علوم دامی ایران
دورهٔ ۴۶، شمارهٔ ۳، پاییز ۱۳۹۴ (ص ۲۸۷-۲۷۷)

بررسی ساختارهای جوامع و خرده‌جوامع به روش خوشه‌بندی شبکه‌ای بدون نظارت با استفاده از نشانگرهای ژنتیکی متراکم

جواد رحمانی‌نیا[۱]، سیدرضا میرائی آشتیانی[۲*] و حسین مرادی شهربابک[۳]
۱، ۲ و ۳. دانشجوی دکتری ژنتیک و اصلاح نژاد، استاد و استادیار، گروه مهندسی علوم دامی، پردیس کشاورزی و منابع طبیعی دانشگاه تهران، کرج
(تاریخ دریافت: ۱۳۹۳/۱۲/۹ - تاریخ تصویب: ۱۳۹۴/۶/۷)

چکیده

رشد روزافزون اطلاعات حاصل از تعیین ژنوتیپ نمونه‌ها به‌ویژه با استفاده از توالی‌یابی چندشکلی‌های تک‌نوکلئوتیدی (SNP) سبب تحول در تجزیه‌وتحلیل دقیق ساختار جوامع در گونه‌های مختلف شده است. تاکنون از روش‌های مختلفی برای بررسی ساختار جمعیتی با استفاده از نشانگرهای موجود در کل ژنوم استفاده شده است که هرکدام نقاط ضعف و قوتی دارند. در بررسی حاضر از خوشه‌بندی شبکه‌ای بدون نظارت یا SPC که روشی مبتنی بر داده‌کاوی است، برای بررسی ساختار جوامع شبیه‌سازی‌شده و کشف خرده‌جوامع موجود استفاده شد. هدف از به‌کاربردن این روش، دستیابی به ساختار جمعیتی بدون هیچ‌گونه آگاهی از اطلاعات شجره‌ای افراد بود. در شبیه‌سازی انجام‌گرفته بدین منظور پس از ویرایش داده‌ها، ۲۹۲۰۹ نشانگر اتوزومی از ۱۵۹ دام، تجزیه‌وتحلیل شدند. نتایج نشان داد که حیوانات براساس شباهت‌ها و تفاوت‌ها به‌خوبی در جوامع مربوطه قرارگرفتند و خرده‌جوامع موجود نیز درون هر جمعیت نمایان شدند. مزیت اصلی این روش، کارایی محاسباتی بالا و نیازنبودن به فرض‌های پیشین در آن است؛ بنابراین به محقق این امکان را می‌دهد که ساختار جوامع متشکل از هزاران حیوان را بدون داشتن هرگونه اطلاعاتی از شجره و نژاد، تجزیه‌وتحلیل کند.

واژه‌های کلیدی: چندشکلی‌های تک‌نوکلئوتیدی، خوشه‌بندی شبکه‌ای بدون نظارت، داده‌کاوی، ساختار جمعیتی.

مقدمه

امروزه نشانگرهای ژنتیکی نقش مهمی در مطالعات مرتبط با منشأ، تاریخچه و سیر تکامل جوامع دامی ایفا می‌کنند (Fernández *et al*., 2013; Lirón *et al*., 2002; Troy *et al*., 2001). پیشرفت‌های کنونی در زمینهٔ توالی‌یابی ژنوم با تراکم بالا با استفاده از تراشه‌های نشانگری، توسعهٔ نرم‌افزارهای مربوطه و علم بیوانفورماتیک، استفاده از چندشکلی‌های تک‌نوکلئوتیدی (SNP) را بسیار مرسوم کرده است؛ هرچند به دلیل دوآللی‌بودن نشانگرهای SNP، اطلاعات ژنتیکی آشکارشده توسط آن‌ها در مقایسه با نشانگرهای ریزماهواره‌ای کمتر است، به‌طوری‌که برای رسیدن به دقتی معادل با دقت ریزماهواره‌ها حداقل دو تا شش برابر نشانگرSNP بیشتر، برای تعیین هویت افراد و مطالعهٔ روابط خویشاوندی نیاز است (Morin *et al*., 2004) اما SNPها دارای مزایایی چشم‌پوشی‌ناپذیر هستند که از آن جمله می‌توان فراوانی بیشتر (Heaton *et al*., 2005)، پایداری ژنتیکی آن در

* تلفن: ۰۲۶-۳۲۲۴۸۰۸۲
E-mail: ashtiani@ut.ac.ir

پستانداران (Markovtsova *et al*., 2000)، نام‌گذاری آسان و مناسب‌بودن برای تجزیه‌وتحلیل‌ها و تفسیر نتایج (Wang *et al*., 1998) را نام برد.

تنوع ژنتیکی موجود در بین جوامع، حاصل عوامل بسیاری ازجمله مهاجرت، جهش، اختلاط جمعیت‌ها و پدیدهٔ گلوگاه جمعیتی است. این تنوع نشانه‌هایی از رانش تصادفی و انتخاب را با خود به همراه دارد. آثار متقابل این عوامل می‌تواند سبب ایجاد جوامع و خرده‌جمعیت‌های مرتبط با آن شود. برای به‌دست‌آوردن دیدگاهی دربارهٔ تاریخچه و سیر تکامل جوامع، مشخص‌کردن خرده‌ساختارهای جمعیتی بسیار بااهمیت است. علاوه بر این معین‌کردن خرده‌ساختارهای جمعیتی و اختصاص افراد به خرده‌جمعیت‌های مربوط به خود در مطالعات ارتباط صفات با ژن‌ها (نشانگرها) که می‌بایست برای کاهش ارتباط‌های کاذب[1] این خرده‌ساختارها را در نظر گرفت (Serre *et al*., 2008) و همچنین فراهم‌کردن اطلاعاتی ارزشمند در جهت حفظ ذخایر ژنتیکی، مفید است (Bowden *et al*., 2012). امروزه مطالعات متنوعی برای مشخص‌کردن خرده‌ساختارهای جمعیتی در دام‌ها (Decker *et al*., 2009; Gautier *et al*., 2010; Kijas *et al*., 2009) صورت گرفته است. ازآنجاکه با گسترش تراشه‌های SNP، امروزه محققان اطلاعات نشانگری با تراکم بالا را در اختیار دارند، داشتن ابزاری برای تجزیه‌وتحلیل ساختار جامعه که قادر باشد از این حجم اطلاعات بهره برده و بدون نیاز به پیشینهٔ اجدادی، خرده‌ساختارهای جمعیتی را معین کند، لازم و ضروری است.

تا به امروز روش‌های آماری بسیاری برای بررسی ارتباط بین جوامع و مشخص‌کردن موقعیت افراد در جوامعی که از آن نشئت گرفته‌اند، پیشنهاد شده است. دو روش مهم از میان این روش‌ها، روش‌های خوشه‌بندی مبتنی بر فاصله و روش‌های مبتنی بر مدل هستند. در روش اول از فنون کاهش ابعاد از قبیل تجزیه‌وتحلیل مؤلفه‌های اصلی (PCA) به همراه ابزارهای خوشه‌بندی همچون K-means استفاده می‌شود (Lee *et al*., 2009; Paschou *et al*., 2007). در روش دوم، از روش‌های استاندارد آماری برای برآورد پارامترهای جمعیتی استفاده می‌شود و معمولاً برای هر جامعه فرض تعادل هاردی وینبرگ لحاظ می‌شود؛ بنابراین اگر نمونه‌های تهیه‌شده کوچک باشند، به دلیل برآورد ناصحیح فراوانی آللی، استنتاج‌های مناسبی به دست نخواهد آمد (Gao & Starmer, 2007). ازجمله نرم‌افزارهای مناسب برای دومین روش می‌توان برنامه‌های STRUCTURE (Pritchard *et al*., 2000)، ADMIXTURE (Alexander *et al*., 2009) و FRAPPE (Tang *et al*., 2005) را نام برد. این دو روش دارای محدودیت در فرض‌های پیشین و عدم توانایی پشتیبانی داده‌های گستردهٔ ژنومی موجود و همچنین مشکل به تصویر درآوردن و تفسیر نتایج هستند.

روشی دیگر که در این بررسی استفاده شده است، کاربرد رویکردی برای خوشه‌بندی شبکه‌ای بدون نظارت در آنالیز ساختار جمعیتی با استفاده از چندشکلی‌های تک‌نوکلئوتیدی است که آن را در اصطلاح خوشه‌بندی Super Paramagnetic (SPC) می‌نامند. ماهیت این روش به لحاظ اصولی با داده‌کاوی[2] و کشف دانش[3] مرتبط است. کشف دانش ابزاری جهت معین‌کردن، کشف داده و استخراج اطلاعات کاربردی بالقوه یا کشف الگوهاست. اصولاً هدف کشف دانش، کمک به درک داده‌ها و کاربرد آن‌هاست. از سویی دیگر داده‌کاوی بخشی ازکشف دانش است که به‌عنوان ابزاری برای معین‌کردن و کسب آگاهی از داده‌هایی که این امر در آن‌ها به سهولت مشهود نیست، استفاده می‌شود. این روش دارای مزایای بسیاری است که از آن جمله می‌توان قوی‌بودن در برابر اختلالات[4] و منشاء[5]، توانایی در تولید ساختار سلسله‌مراتبی خودبه‌خودی و از همه مهم‌تر نیازنداشتن آن به پیش‌آگاهی دربارهٔ توزیع داده‌ها یا ساختار خوشه را نام برد (Radjiman & Sugiarto, 2005).

1. Spurious associations

2. Data mining
3. Knowledge discovery
4. Noise
5. Initialization

هدف از تحقیق حاضر استفاده از خوشه‌بندی شبکه‌ای بدون نظارت در تجزیهٔ ساختار جمعیتی به کمک نشانگرهای متراکم SNP و استنتاج ساختار جمعیتی بدون پیش‌آگاهی از انساب افراد است که خود چالشی پیش‌ِروی تجزیه‌وتحلیل‌های ژنتیکی است.

مواد و روش‌ها

شبیه‌سازی جوامع

به‌منظور بررسی ساختار جمعیتی با استفاده از روش SPC، سه جامعه با ویژگی‌های متفاوت به همراه نشانگرهای آن‌ها شبیه‌سازی شدند. برای اجرای شبیه‌سازی از نرم‌افزار QMsim استفاده شد (Sargolzaei & Schenkel, 2009). این نرم‌افزار برای شبیه‌سازی دامنهٔ گسترده‌ای از انواع ساختارهای ژنتیکی و جوامعی با ساختارهای مختلف از دام‌های اهلی، طراحی شده است. شبیه‌سازی داده‌های ژنوتیپی در مقیاسی گسترده به همراه شجرهٔ پیچیده از قابلیت‌های این نرم‌افزار است.

برای مطالعهٔ حاضر سه جامعه متشکل از دام‌هایی با ۲n=۶۰ شبیه‌سازی شدند. برای این شبیه‌سازی، صفتی با توارث‌پذیری ۰/۲۷ و واریانس فنوتیپی ۱۳۰۰۰۰۰ در نظر گرفته شد (Behzadi *et al.*, 2013; Faraji-Arough *et al.*, 2015; Ojango & Pollott, 2001). شبیه‌سازی در این نرم‌افزار در دو مرحله انجام گرفت. ابتدا نسل‌های تاریخی[1] برای ایجاد سطح مطلوب LD شبیه‌سازی شدند. بدین منظور ۲۶۰ نسل به عقب برگشته و در نسل صفر ۵۰۰۰ دام شبیه‌سازی شدند و سپس این دام‌ها به‌صورت تصادفی آمیزش داده شدند و در شرایط مختلف قرار گرفتند، به‌طوری که در نسل ۲۶۰، تعداد ۱۶۸۰ دام باقی ماند (به نسل‌های ۰ تا ۲۶۰ عنوان نسل‌های تاریخی اطلاق می‌شود). در نسل آخر تاریخی تعداد نرها حداقل ۱۵۰ رأس منظور شد. در مرحلهٔ بعدی، ساختار جوامع اخیر به نحوی تولید می‌شود که این جوامع برحسب پارامترهایی که برای برنامه معین می‌شوند، بتوانند پیچیدگی‌های خاص خود را داشته باشند و داده‌های شبیه‌سازی مدنظر کاربر را فراهم آورند (Sargolzaei & Schenkel, 2009). سپس برای ایجاد سه جامعهٔ قابل تفکیک از هم، انتخاب در نسل‌های بعدی با پیشروی به‌اندازهٔ ۵۰ نسل انجام گرفت. بنابراین افراد بنیان‌گذار برای این سه جامعه از میان افراد نسل ۲۶۰ انتخاب شدند؛ به‌گونه‌ای که در جامعهٔ اول ۱۲ نر و ۱۲۰ ماده با ۲ نتاج برای هر مادر و با این فرض که سهم نتاج نر و ماده برابر است، استفاده شد. جایگزینی حیوانات در هر نسل به‌گونه‌ای برنامه‌ریزی شد که ۴۰ درصد از پدرها و ۲۰ درصد از مادرها در هر نسل جایگزین شوند. طرحی که برای انتخاب افراد در نظر گرفته شد، به‌گونه‌ای بود که در جامعهٔ اول افرادی که فنوتیپ بیشتری ارائه می‌دهند، انتخاب شوند و حذف افراد با توجه به سنشان رخ دهد.

جامعهٔ دوم از ۹ نر و ۹۹ ماده تشکیل شده بود و تفاوت آن با جامعهٔ اول فقط در طرحی بود که افراد با آن انتخاب می‌شدند. در این جامعه طراحی به‌گونه‌ای بود که انتخاب بر اساس فنوتیپ باشد؛ آن‌هم به‌گونه‌ای که افرادی که فنوتیپ کمتر و پایین‌تری ارائه می‌دادند، انتخاب می‌شدند.

در جامعهٔ سوم ۲۰ نر و ۱۶۰ ماده با ۲ نتاج برای هر مادر و با این فرض که سهم نتاج نر و ماده برابر است، لحاظ شد. جایگزینی حیوانات در هر نسل به‌گونه‌ای برنامه‌ریزی شد که ۴۵ درصد از پدرها و ۲۰ درصد از مادرها در هر نسل جایگزین شوند. برخلاف دو جامعهٔ قبلی که سن حیوان معیار حذف بود، در اینجا حیواناتی که فنوتیپ کمتری داشتند، در هر نسل حذف شدند. طراحی آمیزش‌ها و انتخاب به‌گونه‌ای بود که خرده‌جوامعی در درون هر جمعیت پدیدار شد. تفاوت در تعداد حیوانات ایجادکنندهٔ نسل تاریخی، در جوامع مختلف صرفاً به علت تشابه با ساختار طبیعت بوده است؛چراکه در طبیعت هم همهٔ جوامع دارای اندازهٔ یکسان یا تعداد برابری از دو جنس نیستند.

درنهایت برای شبیه‌سازی ژنوم، ژنوم برای ۲۹ کروموزوم اتوزومی به‌گونه‌ای طراحی شد که طول کروموزوم را ۱۰۰ سانتی مورگان و تعداد جایگاه‌های نشانگری را ۱۰۰۰ نشانگر برای هر کروموزوم در نظر

1. Historical generations

گرفتیم. موقعیت نشانگرها بر روی کروموزوم تصادفی در نظر گرفته شد. برای اولین نسل تاریخی شبیه‌سازی‌شده، این‌گونه تعریف شد که تعداد آلل‌های نشانگر در این نسل ۲ عدد باشد، یعنی همهٔ جایگاه‌های نشانگری باید در شروع تعداد برابری آلل (یعنی ۲ عدد) داشته باشند که البته این تعداد می‌تواند با جهش در نسل‌های بعدی افزایش یابد. در ضمن فراوانی آللی نشانگرها در اولین نسل تاریخی برابر در نظر گرفته شد که این تعداد در نسل‌های بعدی با جهش و رانش تغییر می‌کند. نرخ جهش نشانگر $۲/۵ \times ۱۰^{-۵}$ در نظر گرفته شد (Levy & Feingold, 2000; Meuwissen *et al*., 2009; Seo *et al*., 2002).

کنترل کیفیت و انتخاب نشانگرها

۲۹۰۰۰ نشانگر اتوزومی برای هر حیوان طی شبیه‌سازی ایجاد شد. نشانگرها برای از بین رفتن خطای تعیین ژنوتیپ، با نرم‌افزار PLINK (Purcell *et al*., 2007) ویرایش شدند. نشانگرهایی با نرخ فراخوانی[1] کمتر از ۹۵ درصد، حداقل فراوانی آللی[2] کمتر از ۱ درصد حذف شدند. در انتها ۲۸۲۰۹ نشانگر اتوزومی برای تجزیه‌وتحلیل‌های نهایی در نظر گرفته شدند. جهت تعیین ساختار جمعیتی از سه جامعهٔ ایجادی، ۱۵۹ دام (۵۸ دام از جامعهٔ اول، ۵۰ دام از جامعهٔ دوم و ۵۱ دام از جامعهٔ سوم) به‌طور تصادفی انتخاب و بررسی شدند.

اندازه‌گیری فاصلهٔ ژنتیکی

برآورد فاصلهٔ ژنتیکی سبب به‌دست‌آمدن میزان رابطهٔ ژنتیکی بین دو فرد می‌شود. کمّیت فاصلهٔ ژنتیکی میان دو فرد در محدودهٔ صفر و ۱ است. بدین منظور از سهم اشتراک‌گذاری آللی بین دو فرد که از مرسوم‌ترین معیارها برای برآورد فاصلهٔ ژنتیکی است، استفاده شد. از ماتریس فاصله‌ها که ماتریسی با ابعاد n در n و متقارن است و در این تحقیق همان فواصل ژنتیکی بین همهٔ حیوانات موجود در فایل تعیین ژنوتیپ است، به‌عنوان دادهٔ ورودی برای تجزیه‌وتحلیل توسط SPC استفاده شد. این فواصل در نرم‌افزار PLINK و توسط بررسی همسانی وجودی در جایگاه (IBS) با کسر عدد ۱ محاسبه شد.

خوشه‌بندی Super Paramagnetic

در خوشه‌بندی Super Paramagnetic از رویکردهای فیزیک آماری [3] برای خوشه‌بندی استفاده می‌شود. این روش که توسط Blatt و همکاران (Blatt *et al*., 1996a) پیشنهاد شد، روشی نوین برای خوشه‌بندی سلسله‌مراتبی است که الهام‌بخش آن رفتار فیزیکی دانه‌های فرو مغناطیس ناهمگن است. این الگوریتم به گروه الگوریتم‌های ناپارامتری تعلق دارد و مبتنی بر یک تابع هزینه است که در آن هیچ ساختاری برای توزیع مرتبط داده‌ها فرض نمی‌شود. خوشه‌بندی داده‌ها با حل مسئلهٔ فیزیکی مدل فرو مغناطیسی Potts به انجام می‌رسد (Blatt *et al*., 1997). یک متغیر Potts spin به هر الگو در فضای D بعدی محدود اختصاص می‌یابد. آثار متقابل با محدودهٔ کوچک بین spinهای مجاور که قدرتشان تابعی کاهشی از فاصله است، معرفی می‌شود. دو کمیّت ترمودینامیک یعنی حساسیت[4] و تابع همبستگی spin-spin به‌وسیلهٔ شبیه‌سازی مونت‌کارلو محاسبه گردیده و برای معین‌کردن تغییر فاز و جهت تقسیم‌بندی‌کردن الگوها در خوشه، استفاده می‌شوند. در مطالعهٔ حاضر جهت استفاده از خوشه‌بندی شبکه‌ای بدون نظارت در آنالیز ساختار جمعیتی به کمک نشانگرهای متراکم SNP و استنتاج ساختار جمعیتی بدون پیش‌آگاهی از انساب افراد که خود چالشی بزرگ پیش روی تجزیه‌وتحلیل‌های ژنتیکی است، استفاده شد. برای این منظور از بستهٔ نرم‌افزاری SORTING POINTS INTO NEIGHBORHOOD (SPIN) (Blatt *et al*., 1996b; Tsafrir *et al*., 2005) استفاده شد که در آن از مدل Potts Hamiltonian برای معین‌کردن ساختارهای جمعیتی در شبکه‌ها استفاده می‌شود (Neuditschko *et al*., 2012). این

1. Call rate
2. Minor Allele Frequency (MAF)
3. Statistical physics approach
4. Susceptibility

نرم‌افزار با استفاده از ماتریس فواصل، عناصر را به شمایل ذاتی و طبیعی نهفته در این فواصل، مرتب و ساختار بنیادی داده‌های چندین بعدی را آشکار می‌کند. روابط میان عناصر را می‌توان از ماتریسی که توسط این نرم‌افزار دوباره تنظیم شده است، استنتاج کرد. ازآنجایی‌که این ابزار دارای توانایی تجزیه‌وتحلیل بدون نظارت است[1]، نیازمند هیچ اطلاعاتی به‌عنوان معرف خارجی نیست و خود قادر به کشف خصوصیات ذاتی داده‌هاست. از مزایای این نرم‌افزار می‌توان به سادگی کار با آن، سرعت محاسبات، نداشتن نتیجهٔ اشتباه مثبت و قابلیت همکاری با سایر الگوریتم‌ها، به‌گونه‌ای که می‌تواند با سایر فنون بدون نظارت همکاری کند، اشاره کرد. از این نرم‌افزار می‌توان برای منظم کردن در بین و درون خوشه‌های از پیش تعیین‌شده که توسط بیشتر الگوریتم‌های خوشه‌بندی استاندارد ایجاد شده‌اند، استفاده کرد (Tsafrir *et al.*, 2005).

روش کار بر دو مرحله استوار است. مرحلهٔ نخست، فرایند مرتب‌سازی است و سپس عمل خوشه‌بندی انجام می‌گیرد. در نخستین مرحله با استفاده از دو الگوریتم STS[2] یا Neighborhood دوباره مرتب‌سازی ماتریس فواصل از قبل انجام می‌گیرد و عناصری که با فاصلهٔ بیشتری از هم قرار دارند، در گوشهٔ سمت راست بالایی و همچنین سمت چپ پایینی قرار می‌گیرند (Tsafrir *et al.*, 2005)؛ بنابراین عناصری که در ردیف خطی از هم دور هستند، در فضای چندبعدی نیز از هم دور خواهند بود. می‌توان این امر را با تعداد تکرار به‌خصوصی به انجام رسانید. تصویری که در نهایت ارائه می‌شود، بردار امتیازی است که برای آخرین چرخهٔ تکرار محاسبه‌شده و به نمایش در می‌آید. تفاوت این دو الگوریتم در نحوهٔ برخورد با عناصر دور و نزدیک است؛ به‌گونه‌ای که الگوریتم STS شیوهٔ مواجهه را در مرتب‌کردن عناصر دور و اطمینان از فاصلهٔ آن‌ها در فضای چندبعدی قرار داده و الگوریتم Neighborhood تضمین می‌کند که عناصری که در نزدیک قطر اصلی قرار دارند، کمترین عدم‌تجانس را داشته باشند. در این مطالعه از الگوریتم STS استفاده شد.

زمانی که D یا همان ماتریس ورودی، یک ماتریس فواصل باشد، همگرایی STS در نقطه‌ای ثابت پس از اجرای تعداد محدودی مرحله، تضمین‌شده است (Tsafrir *et al.*, 2005)؛ بنابراین هر تکرار STS، تابع هزینه (کمینه‌ساز) را کاهش داده و همگرایی به یک حداقل موضعی را تضمین می‌کند. در مسائل بهینه‌سازی ریاضی و تئوری تصمیم‌گیری، تابع هزینه تابعی است که یک واقعه یا مقادیر یک یا چند متغیر را روی یک عدد حقیقی که به‌طور مستقیم و ذاتی نمایندهٔ برخی «هزینه»های مرتبط با رویداد است، ترسیم و نقشه‌یابی می‌کند. در مسائل بهینه‌سازی به دنبال کمینه‌کردن تابع هزینه هستیم. در مقابل تابع هزینه، تابع هدف یا تابع پاداش[3] قرار دارد که هدف بیشینه‌کردن آن است.

در مرحلهٔ خوشه‌بندی الگوریتم SPC بر ماتریس داده‌هایی که دوباره مرتب شده‌اند، اعمال می‌شود. در این قسمت می‌توان مقادیر K یا همان تعداد نزدیک‌ترین همسایگان (KNN)، حداقل تعداد خوشه‌ها و پارامترهایی در مورد نحوهٔ داده‌های ورودی را لحاظ کرد و تجزیه‌وتحلیل نهایی را به انجام رساند.

نتایج و بحث

با استفاده از ماتریس فواصل ژنتیکی، هر نقطه از داده‌ها با یک متغیر Pott spin به‌نام s مرتبط می‌شود که به‌صورت تصادفی مقادیر صحیح ۱ تا q را می‌گیرد. نتایج خوشه‌بندی به گزینش q حساس نیست (Blatt *et al.*, 1996b)؛ بنابراین با مقدار متعارف ۲۰ (Blatt *et al.*, 1996b; Getz *et al.*, 2000b; Tetko *et al.*, 2005) که پیش‌فرض نرم‌افزار است، محاسبات انجام گرفت. زمانی که Pott spinها مرتبط شدند، شبکهٔ اولیه‌ای به‌وسیلهٔ اتصال جفت‌های کنار هم از طریق لبه‌ها ایجاد می‌شود که اثرات متقابل را به K عدد مجاور (K nearest neighbours (KNN)) محدود می‌کند. ازآنجاکه عملکرد خوشه‌بندی بر یک نمودار متصل (Blatt *et al.*, 1997) مبتنی است که همهٔ

1. Unsupervised analysis tool
2. Side2Side
3. Reward function

داده‌های موجود را به یکدیگر متصل و مرتبط می‌کند، خطوط با توجه به یک درخت پوشای[1] حداقلی مرتبط با داده‌ها اعمال می‌شوند. نتیجهٔ نهایی، گراف SPCای وابسته به خوشه‌هاست.

در نتیجه، عملکرد خوشه‌ای SPC به‌شدت به تعداد KNN وابسته است، به‌طوری‌که تعداد خوشه‌ها با کاهش KNN، افزایش می‌یابد. تعداد KNN در مطالعات مختلف متفاوت گزارش شده است، ولی با توجه به منابع، تعداد KNN ۱۰ در بیشتر مجموعه داده‌ها به‌خوبی عمل می‌کند (Blatt *et al*., 1996b; Blatt *et al*., 1997). خوشه‌بندی‌ها با اختصاص مقادیر مختلف KNN که شروع آن با KNN=۷۵ است، ارزیابی شدند. علت تعیین این عدد، قرارگیری همهٔ حیوانات در یک خوشه است. در مراحل بعدی این عدد در گام‌های پنج‌تایی تا رسیدن به KNN=۱۵ کاهش یافت و پس از آن با گام‌های کوچک‌تر یعنی یک‌به‌یک تا KNN=۳ کار ادامه یافت. با کاهش KNN همان‌گونه که انتظار می‌رفت، تعداد خوشه‌ها افزایش یافت و حیوانات بسته به سطوح متفاوت KNN خوشه‌بندی شدند؛ به‌گونه‌ای که در KNN=۷۵ تا ۶۰، حیوانات در دو خوشه دسته‌بندی شدند، ولی در KNN=۵۵ تا ۲۵ حیوانات در سه خوشه قرارگرفتند (شکل ۱-الف) و به همین ترتیب با کاهش در مقادیر KNN، دسته‌های بیشتری متمایز شدند (شکل ۱). برای تعیین تعداد بهینهٔ KNN پارامتری به‌نام Modularity به‌عنوان معیاری برای اندازه‌گیری کیفیت شبکه‌ها در فرایند پس از ارزیابی، معرفی شد (Newman & Girvan, 2004). قابلیت‌ها و جنبه‌های کلیدی بسیاری از شبکه‌های زیستی را می‌توان سازمان‌دهی آن‌ها به زیرمجموعه‌ای از عناصر (ماژول‌ها) دانست که در مقایسه با سایر عناصر موجود در شبکه، بیشتر به هم مرتبط هستند. به این امر مدولاریته در شبکه گفته می‌شود؛ بنابراین باید ماژول‌ها را شناسایی کرد و قطعاتی را در شبکه یافت که مدولاریته را حداکثر می‌کنند و برای این کار باید از روش‌های بهینه‌سازی استفاده کرد. برای این منظور از نرم‌افزار MODULAR (Marquitti *et al*., 2014) استفاده شد. محدودهٔ Modularity از ۰ تا ۱ است که ۱ نشانگر تقسیمات اجتماعی بهتر است (Holmström *et al*., 2009).

این روش با موفقیت در مسائلی متفاوت ازجمله داده‌های مخابره‌شدهٔ ماهواره‌ای و پروفایل بیان ژن مخمر، استفاده و آزمون شده است. از SPC در خوشه‌بندی توالی‌های پروتئین به‌دست‌آمده از پایگاه‌های دادهٔ SwissProt و SCOP نیز استفاده شد (Tetko *et al*., 2005). از الگوریتم SPC برای آنالیز داده‌های سرطان سینه، سرطان کولون (رودهٔ بزرگ) و سرطان خون استفاده شده است (Domany, 2003; Getz *et al*., 2003; Getz *et al*., 2000a; Getz *et al*., 2000b).

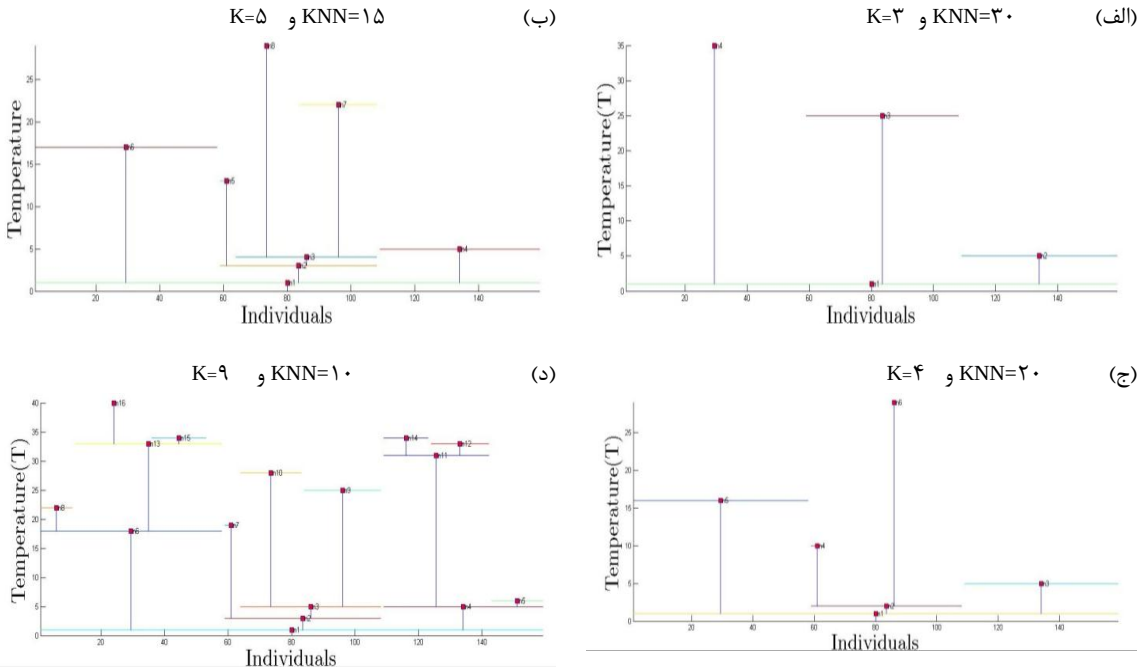
نرم‌افزار SPIN از یک تابع هزینه به‌نام Hamiltonian inhomogeneous ferromagnetic Potts برای ارزیابی عملکرد خوشه‌بندی توسط SPC در فضای چندبعدی استفاده می‌کند.

$$H(S) = -\sum_{ij} J_{ij}\,\delta_{si,sj}$$

در این تابع، طبقه‌بندی (S) با تابعی به‌نام همبستگی Spin-Spin (Gij) و ثابت اتصال J_{ij} که تابعی مثبت و کاهشی از فواصل است، معین می‌شود. مدل‌های ferromagnetic Potts در یک توالی از دماها (T) با داشتن ΔT ثابت، شبیه‌سازی می‌کنند؛ بنابراین خوشه‌بندی می‌تواند در هر سطحی از T بیان شود. به‌طورکلی در این تابع برای مقادیر کوچک T، همهٔ داده‌ها عدم همبستگی دارند. با افزایش دما همبستگی Spin-Spin بین نقاط هم‌جوار (همسایه) افزایش‌یافته و داده‌ها در امتداد دماها ($0 \le T \le T_{MAX}$) با استفاده از تقریب میانگین، خوشه‌بندی می‌شوند. در نهایت سطح خوشه‌های ایجادی بر اساسT یا دما، به‌وسیلهٔ معین‌کردن مقدار حدی Spinهای ij با G_{ij} (T) حداکثری برای هر نقطه از داده، معین می‌شود؛ بنابراین محدودهٔ دمایی $\Delta T = T_2 - T_1$ که یک خوشه از خوشه‌های بالادستی (والدین) جدا می‌شود، معیاری برای پایداری و خوشه‌بندی داده‌های متناظر با آن است. هرچه خوشه پایدارتر باشد، محدودهٔ ΔT بزرگ‌تر است (شکل ۱).

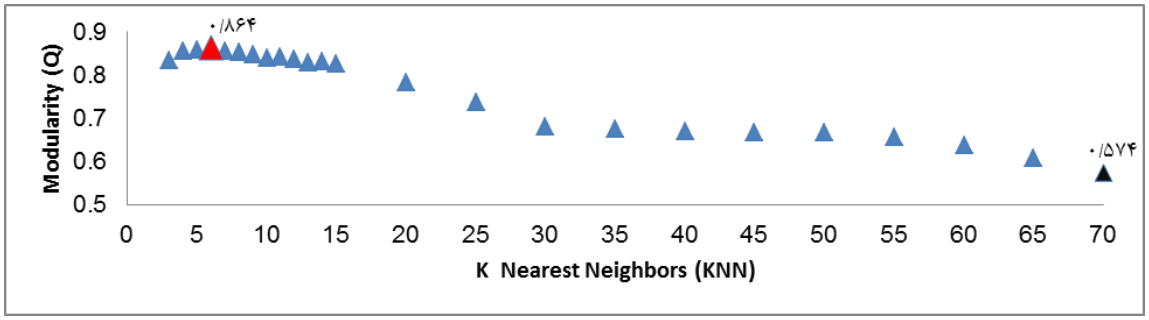
1. Spanning tree

برای به تصویرکشیدن نهایی گراف SPC و نتایج خوشه‌بندی متناظر با آن ابتدا با استفاده از ابزار تحلیلگر شبکه (NeAT) (Brohée *et al.*, 2008) داده‌ها به قالب مختصات تبدیل شدند و سپس از نرم‌افزار Cytoscape استفاده شد (Shannon *et al.*, 2003). در گزارش شبکهٔ نهایی، گره‌ها نمایانگر افراد و خطوط اتصالی بین دو گره نشانگر رابطهٔ بین حیوانات مرتبط است.

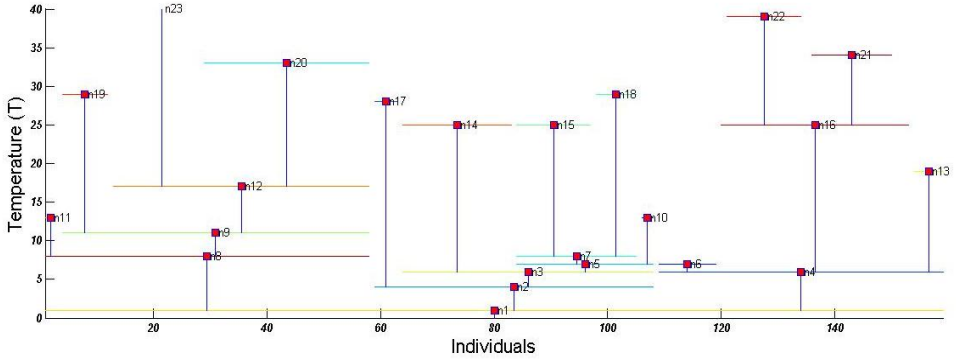


شکل ۱. تعدادی از پاسخ‌های خوشه‌بندی به‌وسیلهٔ SPC برای تعداد متفاوت KNN؛ خوشه‌ها با اعداد متفاوت از یکدیگر مشخص شده‌اند که با توجه به دمای مشخص‌شده، خوشه‌ها ابتدا به تقسیمات بزرگ‌تر و به مرور به تقسیمات خردتر تفکیک می‌شوند. هر عدد بر روی تقسیمات (nها) نشان‌دهندهٔ گروه ایجادی است. اعداد کوچک‌تر برای n، گروه‌های بزرگ ابتدایی و اعداد بزرگتر دسته‌بندی‌های خردتر است. با کاهش KNN، تعداد خوشه‌ها (K) افزایش می‌یابد.

شکل ۲ بیانگر مقدار بهینه KNN به‌وسیلهٔ ترسیم مقادیر Modularity برای محدودهٔ KNN مورد بررسی یعنی KNN=۳ تا KNN=۷۵ است. مشاهده می‌شود که در KNN=۶ یک حداکثر محلی با Q=۰/۸۶۴ به دست آمد (Marquitti *et al.*, 2014). در شکل ۳، خوشه‌بندی درختی بر اساس سلسله‌مراتب به‌دست‌آمده از این تجزیه‌وتحلیل ارائه شده است. همان‌طور که مشاهده می‌کنید، درخت با تقسیم‌شدن به سه گروه آغاز شده و با پیش‌رفتن، حیوانات، بسته به شیب دمایی ارائه‌شده، در خرده‌جوامع مرتبط با خود جای می‌گیرند.

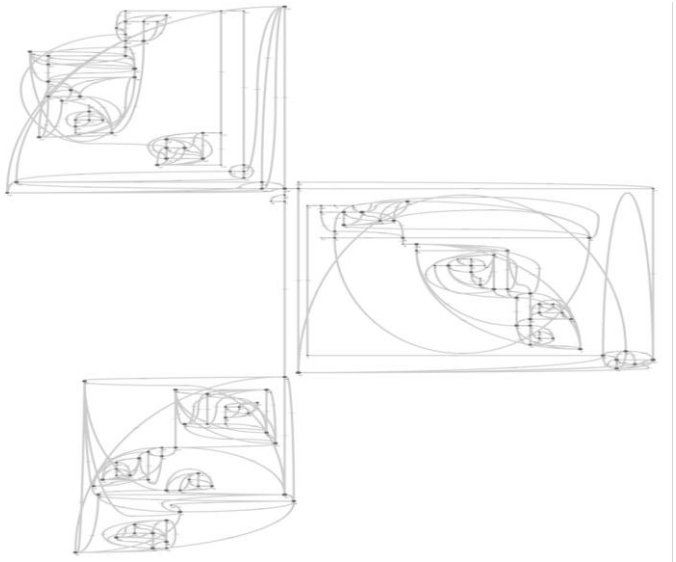


شکل ۲. تعداد بهینهٔ KNN. برای هر پیمایش SPC مقدار Modularity محاسبه شد. این مقدار به‌شدت با تعداد خوشه‌ها برای تعداد KNN مقررشده مرتبط است. مشاهده می‌شود که در KNN=۶ یک حداکثر محلی با Q=۰/۸۶۴ به دست آمد.



شکل ۳. نمودار تأییدشدهٔ خوشه‌بندی به‌وسیلهٔ SPC برای جوامع شبیه‌سازی‌شده موردبررسی. این نمودار نمایانگر سطح بهینهٔ خوشه‌بندی با KNN=۶ است که این سطح بهینه با آزمون مدولاریته معین شده است. حیوانات در ۱۳ خوشه گروه‌بندی شدند که حاصل از ۳ جامعهٔ پایه واصلی هستند که خود به خرده‌جمعیت‌هایی تقسیم شده‌اند. فاصلهٔ افقی نشانهٔ سطح هم‌جواری بین خوشه‌هاست و دما یا همان محور عمودی نشان‌دهندهٔ پایداری هر خوشه است.

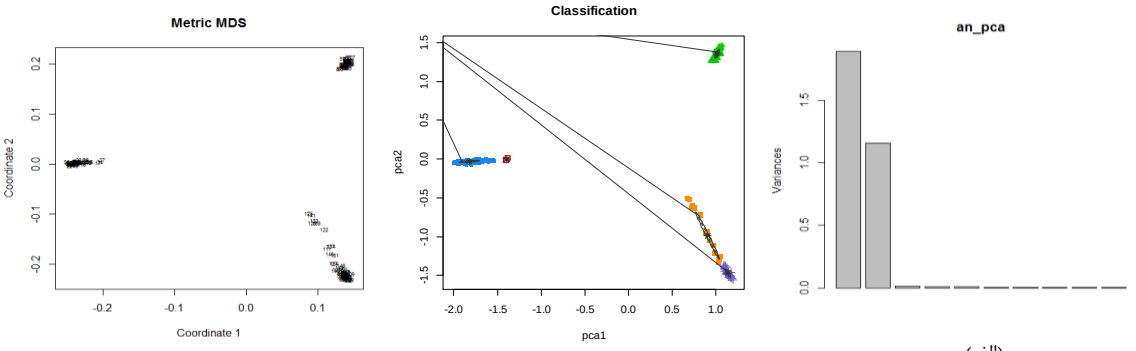
برای یافتن رابطهٔ خویشاوندی دقیق‌تر بین افراد، نمودار استخراجی SPC برای پاسخ‌های نهایی خوشه‌بندی به تصویر کشیده شد (شکل ۴). ساختار جمعیتی ارائه‌شده در شکل ۴، جواب‌های خوشه‌بندی شکل ۳ را منعکس می‌کند. در این شکل به‌وضوح جدایی سه جامعهٔ اصلی مشاهده می‌شود و شاهد تشکیل خرده‌جوامع در هر جامعه هستیم. این تصویر خود بیانگر تفسیر راحت این رویکرد در مقایسه با استفاده از سایر روش‌ها مانند مقیاس‌گذاری چندبعدی[1] (شکل ۵) یا کاربرد رویکرد مبتنی بر مدل با استفاده از نرم‌افزار STRUCTURE (شکل ۶) است که به منظور مقایسه روی همین مجموعه داده اعمال شدند. همان‌گونه که مشاهده می‌شود، تفسیر نتایج در روش‌های سنتی با توجه به فضای موجود، تصویری کامل و دقیق نیست و حتی اگر با داده‌های واقعی از جوامع دامی سروکار داشته باشیم، به‌سبب اختلاط وسیع جمعیتی و عدم وضوح جمعیت‌ها و خرده‌جمعیت‌ها، تفسیر نمودارهای حاصل از بررسی PCA بسیار پیچیده می‌شود.



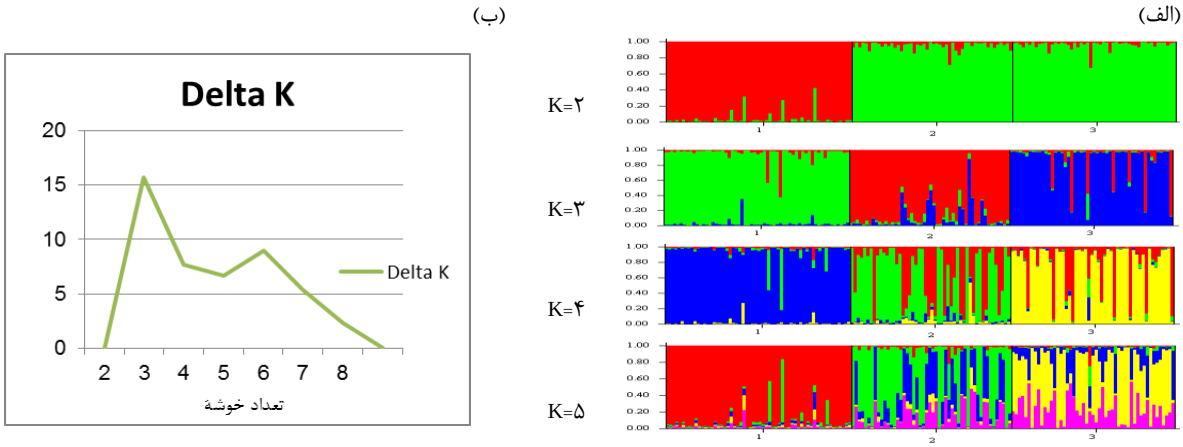
شکل ۴. شبکهٔ ارتباطی بین جوامع و خرده‌جوامع ایجادی از طریق روش SPC. هر حیوان با یک گره مشخص شده است و حیوانات قرار گرفته در هر جامعه به‌خوبی از هم مجزا شده‌اند و خرده‌جوامع هم در درون جامعه‌ها به‌صورت تودهٔ متراکم معین شده‌اند. سه جمعیت ایجادی به‌وضوح از هم جدا شده‌اند و خرده‌جوامع درون هر جمعیت نمایان شده است.

1. Multi Dimensional Scaling (MDS)

از طرفی برخی محققان (Pritchard *et al.*, 2000) معتقدند روش PCA به‌سبب اینکه در آن از برخی اطلاعات ژنتیکی چشم‌پوشی می‌شود (همانند فراوانی آللی)، برای کشف ساختار جمعیتی مناسب نیست. نتایج رویکرد مبتنی بر مدل هم نشان داد که کشف خرده‌ساختارهای جمعیتی در این روش با محدودیت مواجه است و فقط ساختارهای عمدهٔ جمعیتی را نمایان می‌سازد. STRUCTURE که از رویکردهای استاندارد آماری برای برآورد پارامترهای جمعیتی استفاده می‌کند، به‌شدت به فرض‌های مدل وابسته است و زمانی که اندازهٔ نمونه محدود و کوچک باشد، کارایی مناسبی ندارد؛ چراکه فراوانی‌های آللی را بدون دقت برآورد می‌کند. از طرفی به دلیل پیچیدگی محاسباتی، سرعت کمی دارد و محاسبات با تعداد زیاد نشانگر، قابل اجرا نیست. مزیت اصلی روش SPC که در این بررسی به آن اشاره شد، کارایی محاسباتی بالا و کمترین نیاز به فرض‌های پیشین در آن است و بنابراین به محقق این امکان را می‌دهد که اطلاعات هزاران حیوان را بدون داشتن هرگونه اطلاعاتی از شجره و نژاد، برای بررسی ساختار جمعیتی تجزیه‌وتحلیل کند.



شکل ۵. الف) نمودار مقدار واریانس توجیه‌شده نمایانگر تأثیر ۶۰ درصدی مؤلفهٔ اول و تأثیر ۳۷ درصدی مؤلفهٔ دوم است. ب) تجزیه‌وتحلیل مؤلفه‌های اصلی بر روی ماتریس IBS به انجام رسید و برای مشخص‌کردن تعداد بهینهٔ خوشه برای داده‌های موجود از بستهٔ mclust در نرم‌افزار R استفاده شد و ۵ خوشه درنهایت مشخص گردید. ج) اجرای مقیاس‌گذاری چندبعدی به‌عنوان روشی برای تعیین طبقه‌بندی جوامع.



شکل ۶. الف) نمودارهای مربوط به برآوردهای مقادیر Q توسط نرم‌افزار STRUCTURE. هر فرد با یک خط عمودی که به K قطعهٔ رنگی تقسیم شده، نشان داده شده است که طول آن نشانگر سهم هرکدام از K خوشهٔ استنتاجی است. شماره‌های ۱ تا ۳ نشانگر جمعیت‌های از پیش تعیین‌شده هستند. ب) مقادیر ΔK محاسبه‌شده با استفاده از روش Evanno و همکاران (Evanno *et al.*, 2005). توزیع مقادیر، نشان‌دهندهٔ بهترین تعداد خوشه با توجه به ساختار داده‌هاست که برای این کار تعداد ۳ خوشه مناسب شناخته شد.

سپاسگزاری

از پروفسور Eytan Domany و دکتر Assif Yitzhaky از مؤسسهٔ علوم وایزمن و دکتر مهدی ساعتچی از دانشگاه آیوا برای در اختیار گذاشتن و هماهنگی‌های لازم برای استفاده از نرم‌افزار SPIN و دکتر Mehar S. Khatkar از دانشگاه سیدنی به دلیل همکاری در استفاده از نرم‌افزار، تشکر و قدردانی می‌گردد.

REFERENCES

1. Alexander, D. H., Novembre, J. & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9), 1655-1664.
2. Behzadi, B., Amini, A., Slaminejad, A. & Tahmoorespour, M. (2013). Estimation of genetic parameters for production traits of Iranian Holstein dairy cattle. *Livestock Research Rural Development*, 25(9).
3. Blatt, M., Wiseman, S. & Domany, E. (1996a). Clustering data through an analogy to the Potts model. *Advances in Neural Information Processing Systems*, 416-422.
4. Blatt, M., Wiseman, S. & Domany, E. (1996b). Superparamagnetic clustering of data. *Physical review letters*, 76(18), 3251.
5. Blatt, M., Wiseman, S. & Domany, E. (1997). Data clustering using a model granular magnet. *Neural Computation*, 9(8), 1805-1842.
6. Bowden, R., MacFie, T. S., Myers, S., Hellenthal, G., Nerrienet, E., Bontrop, R. E., Freeman, C., Donnelly, P. & Mundy, N. I. (2012). Genomic tools for evolution and conservation in the chimpanzee: Pan troglodytes ellioti is a genetically distinct population. *PLoS genetics*, 8(3), e1002504.
7. Brohée, S., Faust, K., Lima-Mendez, G., Sand, O., Vanderstocken, G., Deville, Y. & van Helden, J. (2008). NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic acids research*, 36(suppl 2), W444-W451.
8. Decker, J. E., Pires, J. C., Conant, G. C., McKay, S. D., Heaton, M. P., Chen, K., Cooper, A., Vilkki, J., Seabury, C. M. & Caetano, A. R. (2009). Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proceedings of the National Academy of Sciences*, 106(44), 18644-18649.
9. Domany, E. (2003). Cluster analysis of gene expression data. *Journal of Statistical Physics*, 110(3-6), 1117-1139.
10. Evanno, G., Regnaut, S. & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology*, 14(8), 2611-2620.
11. Faraji-Arough, H., Aslaminejad, A., Tahmoorespur, M., Rokouei, M. & Shariati, M. (2015). Bayesian Inference of (Co) Variance Components and Genetic Parameters for Economic Traits in Iranian Holsteins via Gibbs Sampling. *Iranian Journal of Applied Animal Science*, 5(1), 51-60.
12. Fernández, M. E., Goszczynski, D. E., Lirón, J. P., Villegas-Castagnasso, E. E., Carino, M. H., Ripoli, M. V., Rogberg-Muñoz, A., Posik, D. M., Peral-García, P. & Giovambattista, G. (2013). Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd. *Genetics and Molecular Biology*, 36(2), 185-191.
13. Gao, X. & Starmer, J. (2007). Human population structure detection via multilocus genotype clustering. *BMC genetics*, 8(1), 34.
14. Gautier, M., Laloë, D. & Moazami-Goudarzi, K. (2010). Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. *PloS one*, 5(9), e13038.
15. Getz, G., Gal, H., Kela, I., Notterman, D. A. & Domany, E. (2003). Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. *Bioinformatics*, 19(9), 1079-1089.
16. Getz, G., Levine, E. & Domany, E. (2000a). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences*, 97(22), 12079-12084.
17. Getz, G., Levine, E., Domany, E. & Zhang, M. (2000b). Super-paramagnetic clustering of yeast gene expression profiles. *Physica A: Statistical Mechanics and its Applications*, 279(1), 457-464.
18. Heaton, M. P., Keen, J. E., Clawson, M. L., Harhay, G. P., Bauer, N., Shultz, C., Green, B. T., Durso, L., Chitko-McKown, C. G. & Laegreid, W. W. (2005). Use of bovine single nucleotide polymorphism markers to verify sample tracking in beef processing. *Journal of the American Veterinary Medical Association*, 226(8), 1311-1314.
19. Holmström, E., Bock, N. & Brännlund, J. (2009). Modularity density of network community divisions. *Physica D: Nonlinear Phenomena*, 238(14), 1161-1167.

20. Kijas, J. W., Townley, D., Dalrymple, B. P., Heaton, M. P., Maddox, J. F., McGrath, A., Wilson, P., Ingersoll, R. G., McCulloch, R. & McWilliam, S. (2009). A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. *PloS one*, 4(3), e4668.
21. Lee, C., Abdool, A. & Huang, C.-H. (2009). PCA-based population structure inference with generic clustering algorithms. *BMC bioinformatics*, 10(Suppl 1), S73.
22. Levy, M. & Feingold, J. (2000). Estimating prevalence in single-gene kidney diseases progressing to renal failure. *Kidney international*, 58(3), 925-943.
23. Lirón, J., Ripoli, M., De Luca, J., Peral-García, P. & Giovambattista, G. (2002). Analysis of genetic diversity and population structure in Argentine and Bolivian Creole cattle using five loci related to milk production. *Genetics and Molecular Biology*, 25(4), 413-419.
24. Markovtsova, L., Marjoram, P. & Tavaré, S. (2000). The age of a unique event polymorphism. *Genetics*, 156(1), 401-409.
25. Marquitti, F. M. D., Guimarães, P. R., Pires, M. M. & Bittencourt, L. F. (2014). MODULAR: software for the autonomous computation of modularity in large network sets. *Ecography*, 37(3), 221-224.
26. Meuwissen, T., Solberg, T. R., Shepherd, R. & Woolliams, J. A. (2009). A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet Sel Evol*, 41(2).
27. Morin, P. A., Luikart, G. & Wayne, R. K. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, 19(4), 208-216.
28. Neuditschko, M., Khatkar, M. S. & Raadsma, H. W. (2012). NetView: A High-Definition Network-Visualization Approach to Detect Fine-Scale Population Structures from Genome-Wide Patterns of Variation. *PloS one*, 7(10), e48375.
29. Newman, M. E. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 026113.
30. Ojango, J. & Pollott, G. (2001). Genetics of milk yield and fertility traits in Holstein-Friesian cattle on large-scale Kenyan farms. *Journal of animal science*, 79(7), 1742-1750.
31. Paschou, P., Ziv, E., Burchard, E. G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. W. & Drineas, P. (2007). PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS genetics*, 3(9), e160.
32. Pritchard, J. K., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945-959.
33. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. & Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3), 559-575.
34. Radjiman & Sugiarto. (2005). *Super Paramagnetic Clustering of DNA Sequences*. Master of Science dissertation. National University of Singapore. (In Farsi)
35. Sargolzaei, M. & Schenkel, F. S. (2009). QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, 25(5), 680-681.
36. Seo, T.-K., Thorne, J. L., Hasegawa, M. & Kishino, H. (2002). Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. *Genetics*, 160(4), 1283-1293.
37. Serre, D., Montpetit, A., Paré, G., Engert, J. C., Yusuf, S., Keavney, B., Hudson, T. J. & Anand, S. (2008). Correction of population stratification in large multi-ethnic association studies. *PloS one*, 3(1), e1382.
38. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11), 2498-2504.
39. Tang, H., Peng, J., Wang, P. & Risch, N. J. (2005). Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol*, 28(4), 289-301.
40. Tetko, I. V., Facius, A., Ruepp, A. & Mewes, H.-W. (2005). Super paramagnetic clustering of protein sequences. *BMC bioinformatics*, 6(1), 82.
41. Troy, C.S., MacHugh, D.E., Bailey, J.F., Magee, D.A., Loftus, R.T., Cunningham, P., Chamberlain, A.T., Sykes, B.C. & Bradley, D.G. (2001). Genetic evidence for Near-Eastern origins of European cattle. *Nature*, 410(6832), 1088-1091.
42. Tsafrir, D., Tsafrir, I., Ein-Dor, L., Zuk, O., Notterman, D. A. & Domany, E. (2005). Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. *Bioinformatics*, 21(10), 2301-2308.
43. Wang, D. G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E. & Spencer, J. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366), 1077-1082.

Iranian Journal of Animal Science, Vol. 46, No. 3, Autumn 2015

Unsupervised clustering analysis of population and subpopulation structure using dense SNP markers

Javad Rahmaninia[1], Seyed Reza Miraei-Ashtiani[2*] and Hossein Moradi Shahrbabak[3]
1, 2, 3. Ph. D. Student, Professor and Assistant Professor, Department of Animal Science, University College of Agriculture & Natural Resources, University of Tehran, Karaj, Iran
(Received: Feb. 28, 2015 - Accepted: Aug. 29, 2015)

ABSTRACT

High through put sequencing of single nucleotide polymorphisms (SNP) has revolutionized the fine scale analysis of the population structure in different species. Various methods have been proposed and used for the study of population structure using whole-genome marker data that each has advantages and disadvantages with respect to their characteristics. Super Paramagnetic Clustering (SPC) which is based on data mining was used in this study in order to investigate the population and sub-population structures in simulated populations. The purpose of applying this method was to achieve population structure without using any information from ancestral population. After editing the data, 29209 autosomal markers from 159 animals were analyzed. The results showed that animals are placed properly in their respective population and sub-populations based on their similarities and dissimilarities. The main advantages of this method are the computational efficiency and not requiring any prior assumptions. Therefore, it might be used to analyze the data from thousands of animals without any pedigree and ancestry information to reveal their population structure.

Keywords: data mining, population structure, single nucleotide polymorphism (SNP), super paramagnetic clustering (SPC).

* Corresponding author E-mail: ashtiani@ut.ac.ir Tel: +98 2632248082