoost
8. http://www.genemania.org

اختصاص داد. البته این موضوع همیشه صدق نکرد. همانطور که در جدول ۳ مشاهده شد اگزون ۸ ژن *TBC1D20* که طول بیشتری نسبت به اگزون ۱ ژن *YWHAH* داشت، فقط در مرتبه سوم مقادیر بالاتری را به خود اختصاص داده است.

اکثر ژن‌های مورد بررسی دارای اگزون‌هایی بودند که مقادیر بالا و پایین آنتروپی در آنها مشاهده شد. ولی در این میان، همه اگزون‌های ژن‌های *DES* و *FAM192A* دارای آنتروپی پایین‌تری از سایر اگزون‌های ژن‌های مورد بررسی بودند. آنتروپی مراتب یک تا چهار تمام ژن‌ها، اگزون‌ها و توالی متناظر تصادفی آنها در جدول ۳ فایل ضمیمه قابل دسترس است.

در مرحله بعد به خوشه‌بندی ژن‌ها و ایجاد درخت ژنی بر اساس روش‌های پیشنهادی پرداخته شد. تعداد زیادی از اشکال مربوط به خوشه‌بندی ژن‌ها با هفت الگوریتم خوشه‌بندی در بخش ضمیمه ارائه شده است.

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*

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

Cells with gray and pink colors indicated the highest and lowest entropy values for genes in corresponding order, respectively

جدول ۲- نتایج آنتروپی کمینه و بیشینه در مراتب یک تا چهار ژن‌های مربوطه

Table 2. Results of maximum and minimum entropy orders of one to four in respected genes

$H(x)$	Minimum entropy			Maximum entropy		
	Name of genes	Lenght of genes	Value	Name of genes	Lenght of genes	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

جدول۳ - نتایج آنتروپی بیشینه و کمینه در مراتب یک تا چهار اگزون‌های ژن‌ها

Table 3. Results of different entropy orders of one to four over gene exones

$H(x)$	Minimum entropy			Maximum entropy		
	Name of exone	Lenght of exone	Value	Name of exone	Lenght of exone	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

روش اول - خوشه بندی بر اساس واگرایی کولبک– لیبلر مبتنی بر تفاوت نوکلئوتیدهای متناظر ژن‌ها و اگزون‌ها (*KL_B*): با این روش و بر اساس نتایج حاصل، ژن‌های *YWHAH* و *COPS6* با مقدار KL_B= ۳/۰۴۲ و ژن‌های *YWHAH* و*NSUN32* با KL_B= ۷۸۲۸۴/۱۴۵ به ترتیب کمترین و بیشترین فاصله را داشتند. همچنین در کل ۲۱۱ اگزون، اگزون ۴ ژن *DGCR8* و اگزون ۹ ژن DALRD3 با KL_B= ۰/۰۰۱۹ و اگزون ۱ ژن *ACTR2* و اگزون ۱ ژن *HPS6* با مقدار KL_B= ۴۶۵۲/۴۳۵۲ به ترتیب کمترین و بیشترین فاصله را بر اساس این معیار داشتند.

در شکل ۲ محور X مقادیر KL_B و محور Y تعداد مقایسات را نشان می‌دهد. با توجه به وجود ۳۰ ژن و ۲۱۱ اگزون، ماتریسی به ابعاد ۳۰*۳۰ برای ژن‌ها و ۲۱۱*۲۱۱ برای اگزون‌ها ایجاد شد. برای روشن شدن موضوع، به طـور مثـال اولین میله در هیستوگـرام سمت راست شکل ۲ تعداد مقایساتی که در آن مقادیر KL_B نزدیک به صفر بودند را نشان می‌دهد که در این میان حدود ۱۹۰ ژن، خیلی شبیه هم بودند.

نتایج KL_B حاصل از ژن‌ها و اگزون‌ها به الگوریتم‌های خوشه‌بندی مختلفی وارد شدند. به علت محدودیت، در این مقاله تنها به نتایج روش *single linkage* در ژن‌ها اشاره شد (شکل ۳). درخت ژنی ایجاد شده بر اساس این روش، دو خوشه مجزا را نشان داد. یک خوشه شامل ژن‌های *HTRA2*، *SLC35A3*، *FAM192A*، *DGCR8*، *CNOT8*، *EIF3L*، *TBC1D20*، *CD34*، *SPSB3*، *DALRD3*، *HPS6*، *RPS3A*، *SNRPG*، *NR1H2*، *COPS6*، *ZNF419*، *CALM1*، *DES*، *ZDHHC4*، *GRWD1*، *CDK9*، *MRPS11*، *S100A16*، *YWHAH*، *C1H21orf59* و *TIMM21* و خوشه دیگر شامل ژن‌های *SMIM14*، *ACTR2*، *NSUN3* و *B4GALT1* بود.

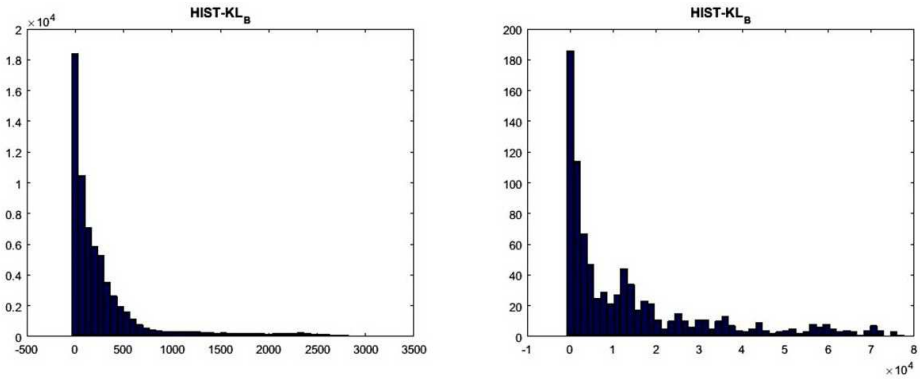


Fig. 2. Histogram of KL_B of genes (right) and exons (left) due to first order relative entropy. About 21% of data (190 data points) in genes and 40% of data (19000 data point) in exon were almost identically lumped around zero value

شکل ۲- هیستوگرام فراوانی مقادیر کولبک- لیبلر (KL_B) ژن‌ها (سمت راست) و اگزون‌ها (سمت چپ) (همانطور که مشاهده می‌شود حدود ۲۱ درصد از برآوردها در ژن‌ها (حدود ۱۹۰ مشاهده) و ۴۰ درصد در اگزون‌ها (حدود ۱۹۰۰۰مشاهده) کاملاً شبیه هم بوده و در مقادیر نزدیک صفر قرار گرفتند)

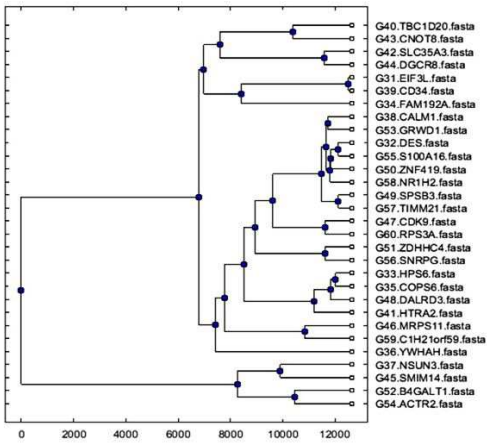


Fig. 3. KLD (KL-B) based gene clustering using single linkage method due to different nucleotide content

شکل ۳- خوشه‌بندی کولبک- لیبلر مبتنی بر تفاوت(KL_B) ژن‌ها با استفاده از روش Single linkage

روش دوم- خوشه‌بندی بر اساس واگرایی کولبک- لیبلر مبتنی بر شباهت نوکلئوتیدهای متناظر ژن ها و اگزون ها (KL_A): بر اساس نتایج حاصل، ژن‌های *HTRA2* و *COPS6* با مقدار KL_A=۱۴/۳۰۹ و ژن‌های *YWHAH* و *ACTR2* با KL_A=۹۳۵۰۴/۷۹۶ به ترتیب کمترین و بیشترین فاصله را داشتند. همچنین در بین ۲۱۱ اگزون، اگزون ۸ ژن *ACTR2* و اگزون ۲ ژن *FAM192A* با KL_A= ۲۲۱/۳۳۸ و اگزون ۱ ژن *ACTR2* و اگزون ۱ ژن *HPS6* با مقدار KL_A= ۷۴۰۷/۲۵۱ به ترتیب کمترین و بیشترین فاصله را بر اساس این معیار داشتند. همانند روش اول، نتایج KL_A به الگوریتم‌های خوشه‌بندی مختلفی وارد شدند. درخت ژنی ایجاد شده، دو خوشه مجزا و مشابه با روش اول را نشان داد با این تفاوت که خوشه اول که شامل ۲۶ ژن بود توپولوژی بسیار متفاوتی نسبت به روش اول داشت.

در مقایسه نتایج دو روش مشاهده شد که هر کدام از ژن‌های *COPS6* و *YWHAH* به ترتیب بیشترین شباهت و تفاوت را با ژن دیگر ایجاد کردند. همچنین مشاهده شد که اگزون ۱ ژن *ACTR2* و اگزون ۱ ژن *HPS6* بیشترین فاصله را در هر دو روش میان اگزون‌ها داشتند.

با مقایسه خوشه‌های روش *single*، در KL_A و KL_B مشاهده شد توپولوژی ژن‌های خوشه اول با هم متفاوت است. ژن *SPSB3* تغییر عمده‌ای از خود نشان داد. این ژن که در KL_B در کنار ژن‌های *CALM1*، *GRWD1*، *DES*، *S100A16*، *ZNF419*، *NR1H2*، *TIMM21*، *CDK9* و *RPS3A* قرار گرفته بود، در KL_A از این خوشه جدا و به

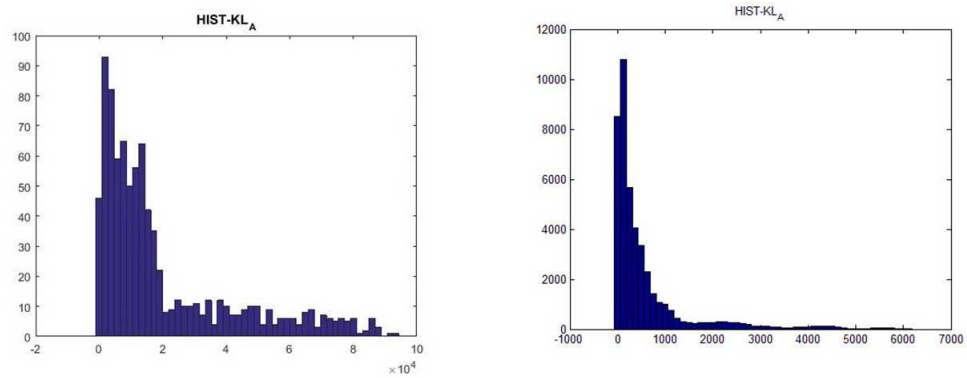


Fig. 4. Histogram of KL_A of genes (left) and exons (right) due to first order relative entropy. About 5% of data (45 data points) in genes and 19% of data (8500 data point) in exon were almost identically lumped around zero value

شکل ۴- هیستوگرام فراوانی مقادیر کولبک- لیبلر (KL_A) ژن ها (سمت چپ) و اگزون‌ها (سمت راست) در مرتبه اول آنتروپی (همانطور که مشاهده می‌شود حدود ۵ درصد از برآوردها در ژن‌ها (حدود ۴۵ مشاهده) و ۱۹ درصد در اگزون‌ها در (حدود ۸۵۰۰ مشاهده) کاملاً شبیه هم بوده و در مقادیر نزدیک صفر قرار گرفتند)

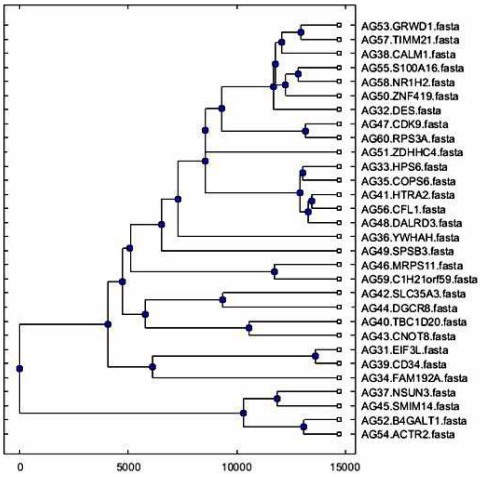


Fig. 5. KLD (KL_A) based gene clustering using single linkage method due to similarity nucleotide content

شکل ۵- خوشه‌بندی کولبک- لیبلر مبتنی بر شباهت (KL_A) ژن‌ها با استفاده از روش Single linkage

طور جداگانه در کنار ژن‌های دیگر قرار گرفت، این تغییر در شکل ۶ با کادر قرمز مشخص شده است. همچنین ژن‌های *SLC35A3* با *DGCR8* ، *FAM192A* با *EIF3L* و *CD34*، *MRPS11* با *C1H21orf59*، *HPS6* با *COPS6* هم در KL_A و هم در KL_B در کنار همدیگر و در یک خوشه و ژن *YWHAH* به صورت جدا قرار گرفت. نکته جالب که در شکل ۶ نیز نشان داده شد قرار گرفتن ژن‌های با طول نزدیک به هم در یک خوشه و نزدیک به هم بودند.

روش سوم- خوشه‌بندی بر اساس واگرایی کولبک- لیبلر مبتنی بر آنتروپی نسبی ژن‌ها و اگزون‌ها (KL_H): پس از محاسبه آنتروپی مراتب یک تا چهار ژن‌ها و اگزون‌ها، واگرایی کولبک- لیبلر برای هر مرتبه از آنتروپی به طور جداگانه محاسبه شد. این روش به ژن‌ها و اگزون‌ها این اجازه را داد که با طول حقیقی و متفاوت و با محتوای واقعی خود نسبت به یکدیگر مورد ارزیابی قرار گیرند. نتایج این بخش در جدول ۴ ارائه شده است. خوشه‌بندی ژن‌ها و اگزون‌ها بر پایه آنتروپی نسبی و بدون اعمال همترازی انجام شد. در شکل ۷، هیستوگرام فراوانی مقادیر آنتروپی ژن‌ها و اگزون‌های آنها بر اساس مرتبه یک نشان داده شد. کلیه نتایج در بخش ضمیمه ارائه شده است.

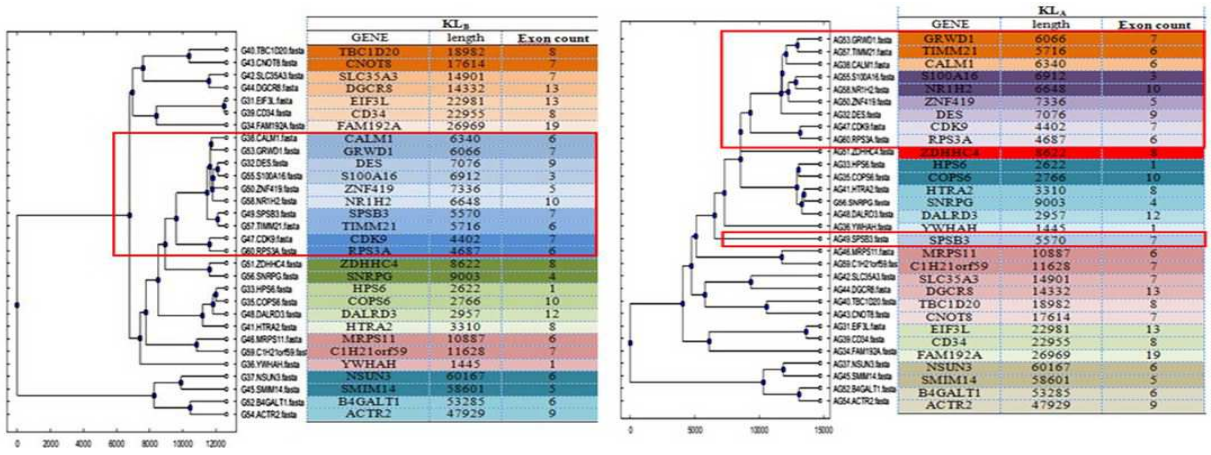


Fig. 6. Comparison between KL_A (right) and KL_B (left) based gene single linkage clustering

شکل ۶- مقایسه خوشه‌بندی کولبک - لیبلر KL_A (راست) و KL_B (چپ) ژن‌ها با استفاده از روش Single linkage

جدول ۴ - نتایج واگرایی کولبک- لیبلر (آنتروپی نسبی) ژن‌ها و اگزون‌ها بر اساس آنتروپی مراتب یک تا چهار

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

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

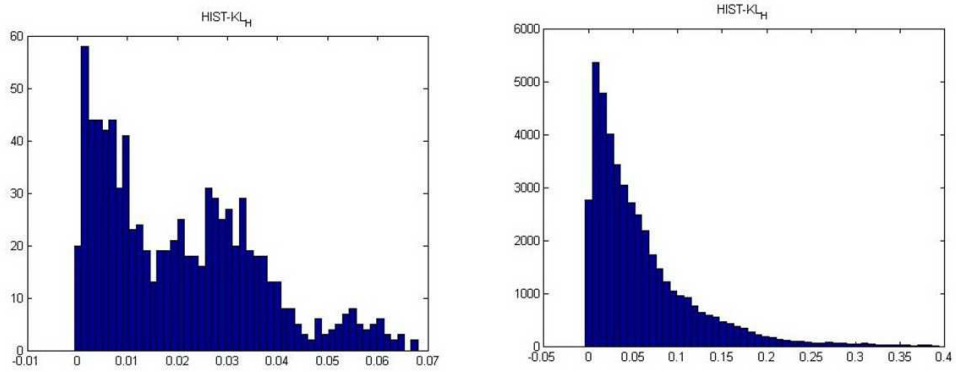


Fig. 7. Histogram of KL_H of genes (left) and exons (right) due to first order relative entropy. About 2.2% of data (20 data points) in genes and 6.6% of data (1800 data points) in exon were almost identically lumped around zero value

شکل ۷- هیستوگرام فراوانی مقادیر کولبک- لیبلر (آنتروپی نسبی) ژن‌ها (سمت چپ) و اگزون‌ها (سمت راست) بر اساس آنتروپی مرتبه یک (همانطور که مشاهده می‌شود حدود ۲/۲ درصد از برآوردها در ژن‌ها (حدود ۲۰ مشاهده) و ۶/۶ درصد در اگزون‌ها (حدود ۱۸۰۰ مشاهده) کاملاً شبیه هم بوده و در مقادیر نزدیک صفر قرار گرفتند)

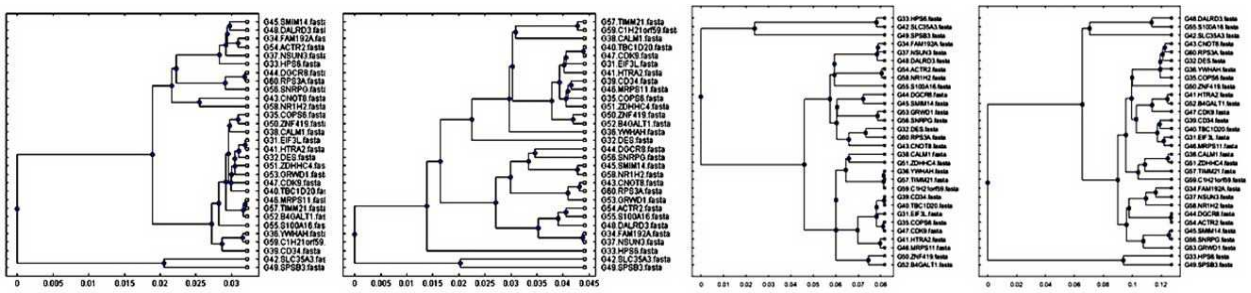
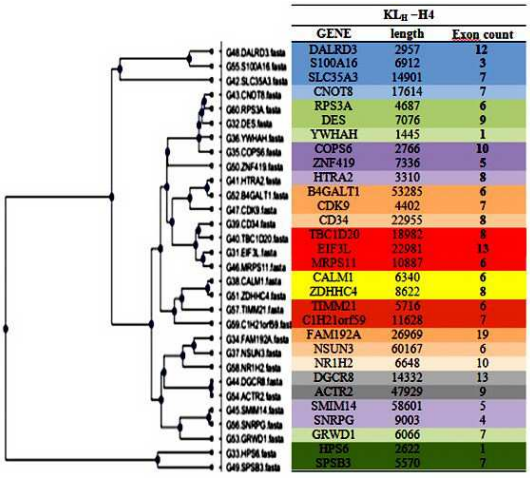


Fig. 8. Result of single clustering of genes due to KL_H. The rightest one is due to first order entropy and it goes to leftist one which is due to fourth entropy order

شکل ۸- نتایج خوشه‌بندی کولبک- لیبلر مبتنی بر آنتروپی نسبی (KL_H) ژن‌ها از مرتبه یک (چپ) به مرتبه چهار (راست) با استفاده از روش Single linkage

در روش سوم با توجه به اینکه در هر مرتبه آنتروپی از هفت الگوریتم خوشه‌بندی استفاده شد در مجموع، تعداد ۲۸ خوشه ایجاد شد. با مقایسه و بررسی همه خوشه‌های ایجاد شده با روش *Single* مشاهده شد 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 سه ژن بود- از دو ژن تشکیل شده بود. ژن *SPSB3* در همه مرتبه‌ها در گروه دوم به طور ثابت قرار گرفت و ژن *SLC35A3* که تا مرتبه سوم در کنار این ژن و در یک خوشه بود در مرتبه چهارم از این گروه جدا و به گروه اول و به ژن‌های *DALRD3* و *S100A16* اضافه شد. همچنین ژن *SLC35A3* در مرتبه سوم جایگزین ژن *HPS6* - که در مرتبه اول به سمت خوشه دوم نزدیک شده بود- شد. ژن‌های گروه اول نیز خود به دو شاخه فرعی عمده تفکیک شدند. ژن‌های *GRWD1* و *S100A16* در مرتبه دوم آنتروپی از شاخه فرعی اول به شاخه فرعی دوم در گروه اول منتقل شدند. همچنین ژن *DES* در مرتبه سوم آنتروپی از خوشه خود جدا و به خوشه فرعی دوم اضافه شد. همچنین 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 خوشه‌ها با تغییر مرتبه آنتروپی از یک به چهار، تغییر یافت که عمده‌ترین تغییرات را در خوشه‌بندی‌های مرتبه چهار شاهد بودیم. همچنین مشاهده شد که روش KL_H مستقل از طول ژن و کاملاً با محتوای ژن و فراوانی نوکلئوتیدها در توالی در ارتباط بود (شکل ۹).



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	2822	1
SPSB3	5570	7

Fig. 9. Lack of dependency of KL_H due to forth order of entrop on the length of genes in single clustering

شکل ۹- عدم وابستگی واگرایی کولبک- لیبلر (KL_H) ناشی از آنتروپی مرتبه چهارم ژن‌ها به طول در خوشه‌بندی به روش Single linkage

تجمیع نتایج حاصل از خوشه‌بندی: در این بخش از پژوهش بر اساس اطلاعات مفیدی که از هفت روش معمول خوشه‌بندی ژن‌ها بدست آمد، نتایج تجمیع شود. در این راه از دسته‌بند *AdaBoost* استفاده شد. در شکل ۱۰ خوشه‌بندی نهایی بر اساس تجمیع نتایج *AdaBoost* در سه روش محاسبه واگرایی کولبک- لیبلر نشان داده شده است. دسته‌بند *AdaBoost* در روش اول (KL_B) ژن‌ها را به چهار دسته، در روش دوم (KL_A) به سه دسته و در روش سوم (KL_H) به دو دسته عمده تقسیم نمود. با توجه به ارتولوگ بودن این ژن‌ها با انسان، ارتباط بین آنها و عملکردشان در هر سه روش با مراجعه به تارگاه *GeneMANIA* مورد بررسی و پیش‌بینی قرار گرفت تا صحت خوشه‌بندی روش‌ها بررسی و مقایسه شود.

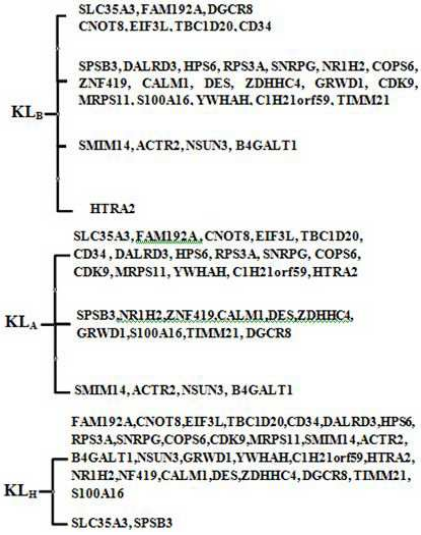


Fig. 10. The results of aggregation of clustering results with the Adaboost algorithm over investigated genes

شکل ۱۰- نتایج حاصل از ترکیب دسته‌بند *AdaBoost* در سه روش و در ژن‌های مورد بررسی

بررسی دقت نتایج حاصل از دسته‌بند AdaBoost در روش واگرایی کولبک- لیبلر مبتنی بر تفاوت (KL_B): طبق نتایج حاصل از دسته بند *AdaBoost* در روش اول (KL_B) ژن‌ها به چهار گروه اصلی تقسیم شدند. با مراجعه به تارگاه *GeneMANIA* در بررسی ژن‌های خوشه اول (۷ ژن)، خوشه دوم (۱۸ ژن)، خوشه سوم (۴ ژن) و خوشه چهارم (۱ژن)، شاهد ارتباط این ژن‌ها با تعدادی از ژن‌های دیگر بودیم که به ترتیب ۶/۱٪، ۸۷/۴۵٪، ۱۳/۵٪ و ۱۳/۵٪ هم-بیانی[1] نشان دادند. به جز ژن *NSUN3* در خوشه سوم که تنها یک ارتباط ضعیف با ژن *SMIM14* داشت ژن‌های *GRWD1*، *ZNF419*، *S100A16*، *SPSB3*، *DALRD3*، *HPS6* و *ZDHHC4* نیز در خوشه دوم ارتباط خیلی ضعیفی با ژن‌های دیگر نشان دادند که ناشی از عدم وجود مسیرهای متابولیکی مشترک بود و صحت خوشه‌بندی گروه دوم و سوم را تا حد زیادی تحت تاثیر قرار داد.

بررسی دقت نتایج حاصل از دسته‌بند AdaBoost در روش واگرایی کولبک-لیبلر مبتنی بر شباهت (KL_A): تارگاه *GeneMANIA*، هم‌بیانی ژن‌های خوشه اول (۱۰ژن)، خوشه دوم (۱۶ ژن) و خوشه سوم (۴ ژن) با ژن‌های دیگر را به ترتیب ۹۱/۱٪، ۶۰/۳۳٪، ۶۷/۶۴٪ نشان داد. ژن‌های *CD34*، *SLC35A3* و *TBC1D20*، در خوشه اول و ژن‌های *SPSB3*، *NR1H2*، *ZNF419*، *DES*، *GRWD1*، *TIMM21*، *DGCR8* در خوشه دوم هیچ ارتباطی با ژن‌های مرتبط دیگر نداشتند، همچنین بین ژن‌های *CALM1* و *S100A16* تنها یک ارتباط ضعیف مشاهده شد. در این میان تنها دو ژن *ZDHHC4* و *DGCR8* به طور مجزا با تعدادی از ژن‌ها ارتباط متقابل و عملکرد مشترک نشان دادند. همچنین ژن *NSUN3* در خوشه سوم فقط با ژن *SMIM14* ارتباط متقابل نشان داد.

بررسی دقت نتایج حاصل از دسته‌بند AdaBoost در روش واگرایی کولبک- لیبلر مبتنی بر آنتروپی نسبی (KL_H): طبق نتایج حاصل از دسته‌بند *AdaBoost* و بر اساس نتایج واگرایی کولبک- لیبلر مبتنی بر آنتروپی نسبی (KL_H) ژن‌ها به دو گروه اصلی تقسیم شدند. با مراجعه به تارگاه *GeneMANIA* و در بررسی ژن‌های خوشه اول (۲۸ ژن) و خوشه دوم (دو ژن)، شاهد ارتباط این ژن‌ها با

1. Co-expression

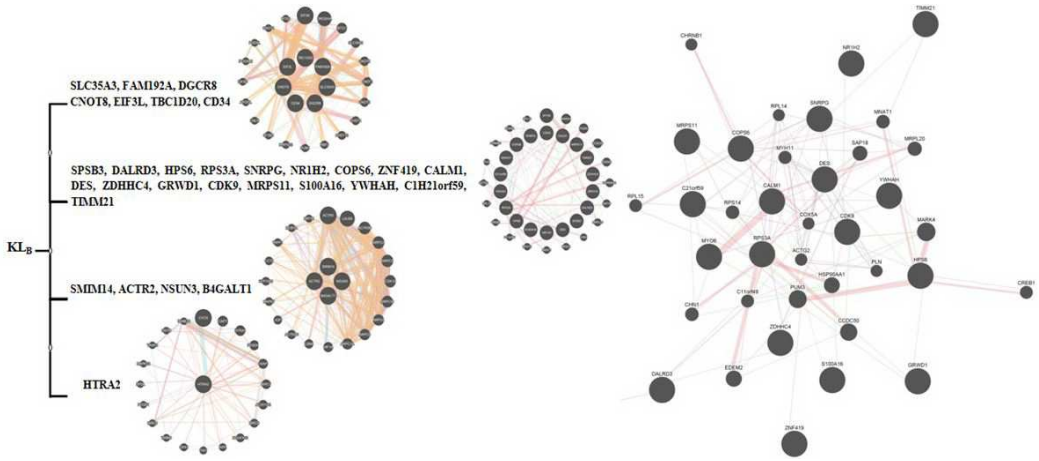


Fig. 11. Results of Adaboost classifier genes based on KL_B (left) and the interrelationship among genes in the second cluster (right)

شکل ۱۱- نتایج دسته‌بند *AdaBoost* ژن‌ها بر اساس روش کولبک- لیبلر KL_B (چپ) و ارتباط درونی بین ژن‌های خوشه دوم (راست)

تعدادی از ژن‌های دیگر بودیم که به ترتیب ۷۵/۷۳٪ و ۶۷/۶۴٪ هم بیانی 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 شد که ژن‌های هر خوشه وظایف متفاوتی داشته و انتظار می‌رود مسیرهای متابولیکی آنها هم متفاوت باشد. بررسی‌ها نشان داد که ژن‌های دو خوشه با یکدیگر مسیر متابولیک مشترک نداشته و هیچ کدام از ژن‌های خوشه اول در ارتباط متقابل با ژن‌های خوشه دو نبودند و بالعکس. همچنین با بررسی عملکرد ژن‌های خوشه اول و دوم مشخص شد که هیچ‌یک عملکرد یکسان و مشابهی نداشته و ژن‌های هر خوشه مسیرهای متابولیکی متفاوت و متمایزی را کددهی نموده، لذا عملکرد مشترک و یکسانی در دو خوشه مشاهده نشد که این خود به نوعی تاییدی بر قدرت و دقت روش ارایه شده بود.

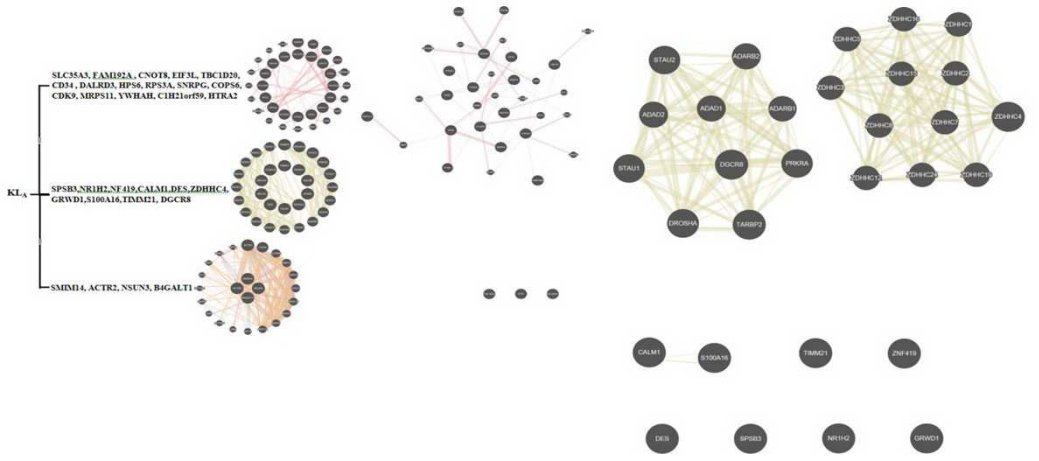


Fig. 12. Results of Adaboost classifier genes based on KL_A (left) and the interrelationship among genes in the first (middle) and second (right) cluster

شکل ۱۲- نتایج دسته‌بند *AdaBoost* ژن‌ها بر اساس روش کولبک- لیبلر KL_A (چپ) و ارتباط درونی بین ژن‌های خوشه‌های اول (وسط) و دوم (راست)

جدول ۵- وظایف و عملکرد ژن‌های خوشه اول و دوم

Table 5. Functions of the first and second gene clusters

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

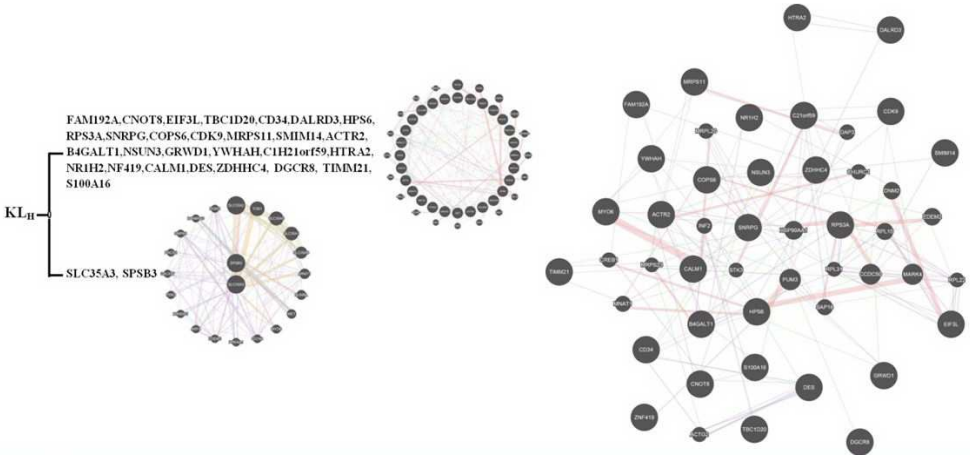


Fig. 13. Results Adaboost classifier genes based on KL_H (left) and the interrelationship among genes in the first cluster (right)

شکل ۱۳- نتایج دسته‌بند *AdaBoost* ژن‌ها بر اساس روش کولبک- لیبلر KL_H (چپ) و ارتباط درونی بین ژن‌های خوشه اول (راست)

مقایسه نتایج حاصل و بررسی عملکرد، ارتباط متقابل و مسیرهای متابولیکی مشترک ژن‌ها، روش خوشه‌بندی بر اساس واگرایی کولبک- لیبلر مبتنی بر آنتروپی (KL_H) را معیاری صحیح‌تر، منطقی‌تر و در عین حال سریع‌تر نشان داد. این روش علاوه بر اینکه معایب همتراز نمودن ژن‌ها را نداشت، به طور مستقیم محتوا و شکل واقعی ژن را مورد بررسی قرار داده و در مقایسه با دو روش دیگر پیچیدگی محاسباتی و نیاز به حافظه بالا برای توالی‌های با طول بزرگ را نداشت لذا با توجه به اینکه محتوای اطلاعات درون توالی DNA از دست نرفت، صحت خوشه-بندی داده‌های توالی افزایش یافت. کلیه محاسبات در روش‌های مورد پژوهش به وسیله نرم‌افزار مهندسی متلب انجام شد. برای محاسبات KL_H از یک رایانه شخصی با مشخصات: پردازشگر *Intel Core i5 CPU 2.40 GHz* و حافظه ۴ گیگابایت و برای محاسبات KL_A و KL_B از ابررایانه مرکز ملی ابررایانش شیخ بهایی دانشگاه اصفهان با مشخصات پردازشگر *Intel Xeon Core i24 CPU 3 GHz* و حافظه ۲۳ گیگابایت استفاده شد. مدت زمان انجام محاسبات در سه روش انجام شده برای ژن‌ها و اگزون‌های مورد بررسی در جدول ۶ نشان داده شده است. همانطور که مشاهده شد محاسبات در KL_H در مدت زمان کمتر و با رایانه با مشخصات فنی پایین‌تر انجام شد.

با توجه به ارتباط مستقیم فرمول محاسبه واگرایی کولبک- لیبلر به آنتروپی و حساسیت به مقادیر آن، در KL_H نشان داده شد که ژن‌ها و اگزون‌هایی که کمترین و بیشترین مقادیر آنتروپی را در رتبه های یک تا چهار کسب نمودند دقیقاً بیشترین فاصله ژنی در کولبک- لیبلر مبتنی بر آنتروپی را داشتند. نتایج نشان داد که اگر آنتروپی دو قطعه DNA مشابه به هم باشند، آنگاه، فاصله کولبک- لیبلر آنها صفر خواهد شد و انتظار می‌رود آن

جدول ۶ - مقایسه زمان پردازش در هر روش
Table 6. Running time comparison in each method

Datasets	KL_H				KL_A	KL_B
	H1	H2	H3	H4		
Genes (30)	0.076s	0.599s	0.778s	0.722s	10h 52min 33s	7 h 24min 18 s
Exone (211)	0.149s	0.143s	0.157s	0.703s	3h 44min 27s	51min 46s

دسته از قطعات DNA که چنین خاصیتی را داشته یا واگرایی کولبک- لیبلر آن‌ها به صفر نزدیک باشد (خصوصاً در آنتروپی‌های مراتب بالا)، احتمالاً یک نقش زیستی مشابه دارند. نتایج حاصل از واگرایی کولبک- لیبلر روی قطعات DNA نشان داد که ساخت و ترکیب DNA ژن‌ها، اگر به فضای دیگری نگاشت شود (مثل فضای آنتروپی)، می‌تواند مشابهت‌های عملکردی آن‌ها را آشکار سازد. آن دسته از توالی‌های DNA که ساختار یکسانی دارند، احتمالاً یک نوع پروتئین را کد می‌کنند و در نتیجه نقش عملکردی و زیستی یکسان داشته، بنابراین امکان استخراج شبکه متابولیتی بین توالی‌های زیستی یک سازواره به طور نسبی وجود خواهد داشت. البته این در صورتی درست می‌باشد که هیچ‌گونه اطلاعات زیستی دیگری موجود نباشد. پژوهش‌های مختلفی در این خصوص انجام گرفته است. در پژوهشی، از نظریه درختچه حیات برای تحلیل مسیرهای متابولیتی استفاده شد. این پژوهش از اولین پژوهش‌هایی محسوب می‌شود که ترکیب داده‌های سطح DNA و مسیرهای متابولیکی با استفاده از درختچه حیات انجام شد (Forst and Schulten, 2001). Lee *et al.* (2009) از واگرایی کولبک- لیبلر به عنوان روشی نو در بازسازی درخت فیلوژنتیک کورنوویروس و سارس ویروس‌ها استفاده کردند. همچنین پژوهش‌هایی جهت استفاده هر چه بیشتر داده‌های متابولیکی برای درک بهتر ارتباط تکاملی گونه‌های مختلف (Clemente *et al.*, 2007)، با توجه به انباشت داده‌های متابولیکی و با استفاده از نظریه گراف انجام شد که نشان‌دهنده همخوانی درختچه فیلوژنی ایجاد شده با نتایج آزمایشگاهی بود (Clemente *et al.*, 2007; Heymans and Singh, 2003). 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 شد که بنیاد نظری ایجاد کننده درختچه تکاملی را می‌توان برای ارتباط متابولیکی نیز به کار برد.

نتیجه‌گیری کلی

در روش ارائه شده از خوشه‌بندی به یک گروه‌بندی زیستی از ژن‌ها دست یافتیم. با توجه به استخراج ویژگی‌های حاصل شده از نتایج خوشه بندی، از این روش نو و بدیع می‌توان در خوشه‌بندی ژن‌های دیگر استفاده نمود. نتایج نهایی خوشه‌بندی و بررسی عملکرد ژن‌های هر خوشه، روش ارائه شده در این پژوهش را برای خوشه-بندی ژن‌ها مورد تایید قرار داد.

همچنین الگوریتم یاد شده می‌تواند 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[1] با استفاده از واگرایی کولبک - لیبلر که مبتنی بر آنتروپی ژن‌ها و دو روش دیگر، که مبتنی بر همترازی اولیه توالی DNA است، 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 دسته‌بندی‌کننده چندگانه[2] و یا سامانه‌های شورایی حل می‌کند. یکی از مطرح‌ترین این روش‌های شورایی *AdaBoost* است (Freund and Schapire, 1996) که این پژوهش از آن بهره برد.

روش ارایه شده در این مقاله می‌تواند برای اختصاص دادن و پیشی‌بینی فعالیت زیستی برای آن دسته از ژن‌هایی که حاشیه‌نویسی ژنومی قوی ندارند، کمک‌کننده 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 حاشیه‌نویسی ژنومی قوی دارند با آن‌هایی که ندارند، می‌تواند ارزش

1. Free-Alignment
2. Multiple classifier system

افزوده تحلیل زیستی به گروه دوم ژن‌ها (بدون حاشیه-نویسی ژنومی) بدهد.

تشکر و قدردانی

نویسندگان مراتب تقدیر و تشکر صمیمانه خود را از مهندس کامیار شیوعی، دکتر سعید انصاری مهیاری، شرکت رایان زیست فنآوری پارس آوید (www.Avidbiotech.com) و همچنین دست اندرکاران مرکز ابررایانش ملی شیخ بهایی به جهت استفاده از امکانات پردازشی آن مرکز اعلام می‌نماید. این مرکز تحت حمایت معاونت علمی و فنآوری ریاست جمهوری و دانشگاه صنعتی اصفهان است.

فهرست منابع

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(79): 1-10.

Changchuan Y., Ying C. and Stephen Y. 2014. A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. Journal of Theoretical Biology, 359: 18–28.

Clemente J. C., Satou K. and Valiente G. 2007. Phylogenetic reconstruction from non-genomic data. Bioinformatics, 23: 110–115.

Edwards S. V., Fertil B., Giron A. and Deschavanne P.J. 2002. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. System Biology, 51: 599-613.

Erill I. 2012. Information Theory and biological sequences: Insights from an evolutionary prespective. 2012 Nova Science Publishers, Inc.

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.

Freund Y. and Schapire R. 1996. Experiments with a new boosting algoritm. Paper read at Proceeding of the Thirteenth Internatioanal Conference on Machine Learning.

Forst C. V. and Schulten K. 2001. Phylogenetic analysis of metabolic pathways. Journal of Molecular Evolution, 52: 471–489.

Ghaderi-Zefrehei M., Bandi Dastjerdi A., Bahreini Behzadi A., 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(2): 1-22.

Gray R. M . 2013. Entropy and Information Theory. First Edition. Springer-Verlag New York publisher.

Heymans M. and Singh A. K. 2003. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics, 19 (1): 138–146.

Jiang S., 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.

Khatib H., Monson R. L., Schutzkus V., Kohl D. M., Rosa G. J. M. 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.

Kim J., Kim S., Lee K. and Kwon Y. 2009. Entropy analysis in yeast DNA. Chaos, Solitons and Fractals, 39: 1565–1571.

Kullback S. and Leibler R. 1951. On information and sufficiency. The Annals of Mathematical Statistics, 22: 79–86.

Lee L. 2009. Used kullback-Liebler measure as a new method for the reconstruction of the phylogenetic tree of the Cornavirus and SARS viruses.

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.

Li C. and Wang J. 2005. Relative entropy of DNA andits application. Physica A, 347: 465–471.

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.

Liu B. 2007. Uncertainty Theory, 2nd ed., Springer-Verlag, Berlin.

Machado J. T. 2012. Shannon entropy analysis of the genome code. Mathematical Problems in Engineering, 2012:1-12.

Monge R. E. and Crespo J. L. 2014. Comparison of Complexity Measures for DNA Sequence Analysis. 2014 International Work Conference on Bio-inspired Intelligence (IWOBI).

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(4): 1078–1087.

Pham T. D., Crane D. I., Tannock D. and Beck D. 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.

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.

Qi J., Wang B. and Hao B. 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. Journal of Molecular Evolution, 58: 1-11.

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.

Shannon C. 1948. A mathematical theory of communication. Bell System Technical Journal, 27: 379-423 and 623-656.

Sherwin B. W. 2010. Entropy and Information Approaches to Genetic Diversity and its Expression: Genomic Geography Entropy, 12: 1765-1798.

Stuart G. W., Moffet K. and Baker S. 2002. Integrated gene species phylogenies from unaligned whole genome protein sequences. Bioinformatics, 18: 100-108.

Stuart G. W., Moffet K. and Leader J. J. 2002. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Molecular Biology and Evolution, 19: 554-562.

Sundekilde U. K., Larsen L. B. and Bertram H. C. 2013. NMR-Based Milk Metabolomics. Metabolites, 3: 204-222.

Tautz D., Trick M., Dover G. A. 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature, 322: 652–656.

Vinga S., Almeida J. 2003. Alignment-free sequence comparison: review. Bioinformatics, 19 (4): 513-523.

Vinga S. 2013. Information theory applications for biological sequence analysis. Briefings in bioinformatics. 15 (3): 376-389.

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, Web Server issue doi:10.1093/nar/gkq537.

Xie X., Yu Y., Liu G., Yuan Z. and Song J. 2010. Complexity and entropy analysis of DNA methyltransferase. Journal of Data Mining in Genom Proteomics, 1(2): 100-105.

Yu Z. G., Anh V. and Lau K. S. 2003. Multifractal and correlation analysis of protein sequences from complete genome, Physics Review E, 68: 021913.

Yu Z. G., Anh V. V. and Zhou L. Q. 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.

Yu Z. G., Zhou L. Q., Anh V., Chu K. H. 2005. 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.

Zhang J. L., Zan L. S., Fang P., Zhang F., Shen G. L. and Tian W. Q. 2008. Genetic variation of PRLR gene and association with milk performance traits in dairy cattle. Canadian Journal of Animal Science, 88: 33-39.

Zhou L. Q., Yu Z. G., Anh V., Nie P. R., Liao F. F. and Chen Y. J. 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 3nd International Conference on Natural Computation (ICNC2007), Haikou, China, August 2007; pp: 304–308.

Animal Production Research

Vol. 6, No. 3, 2017 (11-28)

University of Guilan

Gene tree construction using Kullback-Leibler divergence on milk governing genes in dairy cattle

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

1. Ph.D Student, Department of Animal Science, Faculty of Agricultural Sciences, University of Guilan, Rasht, Iran
2. Professor, Department of Animal Science, Faculty of Agricultural Sciences, University of Guilan, Rasht, Iran
3. Assistant Professor, Department of Animal science, Faculty of Agricultural Sciences, University of Yasouj, Yasouj, Iran
4. Assistant Professor, Department of Electrical Engineering, Faculty of Electrical Engineering, University of Guilan, Rasht, Iran
5. Associate Professor, Department of Biotechnology, Animal Science Research Institute, Agricultural Research, Education and Extension Organization (AREEO), Karaj, Iran

(Received: 27-08-2017 – Accepted: 07-12-2017)

Abstract

Information theory is a branch of mathematics that overlaps with communications, biology. The aim of the current study was to provide a method for clustering a number of Milk Governing Genes in Dairy Cattle using an algorithm based on Kullback-Leibler divergence. In this study, after retrieving gene and exon DNA sequences affecting milk yield in dairy cattle, the entropy in orders one to four was calculated. In order to extract gene distances, Kullback-Leibler divergence over three different methods was calculated. The first and second methods were based on the genes alignment but the third method was based on non-alignment and the relative entropy of the genes. The results of each method of Kullback-Leibler divergence over DNA and exon sequences were entered as input into 7 general clustering algorithms: *Single*, *Complete*, *Average*, *Weighted*, *Centroid*, *Median* and *K-Means*. Integrated result of each clustering algorithm due to AdaBoost algorithm, which implied as gene tree, indicated that the third method was based on the relative entropy of the genes, 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: Information theory, Gene clustering, Dairy cattle, Kullback-Leibler divergence

*Corresponding author: mirhosin@guilan.ac.ir