تجزیه و تحلیل داده‌های مولکولی از دیده‌گاه بررسی تنوع ژنتیکی

سید ابوالقاسم محمدی
دانشگاه تبریز، دانشکده کشاورزی، گروه زراعت و اصلاح نباتات

چکیده

پیشرفت‌های سریع در تولید تکنیک‌های مولکولی به‌خصوص نشانگرهای DNA باب جدیدی را در مطالعات ژنتیکی و اصلاحی باز کرده است. کاربرد این نشانگرها در برنامه‌های مختلف حجم عظیمی از داده‌ها را تولید کرده است که مدیریت و تجزیه و تحلیل آنها نیازمند برنامه‌های آماری خاص است. هرچند که همگام با توسعه تکنولوژی‌های زیستی برنامه‌های کامپیوتری مناسب نیز برای تجزیه داده‌های حاصل ایجاد شده است. ولی عدم آشنایی اکثر کاربران با اساس ژنتیکی و آماری پارامترها و روشهای مورد استفاده، استفاده بهینه از این برنامه‌ها را کاهش داده است و حتی در بیشتر موارد کاربرد روشهای نامناسب در تجزیه و تحلیل داده‌های مولکولی منجر به نتایج غیرقابل اعتماد می‌شود. به‌عنوان مثال تعیین درست روابط ژنتیکی افراد و یا جمعیت‌ها در مجموعه ژرم‌پلاسمی بسیار مهم و نقطه اصلی در گروه-بندی افراد و یا جمعیت است. این امر نه تنها نیازمند استفاده از ابزارها و تکنیک‌های کارآ برای تعیین تفاوت و یا شباهت افراد است بلکه استفاده از ضرایب فاصله و یا شباهت مناسب را باتوجه به نوع تکنیک به‌کار رفته و ماهیت مواد ژنتیکی نیز لازم دارد. لازم به ذکر است که یک روش و یا ضریب کلی و الگوریتم خاص برای تعیین روابط افراد و گروه‌بندی آنها وجود ندارند و باید با درنظر گرفتن اساس ژنتیکی و آماری روشها و تطابق آنها با ماهیت ژنتیکی مواد ژنتیکی، روشی که بیشترین کارآیی را دارد انتخاب شود. هدف این مقاله بررسی روشهای آماری مورد استفاده در تجزیه داده‌های مولکولی با تأکید بر تنوع ژنتیکی و ارتباط داده‌های مولکولی و مورفولوژیکی در شناسایی نشانگرهای مثبت برای صفات مورد بررسی است. در این راستا اساس ژنتیکی و آماری ضرایب شباهت و فاصله مختلف و نیز روشهای آماری چندمتغیره مورد استفاده در تجزیه و تحلیل داده‌های مولکولی مورد معرفی خواهند شد.

مقدمه

توسعه بیولوژی مولکولی، تکنولوژی‌های زیستی و تکنیک‌های مرتبط با آنها و راه‌اندازی پروژه‌های ژنومی در موجودات مختلف با هدف توالی‌یابی کل ژنوم منجر به تولید حجم عظیمی از داده‌ها شده است. مدیریت و تجزیه و تحلیل صحیح این داده‌ها و بهره‌برداری درست از آنها در جهت کاربردهای علمی و عملی از ضروریات امروزه علوم زیستی می‌باشد. تولید و توسعه همگام برنامه‌های کامپیوتری برای مدیریت و تجزیه و تحلیل داده‌های مولکولی تا حدودی مشکل محققین را در این زمینه برطرف

کرده است، ولی استفاده بهینه از این برنامه‌ها و تفسیر صحیح نتاج حاصل علاوه بر آشنایی با اصول و قوانین ژنتیک مولکولی و اساس مولکولی تکنیک‌های مورد استفاده، نیازمند آشنایی و درک صحیح اساس روشهای آماری نیز می‌باشد.

یکی از کاربردهای اصلی تکنیک‌های مولکولی در بررسی و برآورد سطح تنوع ژنتیکی در مجموعه‌های ژرم‌پلاسمی و جمعیت‌های برای استفاده بهینه از آنها در مطالعات ژنتیکی و اصلاح نباتات است. تنوع ژنتیکی رکن اصلی بیشتر برنامه‌های اصلاحی بوده و انجام گزینش منوط به وجود تنوع ژنتیکی مطلوب از حیث صفت مورد بررسی می‌باشد. علاوه بر این، اطلاع از سطح تنوع موجود در رژم‌پلاسم‌ها و خزانه‌های ژنی برای تشخیص تکرارها در بانک‌های ژنی، غنی‌سازی ذخایر ژنتیکی از طریق اینتروگرسیون ژن‌های مطلوب و شناسایی ژنهای مناسب ضروری به نظر می‌رسد.

مطالعه تنوع ژنتیکی فرآیندی است که تفاوت یا شباهت گونه‌ها، جمعیت‌ها و یا افراد را با استفاده از روشها و مدل‌های آماری خاص براساس صفات مورفولوژیک، اطلاعات شجره‌ای یا خصوصیات مولکولی افراد بیان می‌کنند (Mohammadi and Prasanna, 2003). تغییرات ژنتیکی در جمعیت‌های گیاهی می‌تواند به‌وسیله مکانیسم‌های مختلف از قبیل جهش، نوترکیبی ژنتیکی، مهاجرت، جریان ژنی، رانده‌شدگی ژنتیکی و گزینش به‌وجود آید. منبع اصلی تغییرات ژنتیکی جدید، تغییرات ژنتیکی حاصل از تولید آلل‌های جدید یا آلل‌های تغییر شکل یافته با عمل ژنی جدید به‌نام تغییرات جدید[1] می‌باشد. فرآیندهای ژنتیکی مختلف مانند کراسینگ‌اور نامساوی، نوترکیبی درون ژنی، ترانسپوزن‌ها، میتله شدنDNA، پاراموتاسیون و تکثیر ژنی از عوامل مهم به‌وجود آورنده تغییرات جدید در مجموعه‌های ژرم‌پلاسمی هستند. بنابراین باتوجه به هدف مطالعه، ماهیت تغییرات ژنتیکی موجود، برای برآورد دقیق سطح تنوع ژنتیکی موجود در یک مجموعه و تفکیک آن به تغییرات درون و بین مجموعه‌ها، موارد زیر را باید مد نظر قرار داد :

- نوع نشانگر مورد استفاده در اندازه‌گیری تنوع ژنتیکی
- واحد ارزیابی تنوع ژنتیکی
- روشهای آماری مورد استفاده در تجزیه و تحلیل داده‌های حاصل

انتخاب روش مناسب برای تعیین تنوع و برآورد روابط بین ژنوتیپ‌ها و جمعیت‌ها

یکی از مسایلی که در سال‌های اخیر ذهن محققین را به خود مشغول کرده است، نقش، اهمیت و کارآیی داده‌های مورفولوژیکی و مولکولی در تعیین سطح تنوع ژنتیکی و برآورد روابط بین افراد، ژنوتیپ‌ها و جمعیت‌ها می باشد. در رابطه با اینکه آیا داده‌های حاصل از ارزیابی‌های مورفولوژیکی و یا

[1] - de novo generated variation

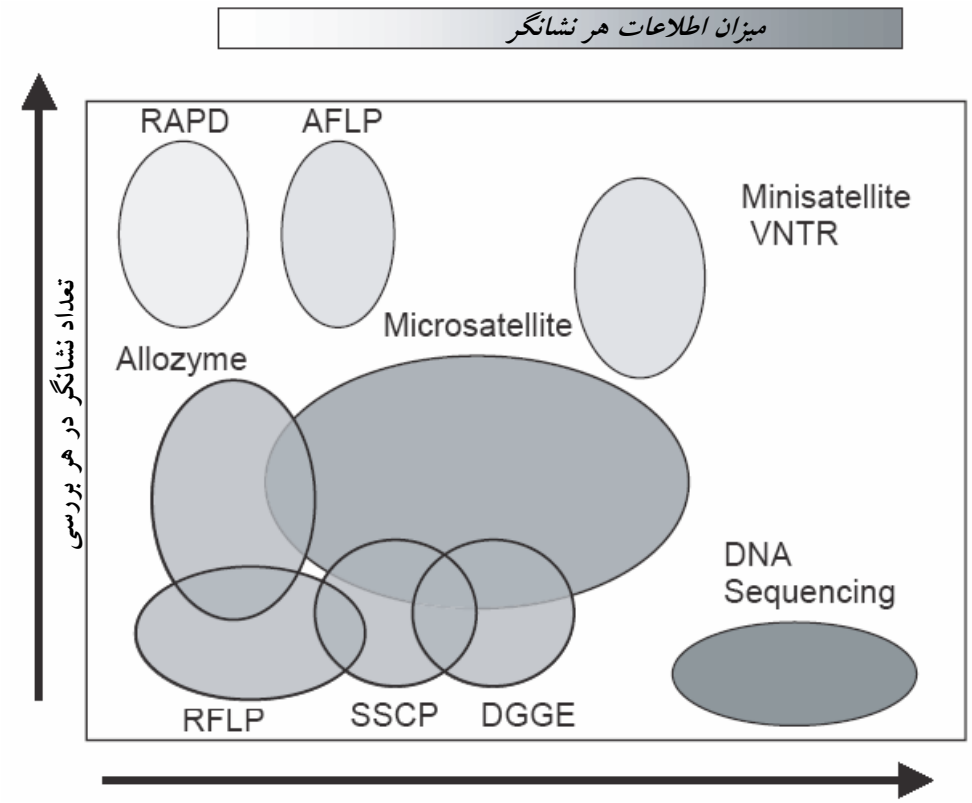
مولکولی از نظر توارثی منبع اطلاعاتی بهتر و مناسب‌تر در این زمینه می‌باشند، اختلاف نظر وجود دارد. برخی عقیده دارند که اکثر نشانگرهای مولکولی بیشتر تنوع افراد را از نظر بخش‌های غیر رمزکننده ژنوم مشخص می‌کنند. بنابراین از این نقطه نظر ممکن است نتوان ارتباط مستقیم بین تنوع مولکولی و مورفولوژی ایجاد کرد. عده‌ای دیگر معتقد هستند که تعیین تنوع فقط بر اساس اطلاعات موفولوژیکی گمراه کننده و فاقد اطلاعات کافی جهت برآورد روابط بین ژنوتیپ‌ها و یا جمعیت‌ها جهت استفاده در مطالعات کاربردی است. با این وجود باید در نظر داشت که روشهای مورفولوژیکی و مولکولی هر کدام دارای مزایا و معایب خاص خود هستند. به‌عنوان مثال، اکثر (نه همه) داده‌های مولکولی دارای اساس ژنتیکی شناخته شده بوده و توارث مندلی دارند و تعداد کل داده‌های حاصل از یک مطالعه بستگی به اندازه ژنوم و تعداد نشانگرهای مورد استفاده دارد. از طرف دیگر، داده‌های مورفولوژیکی به آسانی قابل اندازه‌گیری بوده و تغییرات مربوط به قسمت‌های رمزکننده ژنوم را همراه با اثر محیط نشان می‌دهند. باتوجه به اینکه تنوع تعیین شده توسط نشانگرهای DNA اکثراً تنوع خاموش می‌باشد و داده‌های مورفولوژیکی علاوه بر تغییرات ژنتیکی شامل تغییرات محیطی غیر قابل توارث هستند. درنتیجه استفاده همزمان از هر دو نوع داده، توصیف و تفسیر بهتر و جامع‌تری از تنوع بیولوژیکی را فراهم خواهد کرد. در سال‌های اخیر به‌دلیل پیشرفت‌های وسیع در زمینه تکنولوژی نشانگرهای *DNA*، استفاده از این نشانگرها به‌عنوان ابزارهای کاراً و مکمل روشهای مورفولوژیکی جهت بررسی و تعیین سطح تنوع ژنتیکی و روابط بین افراد، گونه و جمعیت به‌طور چشمگیری افزایش یافته است.

نشانگرهای DNA

نشانگرهای DNA تفاوت افراد را در سطح مولکول DNA نشان می‌دهند و دارای چندشکلی بالا، توزیع تصادفی در ژنوم، عدم تأثیرپذیری از عوامل محیطی، خنثی بودن از نظر فنوتیپی، فراوانی زیاد، غیر وابسته بودن به مرحله رشد، بافت و ارگان و قابل ارزیابی بودن در هر مرحله و در آزمایشگاه می‌باشند. امروزه طیف وسیعی از نشانگرهای DNA در دسترس هستند که در انتخاب آنها باید معیارهایی مانند محل ژنومی آنها (رمزکننده، غیررمزکننده)، توزیع نسبی آنها در ژنوم (در مورد نشانگرهایی با محل ژنومی معینی) درنظر گرفته شوند (Karp et al., 1997). نشانگرهایی از قبیل RFLP[1]، RAPD[2]، SSR[3]، AFLP[4] و ... به‌طور وسیعی در مطالعه تنوع ژنتیکی استفاده شده اند. در بررسی تنوع ژنتیکی معیارهای

[1] Restriction Fragment Length Polymorphism
[2] Random Amplified Polymorphic DNA
[3] Simple Sequence Repeat
[4] Amplified Fragment Length Polymorphism

اولیه برای گزینش نشانگرها عبارت از هزینه، ابزار و نیازهای تکنیکی، تعداد نشانگر در هر بررسی و میزان اطلاعات هر نشانگر است (شکل ۱).



شکل ۱ - معیارهای لازم برای انتخاب نشانگر مناسب در بررسی تنوع

علاوه بر معیارهای فوق، پارامترهایی مثل درجه چندشکلی، تکرارپذیری و درجه اعتبار نیز باید مد نظر قرار گیرد. درجه چندشکلی نشانگرهای توسط دو پارامتر درصد مکان‌های چند شکل و متوسط تعداد آلل‌ها در هر مکان ژنی تعیین می‌شود. درجه چندشکلی با استفاده از شاخص تنوع[1] (نی، ۱۹۷۳) و میزان اطلاعات چند شکل[2] *(PIC)* نشان داده می‌شود. میزان اطلاعات چند شکل از فرمول $PIC = 1 - \sum p_i^2$ برای نشانگرهای همبارز که در آن‌ها روابط اللی در هر مکان ژنی مشخص

[1] Diversity index
[2] Polymorphic information content

است و از فرمول $PIC = \sum[2p_i(1-p_i)]$ برای نشانگرهای غالب که در آنها رابطه آللی بین نوارها قابل دسترس نیست، محاسبه می‌گردد. درصورتی‌که در بین افراد مورد بررسی ژنوتیپ‌های هتروزیگوت وجود داشته باشد، در نشانگرهای همبارز، میزان اطلاعات چند شکل از فرمول $PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2 p_i^2 p_j^2$ برآورد خواهد شد. در این فرمولها p_i فراوانی آلل i ام برای نشانگرهای همبارز و فراوانی نوار i ام برای نشانگرهای غالب است. یکی از پارامترهای متداول برای اندازه‌گیری تنوع ژنتیکی، تنوع در سطح مکان ژنی منفرد است که به‌صورت $D_l = ۱ - \sum p_{li}^۲$ برآورد می‌گردد که در آن p_{li} فراوانی آلل i ام در مکان ژنی l ام می‌باشد. این پارامتر مشابه PIC در لاین-های خالص است. متوسط تنوع در مجموع چندین مکان ژنی از فرمول $D = ۱ - \frac{۱}{L}\sum_{l} D_{l} = ۱ - \frac{۱}{L}\sum_{l}\sum_{i} p_{li}^۲$ حساب می‌شود. این پارامتر مشابه هتروزیگوتی مورد انتظار در یک مکان ژنی برای موجودات دیپلوئید با تعادل هاردی-واینبرگ است. هتروزیگوتی مورد انتظار در یک مکان ژنی برای موجودات دیپلوئید به‌صورت $H_l = ۲\sum_{i \neq j} p_{li} p_{lj} = ۱ - \sum p_{li}^۲ = D_l$ قابل محاسبه می‌باشد. رابطه بین این دو پارامتر روشی را برای مقایسه سطح تنوع ژنتیکی در موجودات هاپلوئید، جایی‌که امکان تعیین دقیق سطح تنوع امکان‌پذیر است و موجودات دیپلوئید که با استفاده از آنها امکان تعیین سطح هتروزیگوتی وجود دارد، فراهم می-کند.

در مورد جمعیت‌های ناهمگن مثل توده‌های بومی یا ارقام محلی شاخص تنوع ژنی(Nei, 1983) به صورت $H_e^k = \frac{1}{L}\sum_{l}[\frac{2n_l}{2n_l - 1}(1 - \sum_{i} p_{li}^k)^2$ محاسبه می‌شود. که در آن L نشان‌دهنده تعداد کل مکان‌ها یا نشانگرهای بررسی شده در جمعیت k ام ، n_l تعداد افراد بررسی شده برای مکان یا نشانگر l ام و p_{li}^k فراوانی آلل i ام مکان l در جمعیت k ام است.

واحد ارزیابی تنوع ژنتیکی

مطالعات مختلف نشان دادند که تنوع ژنتیکی درون گونه‌ای به‌طور یکنواخت در محدوده زیستگاه-ها توزیع نشده است. هرچند که در مورد گونه‌های وحشی توزیع جغرافیایی سهم بیشتر در تنوع ژنتیکی دارد ولی در گونه‌ها زراعی، الگوهای جغرافیایی نشان‌دهنده اثر و نقش گزینش مصنوعی در مناطق خاص و تاریخچه تولید گیاهان زراعی در نواحی مختلف است. بنابراین در بررسی تنوع ژنتیکی، تعیین سطح تنوع ژنتیکی درون و بین جمعیت‌های حائز اهمیت است. در این بررسی‌ها لازم است شباهت یا

فاصله ژنتیکی افراد یا جمعیت ها تعیین شود که معیارهای متفاوتی برحسب داده‌های مولکولی برای برآورد شباهت یا فاصله ژنتیکی وجود دارد (Mohammadi & Prasanna, 2003).

معیارهای برآورد شباهت یا فاصله ژنتیکی افراد براساس داده‌های مولکولی :

الگوی نواری نشانگرهای به‌صورت وجود یا عدم وجود نوار خاص (یک و صفر) برای نشانگرهای غالب و نیز همبارز و به‌صورت فراوانی‌های آللی برای نشانگرهای همبارز امتیازدهی می‌شود. زمانی‌که هدف برآورد فاصله یا شباهت ژنتیکی بین افراد باشد. در اکثر موارد نشانگرهای همبارز نیز مثل نشانگرهای غالب به‌صورت وجود یا عدم وجود آلل خاص امتیازدهی می‌شوند (یک و صفر).

ضرایب مبتنی بر داده‌های صفر و یک: معیارهای متفاوتی برای برآورد فاصله یا شباهت ژنتیکی افراد براساس داده‌های صفر و یک وجود دارد که در جدول (۱) آورده شده‌اند. از بین معیارهای مختلف ضرایب شباهت ژاکارد (GS_J)، نی و لی یا دایس(GS_{NL}) و تطابق ساده (GS_{SM})متداول‌ترین ضرایب مورد استفاده برای داده‌های صفر و یک هستند (محمدی، ۱۳۸۱). برحسب خصوصیات آماری و ژنتیکی ضرایب، هر کدام برای شرایط خاصی مناسب‌تر می‌باشند. به‌عنوان مثال ضرایب ژاکارد و نی و لی در وزن تعداد نوارهای مشترک بین دو فرد متفاوت هستند و در ضریب نی و لی این وزن دو برابر ضریب ژاکارد است. بنابراین اگر افراد مورد بررسی خالص باشند، برآورد شباهت ژنتیکی آنها با استفاده از این دو ضریب یکسان خواهد بود. درصورتی‌که مواد ژنتیکی مورد مطالعه شامل افراد هتروزیگوت نیز باشد، برآورد این دو ضریب با استفاده داده‌های نشانگرهای هم‌بارز که قدرت تفکیک افراد هموزیگوت و هتروزیگوت را دارند متفاوت خواهد بود. مطالعه روی گیاهان مختلف نشان داده است که برآورد شباهت ژنتیکی دو فرد براساس ضریب نی و لی با استفاده نشانگرهای همبارز تابع خطی از ضریب همنسبی آنها است. درحالی‌که در استفاده از نشانگرهای غالب این رابطه برای ضریب ژاکارد صادق است (Melchinger, 1993). ضرایب "نی و لی" و ژاکارد به‌دلیل عدم تصادفی بودن آنها دارای توزیع آماری مشخص نیستند که محاسبه واریانس نمونه‌برداری و درنتیجه حدود اطمینان را در مورد این ضرایب مشکل می‌کند (Lombard et al., 2000)، هرچند که استفاده از تکنیک Bootstrap این مشکل را برطرف خواهد کرد (Tivang et al., 1994). ضریب تطابق ساده دارای توزیع آماری دوجمله‌ای است. ولی عیب عمده این ضریب وزن یکسان برای نوارهایی که در هردو فرد مشترک هستند و نوارهایی که در هر دو فرد وجود ندارند می‌باشد.

جدول ۱ - ضرایب شباهت و فاصله مورد استفاده در برآورد روابط ژنتیکی افراد با استفاده از داده‌های مولکولی امتیازدهی شده به صورت صفر و یک

ردیف	نام ضریب	فرمول	ردیف	نام ضریب	فرمول
۱	Jaccard/Tanimoto	$\frac{a}{a+b+c}$	۱۲	Ochiai/Cosine	$\frac{a}{\sqrt{(a+b)(a+c)}}$
۲	Dice	$\frac{2a}{2a+b+c}$	۱۳	Kulczynski (2)	$\frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)}$
۳	Russell/Rao	$\frac{a}{n}$	۱٤	Forbes	$\frac{n\times a}{(a+b)(a+c)}$
٤	Sokal/Sneath (1)	$\frac{a}{a+2b+2c}$	۱٥	Fossum	$\frac{n\left[a-\frac{1}{2}\right]^2}{(a+b)(a+c)}$
٥	Kulczynski (1)	$\frac{a}{b+c}$	۱٦	Simpson	$\frac{a}{\min(a+b,a+c)}$
٦	Simple Matching	$\frac{a+d}{n}$	۱۷	Pearson	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
۷	Hamann	$\frac{a+d-b-c}{n}$	۱۸	Yule	$\frac{ad-bc}{ad+bc}$
۸	Sokal/Sneath (2)	$\frac{2a+2d}{a+d+n}$	۱۹	McConnaughey	$\frac{a^2-bc}{(a+b)(a+c)}$
۹	Rogers/Tanimoto	$\frac{a+d}{b+c+n}$	۲۰	Stiles	$\log_{10}\frac{n\left[\lvert ad-bc\rvert-\frac{n}{2}\right]^2}{(a+b)(a+c)(b+d)(c+d)}$
۱۰	Sokal/Sneath (3)	$\frac{a+d}{b+c}$	۲۱	Dennis	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$
۱۱	Baroni-Urbani/Buser	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	۲۲	Mean Manhattan	$\frac{b+c}{n}$

d, c, b, a و n به ترتیب تعداد نوارهای مشترک بین دو فرد، تعداد نوارهای مختص فرد اول، تعداد نوارهای مختص فرد دوم، تعداد نوارهایی که در هیچ یک از دو فرد وجود ندارد و تعداد کل نوارها می باشند.

ضرایب ۱ تا ۱٦ و ۲۲ ضرایب شباهت و ۱۷ تا ۲۱ ضرایب فاصله هستند.

ضرایب مبتنی بر فراوانی‌های آللی: همانند داده‌های صفر و یک، در استفاده از فراوانی‌های آللی برای برآورد فاصله ژنتیکی نیز ضرایب متعددی به کار رفته است که مهمترین آنها در جدول (۲) آورده شده‌اند (Reif et al., 2005).

جدول ۲ - ضرایب فاصله مورد استفاده در برآورد روابط ژنتیکی با استفاده از فراوانی‌های آللی

ردیف	نام ضریب	فرمول	محدوده
۱	**Euclidean** اقلیدسی	$\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}$	$0,\sqrt{2m}$
۲	Rogers (1972) راجر ۷۲	$\frac{1}{m}\sum_{i=1}^{m}\sqrt{\frac{1}{2}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}$	0,1
۳	Modified Rogers (Wright, 1978; Goodman and Stuber, 1983) راجر تغییر شکل یافته	$\frac{1}{\sqrt{2m}}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2}$	0,1
٤	Cavalli-Sforza and Edwards (1967) کاوالی-اسفورزا و ادوارد	$\frac{1}{m}\sum_{i=1}^{m}\left(1-\sum_{j=1}^{n_i}\sqrt{p_{ij}q_{ij}}\right)$	0,1
٥	Reynolds et al. (1983) رینولد	$-\mathrm{Ln}(1-\sum_{i=1}^{m}(a-b)/c)$ $a=1/2\sum_{j=1}^{n_i}(p_{ij}-q_{ij})^2$ $b=[1/(2(2n-1))]\times(2-\sum_{j=1}^{n_i}(p^2{}_{ij}+q^2{}_{ij}))$ $c=\sum_{i=1}^{m}(1-\sum_{j=1}^{m}p_{ij}q_{ij})$	$0,\infty$
٦	Nei (1972) نی ۷۲	$-\ln\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}p_{ij}q_{ij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n_i}p^2{}_{ij}\sum_{i=1}^{m}\sum_{j=1}^{n_i}q^2{}_{ij}}}$	$0,\infty$
۷	Nei et al. (1983) نی ۸۳	$\frac{1}{m}\sum_{i=1}^{m}(1-\sum_{j=1}^{n_i}\sqrt{p_{ij}q_{ij}})$	0,1
۸	Shared Allele الل های مشترک	$D_{SA}=\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{a_j}\min(p_{ij},q_{ij})$	$0,\infty$

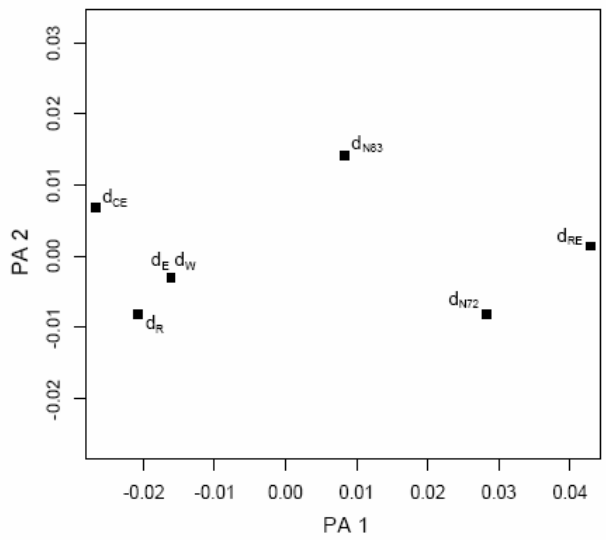
p_{ij} : فراوانی آلل i در مکان j در جمعیت X، q_{ij} : فراوانی آلل i در مکان j در جمعیت Y، a_j : تعداد آلل‌ها در مکان j، m : تعداد مکان‌های مورد آزمایش است.

ضرایب فاصله فوق‌الذکر دارای خصوصیات ژنتیکی و آماری متفاوت است که براساس آن باید مورد استفاده قرار گیرند. به‌عنوان مثال، فاصله اقلیدسی دارای جنبه ژنتیکی خاص نیست و مناسب برای

بررسی روابط افراد یا جمعیت‌ها با استفاده از روشهای آماری چندمتغیره‌ای از قبیل تجزیه بـه هماهنـگ-های اصلی، تجزیه خوشه‌ای آشیانه‌ای، طبقه‌بندی آشیانه‌ای و تئوری گراف است که فاصله‌های اقلیدسـی را لازم دارند. ضریب راجر دارای رابطه خطی با ضریب همنسبی افراد است و برای ارزیابی روابط افـراد در مجموعه‌های ژرم‌پلاسمی، برای تعیین روابط شجره‌ای ارقام مشتق شده ضروری[1] در اصلاح نباتات و نیز تشخیص ژنوتیپ‌های تکرار شده در کلکسیون‌ها و بانـک ژن‌هـا مناسـب مـی‌باشـد. ضـریب راجـر تغییریافته دارای کارآیی بالا در پیش‌بینی عملکرد هیبریدها براساس فاصله ژنتیکی والدین و تعیین گروه-های هتروتیک دارد. به‌طوری‌که رابطه خطی بین این ضریب و هتـروزیس میـانگین والـدین پانمیکتیـک گزارش شده است. با فرض مدل آللی نامحدود[2]، ضریب نی ۷۲ برای بررسـی روابـط فیلـوژنتیکی بـین جمعیت‌ها مناسب است. زمانی‌که مواد ژنتیکی مورد بررسی لاین‌های خالص باشند، ضریب نی ۸۳ برابر با ضریب راجر بوده و درنتیجه دارای رابطه خطی با ضریب همنسبی لاین‌ها خواهـد بـود (Reif et al., 2005).

ریف و همکاران (۲۰۰۵) با استفاده از داده‌های مولکولی هفت جمعیت ذرت از مرکـز بـین‌المللـی سمیت ضرایب فاصله مختلف را با اسـتفاده از روش مقیـاس‌بنـدی چنددامنـه‌ای غیرمتریـک کروسـکال مقایسه و روابط آنها را به‌صورت گرافیکی ارایه کرد (شکل ۲). در این بررسی فاصله ضرایب اقلیدسی و راجر تغییر یافته بسیار نزدیک به یکدیگر با ضریب راجر گروه‌بندی شـدند. همچنـین مـشاهده شـد کـه آنـالوگی ضـرایب نـی ۷۲ و رینولـد در بـرآورد فراوانـی‌هـای آللـی جمعیـت اجـدادی شـدیداً تحـت تأثیرخصوصیات ضریب درمقایسه با مدل تکاملی انتخابی با جهش و رانـده‌شـدگی ژنتیکـی و یـا فقـط رانده‌شدگی ژنتیکی است.

[1] Essentially derived varieties
[2] Infinite-allele model



شکل ۲ . رابطه فواصل ژنتیکی مبتنی بر فراوانی‌های اللی در هفت جمعیت ذرت با استفاده از روش مقیاس-بندی چنددامنه‌ای غیرمتریک کروسکال (d_E: فاصله اقلیدسی، d_R: ضریب راجر، d_{RE}:ضریب راجر تغییریافته، d_W: ضریب رینولد، d_{CE}: ضریب کاوالی-اسفورزا و ادوارد، d_{N72}: ضریب نی ۷۲، d_{N83}: ضریب نی ۸۳).

گروه‌بندی افراد یا جمعیت‌ها

پس از برآورد روابط ژنتیکی افراد و یا جمعیت‌ها، گام بعدی گروه‌بندی آنها براساس درجه شباهت یا تفاوت آنها است. در این راستا روشهای آماری چند متغیره از قبیل تجزیه خوشه‌ای[1]، تجزیه به مولفه-های اصلی[2]، تجزیه به هماهنگ‌های اصلی[3]، از متداول‌ترین روشهای آماری مورد استفاده در این راستا می‌باشند (محمدی، ۱۳۸۱).

تجزیه خوشه‌ای: تجزیه خوشه‌ای یکی از متداول‌ترین روشهای آماری چندمتغیره در بررسی تنوع ژنتیکی و گروه‌بندی افراد و جمعیت‌ها می‌باشد. این روش در مواردی استفاده می‌شود که الگوی گروه-بندی قبلی در مورد مواد ژنتیکی مورد مطالعه وجود نداشته باشد. الگوریتم‌های متعددی برای تجزیه خوشه‌ای پیشنهاد شده‌اند. این الگوریتم‌ها به دو گروه اصلی تقسیم می‌شوند: (۱) روشهای مبتنی بر فاصله[4] که در آنها ماتریس فاصله یا شباهت دو به دو افراد یا جمعیت‌ها برای گروه‌بندی استفاده می‌شود

[1] Cluster analysis
[2] Principal component analysis
[3] Principal coordinate analysis
[4] Distance based methods

(Johnson and Wichern, 1992) و (۲) روشهای مبتنی بر مدل[1] که در آنها فرض بر اینست که افراد درون هر خوشه مشاهدات تصادفی از برخی مدل های پارامتری بوده و استنباطات آماری مربوط به پارامترهای هر خوشه و تعیین عضویت در هر خوشه با استفاده از روشهای آماری استاندارد مانند حداکثر درست‌نمایی و ... انجام می‌گیرد (Pritchard et al., 2000). تاکنون در مطالعات ژنتیکی بیشتر روشهای مبتنی بر فاصله به‌خصوص الگوریتم‌های طبقاتی[2] استفاده شده‌اند. در استفاده از تجزیه خوشه-ای در مطالعات بیولوژیکی محققین با سوالات متعددی مواجه می‌شوند که در اینجا مهمترین آنها مطرح و راه‌حل‌های ممکن ارائه می‌شود.

- ***انتخاب معیار فاصله یا شباهت:*** این مبحث در بخش مربوط به ضرایب شباهت یا فاصله مورد بررسی قرار گرفت.

- ***انتخاب الگوریتم گروه‌بندی:*** الگوریتم‌های متفاوتی برای تجزیه خوشه‌ای وجود دارد. از بین این الگوریتم‌ها، UPGMA[3] و روش Ward متداول‌ترین روشهای تجزیه خوشه‌ای مورد استفاده در مطالعات ژنتیکی بخصوص بررسی تنوع می‌باشند. رینکون و همکاران (۱۹۹۶) کارآیی الگوریتم‌های مختلف تجزیه خوشه‌ای مانند UPGMA ، Ward ،SLINK[4] ،CLINK[5] و ... را در بررسی تنوع ژنتیکی و گروه‌بندی مواد گیاهی بررسی و نتیجه گرفتند که UPGMA نتایج معتبری را در انطباق با روابط شجره-ی مواد ژنتیکی فراهم می‌کند. ولی یکی از معایب اصلی این روش وجود اثر زنجیره‌ای[6] در دندروگرام حاصل است که تفسیر نتایج و تشخیص گروه‌ها را مشکل می‌سازد. روش Ward نتایج مشابه با UPGMA فراهم می‌آورد ولی مشکل اثر زنجیره‌ای در دندروگرام حاصل از آن به‌ندرت دیده می‌شود.

معیارهای متفاوتی برای بررسی کارآیی الگوریتم‌های مختلف در تجزیه خوشه‌ای استفاده می‌شوند که از متداول‌ترین آنها می‌توان به ضریب همبستگی کوفنتیک[7] اشاره کرد. ضریب همبستگی کوفنتیک بزرگتر یا مساوی ۰/۹ نشان‌دهنده کارآیی الگوریتم در گروه‌بندی ژنوتیپ‌ها می‌باشد. با وجود این، رینکون و همکاران (۱۹۹۶) در تجزیه خوشه‌ای با استفاده از داده‌های مولکولی نشان دادند که ضریب کوفنتیک پایین دلیل بر عدم کارآیی دندروگرام حاصل نمی‌تواند باشد، بلکه ضریب همبستگی کوفنتیک پایین ممکن است به دلیل شرایط غیرعادی در داده‌ها به‌خصوص داده‌های مولکولی گمشده باشد.

---

[1] Model based methods
[2] Hierarchical algorithms
[3] Unweighted Paired Group Method using Arithmetic average
[4] Single linkage
[5] Complete linkage
[6] Chaining effect
[7] Cophenetic correlation coefficient

- **تعداد مطلوب خوشه**: یکی از مهمترین جنبه‌های تجزیه خوشه‌ای تعیین تعداد مطلوب خوشه است. این موضوع به‌خصوص در مطالعات ژنتیکی برای گروه‌بندی ژنوتیپ‌ها و تعیین گروه‌های معین جهت استفاده در برنامه‌های اصلاح نباتات بسیار مهم و اساسی است. روشهای متعددی برای تعیین تعداد مطلوب خوشه پیشنهاد شده‌اند. ویشارت (۱۹۸۷) روش حد دامنه بالا[1] را برای این منظور پیشنهاد کرد. در این روش، میانگین و انحراف استاندارد مقادیر فاصله بین ژنوتیپ‌ها در محل ادغام خوشه‌ها به‌عنوان معیاری برای تعیین تعداد مطلوب خوشه استفاده می‌شوند. با n فرد در یک روش طبقه‌ای n-1 نقطه ادغام وجود خواهد داشت. اگر مقادیر فاصله در این نقاط d_1، d_2 و ... d_{n-1} با میانگین d و انحراف استاندارد S_d باشد، نقطه برش دندروگرام نقطه‌ای است که مقدار ضریب فاصله در آن نقطه ادغام در محدود $100\%(1-\alpha)$ توزیع فاصله‌ها باشد. به این ترتیب که در تمام نقاط ادغام در این محدوده مقادیر ضریب فاصله خارج از حد فاصل تعریف شده نباشد (Franco et al., 1997).

تجزیه واریانس مولکولی[2] (AMOVA) که در آن هر گروه در نقطه برش دندروگرام به‌عنوان یک جمعیت و ژنوتیپ‌های درون آن به‌عنوان افراد جمعیت درنظر گرفته شده و برای هر نقطه برش یک تجزیه واریانس انجام می‌گیرد. نقطه‌ای که در آن بیشترین تمایز بین گروه‌ها به‌وجود می‌آید، نقطه‌ای مناسب برای برش دندروگرام خواهد بود. علاوه بر این‌ها، تعداد مطلوب خوشه را می‌توان با استفاده از تکنیک Bootstrapping و یا تجزیه تابع تشخیص[3] نیز تعیین کرد. بدین ترتیب در تجزیه تابع تشخیص، در هر نقطه برش که بیشترین تمایز بین گروه‌های حاصل به‌وجود بیاید، آن نقطه محلی مناسب برای برش دندروگرام خواهد بود.

استفاده از داده‌های چندگانه: تنوع ژنتیکی یا روابط بین ژنوتیپ‌ها یا افراد را می‌توان براساس داده‌ها مورفولوژیکی (کمی یا کیفی) و یا داده‌های مولکولی براساس وجود یا عدم وجود باند یا فراوانی‌های آللی مطالعه کرد. بنابراین سوال اصلی اینست که در صورت وجود داده‌های چندگانه در یک مطالعه به چه صورت باید از آنها در محاسبه فاصله یا شباهت بین افراد استفاده کرد. از نقطه نظر آماری امکان ادغام داده‌های کمی و داده‌های کیفی (داده‌هایی که به‌صورت صفر و یک بیان می شوند) وجود دارد که در این رابطه فرمول‌هایی نیز برای محاسبه فاصله بین دو فرد ارایه شده‌اند. فرمول گوور (Franco et al., 1997) برای محاسبه فاصله ژنتیکی دو فرد به صورت $GD_{ij}=1/p\Sigma W_k d_{ijk}$ می‌باشد که در آن p تعداد صفت (کمی یا کیفی)، $d_{ijk}=|X_{ik}-X_{jk}|$ نشان دهنده سهم صفت k ام در فاصله ژنتیکی دو فرد، X_{ik} و X_{jk} به ترتیب ارزش صفت k ام برای فرد i و j، $W_k=1/R_k$ و R_k دامنه تغییرات صفت k ام است. در این ضریب ابتدا مقادیر صفات یا داده کمی به محدوده صفر و یک تبدیل شده و سپس در محاسبه

[1] Upper tail limit
[2] Analysis of molecular variance
[3] Discriminant analysis

فاصله افراد استفاده می‌شوند. برای تبدیل داده‌های کمی ابتدا هر داده از میانگین کل صفت کم و به دامنه تغییرات آن تقسیم می‌شود. هر چند که از نقطه نظر آماری در ادغام داده‌های مختلف برای برآورد فاصله یا شباهت ژنتیکی مشکل خاصی وجود ندارد ولی از نقطه نظر ژنتیکی ادغام داده‌هایی با ماهیت متفاوت به علت وزن غیر یکسان صفات مختلف در برآورد روابط بین افراد که در جهت عکس اهمیت بیولوژیکی آنها است، سبب نتیجه‌گیری و گروه‌بندی غیرواقعی ژنوتیپ‌ها می‌شود. یعنی صفات کمی که توسط تعداد زیادی ژن کنترل می شوند وزن کمتری درمقایسه با صفات کیفی که تحت کنترل تعداد معدودی ژن می‌باشند، در برآورد فاصله ژنتیکی افراد خوهدند داشت. بنابراین، در مطالعات تنوع ژنتیکی براساس داده‌های مختلف توصیه می‌شود که تجزیه خوشه‌ای جداگانه براساس داده‌های مختلف انجام و پس از مقایسه دندروگرام‌ها با یکدیگر و با اطلاعات شجره‌ای، نهایتاً یک دندروگرام اجمالی[1] براساس نتایج تمام داده‌ها ترسیم شود. لازم به ذکر است که در اکثر موارد ممکن است انطباق بالایی بین دندروگرام های حاصل از داده‌های مختلف و حتی روشهای مختلف روی همان داده‌ها وجود نداشته باشد.

تجزیه به مولفه‌های اصلی (PCA)

در کنار تجزیه خوشه‌ای، PCA از متداول‌ترین روشهای آماری چندمتغیره در مطالعات روابط ژنتیکی ژنوتیپ‌ها است. هدف از این تجزیه، یافتن ترکیباتی از p متغیر x_1، x_2 و ... و x_p جهت ایجاد شاخص‌های مستقل (غیر همبسته) z_1، z_2، ... و z_p به‌نام مولفه‌های اصلی[2] (PC) می‌باشد. PC ها به‌علت عدم همبستگی جنبه‌های متفاوتی از خصوصیات داده‌های اولیه را بیان می‌کنند و طوری مرتب می‌شوند که z_1 بیشترین مقدار تغییرات را داشته باشد، z_2 در مرتبه بعدی قرار می‌گیرد و الا آخر ... . PCAبا استفاده از ماتریس واریانس–کوواریانس و یا ماتریس همبستگی انجام می‌شود. ولی در مورد داده‌های مولکولی ماتریس شباهت یا فاصله بین افراد به‌عنوان ورودی PCA استفاده می‌شود. در استفاده از داده-های مولکولی در تجزیه به مولفه‌های اصلی، احتمال ریشه‌های راکد[3] منفی وجود دارد که برای برطرف کردن این مشکل، های فرانکو و همکاران (۱۹۹۷) پیشنهاد کردند که باید ماتریس شباهت یا فاصله با استفاده از فرمول $S'_{ij}=S_{ij}-S_{i.}-S_{.j}+S_{..}$ تبدیل داده شود. که در آن S_{ij} ضریب شباهت بین فرد i و j، $S_{i.}$ و $S_{.j}$ به ترتیب میانگین ضرایب شباهت برای فردi ام و j ام و $S_{..}$ میانگین کل ضرایب شباهت می‌باشد.

[1] Consensus cluster
[2] Principal components
[3] Eigen value

درمورد صفات کمی در بیشتر مواقع دو یا سه PC اول بیشترین مقدار تغییرات مربوط بـه داده‌هـای اولیه را توجیه می‌کنند (حـدود ۸۰٪-۷۵٪) و ایـن PC هـا بـرای نمـایش گرافیکـی جهـت گـروه‌بنـدی ژنوتیپ‌ها استفاده می‌شوند. ولی در مورد داده‌های مولکولی دو یا سه PC اول حداکثر حـدود ۲۰٪-۱۰٪ تغییرات مربوط به تغییرات اولیه نشانگرها را توجیه می‌کنند. هر چند که این نتایج ممکن اسـت از نقطـه نظر آماری برای PCA و نمایش گرافیکی مناسب نباشد ولی از نقطه نظـر ژنتیکـی نـشان‌دهنـده نمونـه-برداری مطلوب نشانگرها از کل ژنوم می‌باشد. به این ترتیب که هر یک از نشانگرهای مـورد اسـتفاده از بخش‌های متفاوت ژنوم بوده، بنابراین دارای همبستگی کمتر هستند. در تجزیه به مولفـه‌هـای اصـلی بـا استفاده از داده‌ها مولکولی باید توجه کرد که نمـایش گرافیکـی و گـروه‌بنـدی براسـاس دو یـا سـه PC نمی‌تواند نشان‌دهنده تغییرات کل متغیرهای اولیه باشد. درنتیجه توصیه می‌شود که گروه‌بنـدی براسـاس تعداد زیاد PC که تغییرات بیشتری را توجیه می‌کننـد انجـام گیـرد. انتخـاب تعـداد PC در ایـن حالـت می‌تواند براساس ریشه‌های راکد PC ها باشد. PC هایی با ریشه راکد بزرگتر از یک بـه‌عنـوان PC هـای موثر در گروه‌بندی استفاده شوند (ایزونـی و پـریتس، ۱۹۹۱). مـشکل دیگـر در PCA مـشکل داده‌هـای گمشده است که امری متداول در مورد داده‌های مولکولی است. در PCA، داده گمشده به‌طـور سـاده بـا میانگین صفت (متغیر) مربوطه در موقع محاسبه ماتریس فاصله یا شباهت جایگزین مـی‌شـود. درنتیجـه ژنوتیپ‌هایی که دارای داده گمشده بیشتری هستند نزدیک به مرکز گروه مربوطه قـرار مـی‌گیرنـد. بـرای برطرف کردن این مشکل در PCA، توصیه می‌شود که ضریب شباهت یا فاصله دو فرد جداگانه فقـط بـا استفاده از صفات یا متغیرهایی که دارای داده کامل برای هر دو فرد می‌باشند محاسبه گردد و متغیرها یـا صفاتی که دارای داده گمشده برای یک یـا هـر دو فـرد هـستند از محاسـبه ضـریب آن دو فـرد حـذف شود (Rohlf, 1972).

تجزیه به مولفه‌های اصلی می‌تواند به‌عنوان روشی جهت تعیین تعداد مطلوب خوشـه نیـز اسـتفاده شود. تعداد مطلوب خوشه تعدادی خواهد بود که با آن تعداد خوشـه اولـین PC در هـر خوشـه بتوانـد حداکثر تغییرات را توجیه کند. بدین ترتیب که ابتدا همه افراد به‌عنوان یک خوشه درنظر گرفتـه شـده و سپس در هم ادغام می‌شوند تا جایی‌که در همه خوشه‌های تشکیل شده تغییرات توجیه شده توسـط PC دوم کمتر از مقدار تعیین شده باشد (Thompson et al., 1998) .

مقایسه کارآیی تجزیه خوشه‌ای و تجزیه به مولفه‌های اصلی

نتایج حاصل از تجزیه خوشه‌ای و تجزیه به مولفه‌های اصلی در بررسی تنوع ژنتیکی بـا اسـتفاده از داده‌های مولکولی نشان داده است که در حالت ادغام منـابع ژرم‌پلاسـمی الگـوی تنـوع ژنتیکـی حالـت طبقه‌ای نداشته و استفاده از الگوریتم‌های تجزیه خوشه‌ای دارای محـدودیت در گـروه‌بنـدی ایـن مـواد ژنتیکی می‌باشند. بنابراین در چنین موارد تجزیه به مولفه‌های اصلی و سایر تکنیک‌های مرتبط مـی‌توانـد

روشی جایگزین برای گروهب‌ندی افراد باشند (Mohammadi and Prasanna, 2003). با وجود این، ملچینگر (۱۹۹۳) با مقایسه پنج سری داده در ذرت و جو نتیجه‌گیری کرد که اگر دو یا سه PC کمتر از ۲۵ درصد تغییرات کل را توجیه کنند، تجزیه به مولفه‌های اصلی در گروه‌بندی ژنوتیپ‌ها دچار اشکال می‌شود. بنابراین، در این شرایط تجزیه خوشه‌ای روشی مناسب برای تعیین روابط شجره‌ای افراد می‌باشد. مزیت اصلی تجزیه به مولفه‌های اصلی درمقایسه با تجزیه خوشه‌ای در گروه‌بندی افراد این است که در تجزیه خوشه‌ای یک فرد ممکن است همزمان در گروه‌های متفاوت قرار گیرد ولی در تجزیه به مولفه اصلی این مشکل وجود ندارد. درکل باتوجه به محدودیت‌ها و محاسن روشهای مختلف، استفاده از روشهای متفاوت با آگاهی از اصول آماری آنها و جنبه بیولوژیکی داده‌ها می‌تواند محققین را در تجزیه و تحلیل صحیح و کارآیی دادهها در جهت گروه‌بندی منطقی و واقعی ژنوتیپ‌ها یاری نماید.

بررسی تغییرات ژنتیکی جمعیت‌ها در سطوح مختلف با استفاده از داده‌های مولکولی

زمانی‌که جمعیتی به زیر جمعیت‌های مجزا تقسیم می‌شود در بیشتر موارد انتظار می‌رود به‌علت فراوانی بالای تلاقی‌های خویشاوندی، سطح هتروزیگوتی یا به عبارت سطح تنوع ژنی در زیر جمعیت‌ها نسبت به جمعیت پایه کاهش یابد. همچنین در زیر واحدهای جمعیت به دلیل اندازه کوچک آنها، اشتباه نمونه‌گیری ژنی بیشتر خواهد بود. بنابراین راندهشدگی ژنتیکی همراه با اندازه کوچک زیر جمعیت سبب تثبیت برخی از آلل‌ها و درنتیجه باعث تفاوت فراوانی ژنی زیرجمعیت‌ها با جمعیت پایه و اصلی می‌شود. تفاوت جمعیت‌های مختلف در سطح مولکولی را می‌توان براساس تفاوت آنها در فراوانی‌های آللی در جایگاه‌های نشانگری و نیز به علت تفاوت در سطح هتروزیگوتی آنها مشخص کرد.

کاهش در سطح هتروزیگوتی جمعیت‌ها و زیرجمعیت‌ها معمولاً توسط شاخصی به نام آماره F رایت[1] که به شاخص تثبیت[2] نیز مشهور است، اندازه‌گیری می‌شود. آماره F شاخص اندازه‌گیری تفاوت بین میانگین سطح هتروزیگوتی جمعیت‌ها و یا زیرجمعیت‌ها با سطح هتروزیگوتی جمعیت پایه با تلاقی تصادفی است (Hartl and Clark, 1997) . شاخص تثبیت از نظر تئوریکی در محدود صفر (عدم وجود تفاوت بین زیر جمعیت‌ها و جمعیت پایه) و یک (حداکثر تمایز بین زیر جمعیت‌ها و جمعیت پایه یا به عبارت دیگر تثبیت آلل‌های متفاوت در جمعیت‌های مختلف) تغییر می‌کند. ولی در عمل مقدار امکان تثبیت آلل‌های متفاوت در جمعیت‌های مختلف یعنی به دست آوردن شاخص یک وجود ندارد. هر چند که معنی‌دار بودن این شاخص با استفاده از روشهایی مثل Bootstrap و Permutation قابل

[1] - Wiright's F Statistics
[2] - Fixation index

آزمون است، ولی میزان آماره F در دامنه‌های مختلف نشان‌دهنده درجات مختلف تمایز است. بدین ترتیب که مقدار آماره F در محدوده‌های صفر تا ۰/۰۵، ۰/۰۵ – ۰/۱۵، ۰/۱۵–۰/۲۵ و ۰/۲۵ < نشان-دهنده تمایز ژنتیکی بسیار پایین، متوسط، بالا و بسیار بالا است. جدول (۳) برآورد سطح هتروزیگوتی را در سطوح مختلف جمعیت نشان می‌دهد. زمانی‌که سطح هتروزیگوتی در سطوح مختلف جمعیت برآورد شد، شاخص تثبیت در هر سطح از جمعیت قابل محاسبه خواهد بود. با فرض این‌که جمعیت متشکل از گروه‌ها و گروه‌ها متشکل از زیرجمعیت‌ها باشد، آنگاه شاخص تثبیت در سطوح مختلف به شرح جدول (٤) خواهد بود.

جدول ۳ - برآورد سطح هتروزیگوتی در درون زیرجمعیت‌ها، بین زیر جمعیت و کل جمعیت پایه

سطح تقسیم	**هتروزیگوتی**
زیر جمعیت‌ها	$H_S = \sum_i \sum_j ۲ p_{ij}(۱ - p_{ij})$
گروه زیر جمعیت‌ها	$H_G = \sum_i ۲ \frac{\sum_j p_{ij}}{n_i}(۱ - \frac{\sum_j p_{ij}}{n_i})$
جمعیت پایه (کل)	$H_T = \sum ۲ \frac{\sum_i \sum_j p_{ij}}{n_i}(۱ - \frac{\sum_i \sum_j p_{ij}}{n_i}$

که در آن p_{ij} فراوانی الل jام در جمعیت i ام و n_i اندازه زیر جمعیت i ام می‌باشد.

جدول ٤ - شاخص تثبیت برای سطوح مختلف جمعیت

سطح تقسیم	**شاخص تثبیت (آماره F)**
بین زیر جمعیت‌ها درون گروه	$F_{SG} = \frac{H_G - H_S}{H_G}$
بین گروه‌های درون جمعیت	$F_{GT} = \frac{H_T - H_G}{H_T}$
بین زیر جمعیت‌های درون جمعیت	$F_{ST} = \frac{H_T - H_S}{H_T}$

تجزیه واریانس مولکولی (AMOVA)

آماره F رایت براساس مقایسه فراوانی‌های ژنی بین زیرجمعیت‌ها و زیرگروه‌ها می‌باشد، با این وجود در تجزیه واریانس مولکولی نه تنها تفاوت‌ها براساس فراوانی نشانگرهای مولکولی بیان می‌شود بلکه می‌توان تفاوت جهش‌های ژن‌های مختلف را نیز در تفکیک واریانس کل به اجزای آن در نظر گرفت. روش تجزیه واریانس اولین بار توسط کوکرهام (۱۹۷۳) برای مطالعه ساختار ژنتیکی جمعیت‌ها براساس فراوانی‌های ژنی استفاده شد. اکسکوفیر و همکاران (۱۹۹۲) روش تجزیه واریانس مولکولی را برای مطالعه ساختار ژنتیکی جمعیت‌ها براساس داده‌های مولکولی به‌خصوص داده‌های حاصل از نشانگرهای مولکولی پیشنهاد کردند. این روش اساساً مشابه سایر روشهای تجزیه واریانس مبتنی بر فراوانی‌های ژنی است، با این تفاوت که داده‌های اولیه به‌صورت وجود (یک) و یا عدم وجود (صفر) یک نشانگر یا یک جهش درنظر گرفته می‌شود. بدین ترتیب در تجزیه واریانس مولکولی داده‌ها به صورت ماتریس صفر و یک مرتب شده و مربع فاصله اقلیدسی بین افراد به صورت زیر محاسبه می‌شود :

$$D^{2}_{jk} = (p_j - p_k)' \, w(p_j - p_k)$$

که در آن p_j و p_k بردار صفر و یک برای فرد j ام و k ام و w ماتریس واحد است. درصورتی‌که برخی از نشانگرهای مستقل نبوده و یا دارای اهمیت بیشتری باشد. ماتریس w یک ماتریس وزنی خواهد بود. پس از محاسبه فاصله اقلیدمی بین افراد، آنها در یک ماتریس فاصله مرتب می‌گردند :

$$D^2 = \begin{bmatrix} \begin{bmatrix} D^2_{11} & D^2_{12} \\ D^2_{21} & \end{bmatrix} & \dots D^2_{1k} & . \\ D^2_{21} & D^2_{22} & \dots D^2_{2k} \\ \vdots & \vdots & \vdots \\ D^2_{j1} & D^2_{j2} & \dots D^2_{jk} \end{bmatrix}$$

در این ماتریس زیرماتریس‌های روی قطر اصلی نشان‌دهنده فاصله افراد در یک گروه یا زیرجمعیت و زیرماتریس‌های خارج قطر نشان‌دهنده فاصله افراد از گروه‌ها یا زیرجمعیت‌های مختلف می‌باشند. با استفاده از این ماتریس فاصله تجزیه واریانس آشیانه‌ای برای تفکیک تغییرات کل ژنتیکی به اجزای تشکیل‌دهنده آن برحسب تقسیمات جمعیت انجام می‌گیرد (Excoffier et al., 1992). به‌عنوان مثال اگر چند جمعیت از مناطق مختلف مورد مطالعه باشند، جدول تجزیه واریانس مولکولی به شرح زیر خواهد بود (جدول ۵).

جدول ۵ - جدول تجزیه واریانس مولکولی برای مطالعه ساختار ژنتیکی جمعیت‌ها

منابع تغییر	درجات آزادی	میانگین مربعات	ارزش مورد انتظار
بین مناطق	$G_g - ۱$	MS_{AG}	$\sigma_c^۲ + n'\sigma_b^۲ + n''\sigma_a^۲$
بین جمعیت‌ها درون مناطق	$\sum_{g=۱}^{G} I_g - G$	$MS_{AP/WG}$	$\sigma_c^۲ + n\sigma_b^۲$
بین افراد درون جمعیت	$N - \sum_{g=۱}^{G} I_g$	MS_{WP}	$\sigma_c^۲$
کل	$N-۱$		

که در آن G تعداد مناطق، I_g تعداد جمعیت از منطقه g ام و N تعداد کل افراد در مجموع جمعیت است و n و n' و n'' به صورت زیر محاسبه می‌گردند :

$$n = \frac{\sum_{g=۱}^{G}\sum_{i=۱}^{I_g} N_{ig} - \sum_{g=۱}^{G}\left(\frac{\sum_{i=۱}^{I_g} N_{ig}^۲}{\sum_{i=۱}^{I_g} N_{ig}}\right)}{\sum_{g=۱}^{G} I_g}$$

$$n' = \frac{\sum_{g=۱}^{G}\left(\frac{\sum_{i=۱}^{I_g} N_{ig}^۲}{\sum_{i=۱}^{I_g} N_{ig}}\right) - \frac{\sum_{g=۱}^{G}\sum_{j=۱}^{I_g} N_{ig}^۲}{\sum_{g=۱}^{G}\sum_{i=۱}^{I_g} N_{ig}}}{G-۱}$$

$$n'' = \frac{\sum_{g=1}^{G}\sum_{i=1}^{I_g} N_{ig} - \frac{\sum_{g=1}^{G}\left(\sum_{i=1}^{I_g} N_{ig}\right)^2}{\sum_{g=1}^{G}\sum_{i=1}^{I_g} N_{ig}}}{G-1}$$

در فرمول‌های فوق N_{ig} نشان‌دهنده اندازه نمونه در یک سطح خاص جمعیت مثل گروه یا زیرجمعیت می‌باشد.

پس از برآورد واریانس واقعی هر سطح جمعیت، آماره Φ برای هر سطح محاسبه می‌گردد :

$$\Phi_{AG} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2} \quad , \quad \Phi_{AP/WG} = \frac{\sigma_a^F}{\sigma^F} \quad , \quad \Phi_{WP} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma^2}$$

$$\sigma^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2$$

آماره Φ به‌عنوان معیاری برای آزمون فرض تمایز جعیت در سطح مربوطه استفاده می‌شود. به‌عنوان مثال Φ_{WP}میزان تمایز بین جمعیت و افراد آن را نشان می‌دهد. چون داده‌های مولکولی در تجزیه واریانس مولکولی به‌صورت صفر و یک امیتازدهی می‌شوند و دارای توزیع نرمال نیستند، بنابراین آزمون فرض‌های در این مورد از طریق روشهای resampling مثل bootstrap انجام می‌شود (Excoffier et al., 1992).

پیش‌بینی عملکرد هیبریدها براساس داده‌های مولکولی

روشهای مختلفی برای پیش‌بینی عملکرد هیبریدها در گیاهان مختلف به‌خصوص ذرت استفاده شده است. باتوجه به مشکلات موجود در این روشها و عدم کارایی آنها در پیش‌بینی قابل اعتماد، در سال‌های اخیر با توسعه نشانگرهای مولکولی متعدد، استفاده از مدل‌های مبتنی بر فاصله والدین براساس داده‌های مولکولی در پیش‌بینی عملکرد هیبریدها مورد توجه قرار گرفت. پتانسیل مدل فاصله براساس نشانگرهای DNA در مطالعات مختلف در ذرت جایی‌که فاصله ژنتیکی والدین با استفاده از نشانگرهای RFLP (Lee *et a*l., 1989; Melchinger *et al*., 1992; Burstin *et al.,* 1995) SSRs (Mohammadi *et al*., 2002, 2003) و AFLP (Ajmone-Marsan *et al.*, 1998; Vuylsteke *et al.,* 2000) برآورد شده، مورد بررسی قرار گرفته است. در برنج نیز مطالعاتی با استفاده از نشانگرهای ریزماهواره در این راستا انجام شده است (Saghai-Maroof et al., 1997; Zhang et al., 1994, 1996). همبستگی‌های پایین تا بالا در مطالعات مختلف گزارش شده است. به‌طورکلی بدون در نظر گرفتن نوع تلاقی (درون گروهی و یا بین گروهی) همبستگی پایین بین فاصله ژنتیکی والدین براساس نشانگرهای مولکولی و عملکرد

هیبریدهای آنها می‌تواند ناشی از انتخاب تصادفی نشانگرها و برحسب پوشش ژنومی آنها و نه براساس ارتباط و یا پیوستگی آنها با QTL[1]ها دخیل در هتروزیس صورت باشد. بررسی‌ها نشان داده‌اند که وجود نشانگرهای تصادفی غیرمرتبط با QTLها باعث کاهش همبستگی می‌گردد. محمدی و همکاران (۲۰۰۲، ۲۰۰۳) در بررسی رابطه تنوع ژنتیکی والدین براساس نشانگرهای SSR و عملکرد هیبریدها در سه سری تلاقی دی آلل، لاین در تستر و تست کراس با دو تستر هتروتیک با شناسایی نشانگرهای *informative* برای عملکرد و اجزای آن (نشانگرهای مرتبط با نواحی ژنومی دخیل در عملکرد و اجزای آن) نشان دادند که میزان همبستگی به‌خصوص در تلاقی دی آلل بهبود می‌یابد. نتایج مشابه توسط زنگ و همکاران (۱۹۹۴) نیز در برنج به‌دست آمده است. آنها گزارش کردند محاسبه فاصله ژنتیکی والدین با استفاده از نشانگرهای مثبت یا *informative* برآورد بهتر عملکرد هیبریدها را فراهم می‌کند. بنابراین می‌توان نتیجه گرفت که افزایش تعداد نشانگر به تنهایی و بدون درنظر گرفتن موقعیت ژنومی آنها و ارتباط آنها با QTLها موجب بهبود در همبستگی نخواهد شد.

جهت بهبود همبستگی بین فاصله والدین و عملکرد هیبریدها، برخی از محققین پیشنهاد کردند در تجزیه دی آلل که والدین با استفاده از نشانگرهای مولکولی ارزیابی شده‌اند، تفکیک فاصله ژنتیکی لاین‌های والدینی براساس داده‌های مولکولی به فاصله ژنتیکی عمومی و خصوصی می‌تواند اطلاعات بیشتری را در ارائه کند (Melchinger *et al.*, 1992, Mohammadi, 2001). فاصله ژنتیکی والدین به فاصله ژنتیکی عمومی و خصوصی مطابق با مدل یک و روش چهار گریفینگ به صورت زیر امکان‌پذیر است (Melchinger *et al.*, 1992) :

$$GD_{ij} = \overline{GD} + GGD_i + GGD_j + SGD_{ij}$$

که در آن GD_{ij} : فاصلهٔ ژنتیکی دو والد i و j ، $\overline{GD}$: میانگین فواصل ژنتیکی بین والدین مختلف، GGD_i : فاصلهٔ ژنتیکی عمومی والد i ام، GGD_j : فاصلهٔ ژنتیکی عمومی والد j ام، SGD_{ij} : فاصلهٔ ژنتیکی خصوصی دو والد i و j

GGD : انحراف میانگین فاصلهٔ ژنتیکی یک والد با سایر والدین از میانگین فواصل ژنتیکی کل والدین است.

تعیین نشانگرهای مثبت[2] یا Informative

شناسایی نشانگرهای مثبت بر مبنای عدم تعادل پیوستگی در مجموعه‌ای از افراد یا جمعیت می‌باشد. برای شناسایی نشانگرهای مثبت یا Informative مرتبط با صفت مورد اندازه‌گیری، می‌توان از

[1] Quantitative Trait Loci
[2] - Positive markers

رگرسیون چندمتغیره گام به گام استفاده کرد. بدین ترتیب که ارزش افراد برای صفت به‌عنوان متغیر وابسته و ارزش افراد برای نشانگرهای مولکولی به‌صورت صفر و یک و یا فراوانی به‌عنوان متغیرهای مستقل در مدل استفاده می‌شوند. در تجزیه متغیرهایی که دارای ضرایب رگرسیون معنی‌دار هستند و وارد مدل شدند به‌عنوان نشانگرهای مثبت برای صفت درنظر گرفته می‌شوند (Mohammadi, 2001). برای کاهش میزان خطای نوع اول، سطح احتمال ورود و خروج متغیر در مدل پایین درنظر گرفته می‌شود. درصورتی‌که آزمایش درقالب تلاقی‌های دی آلل باشد، ژنوتیپ هیبریدها براساس ژنوتیپ والدین آنها تعیین می‌شود. بدین ترتیب که اگر آلل مورد نظر در هیچ یک از والدین هیبرید وجود نداشته باشد، برای آن هیبرید برای آلل موردنظر عدد صفر، اگر آلل در یکی از والدین باشد، عدد ۰/۵ و اگر آلل در هر دو والد هیبرید وجود داشته باشد، به هیبرید امتیاز یک داده می‌شود. پس از تعیین نشانگرهای مثبت می‌توان اثر افزایشی و درجه غالبیت نواحی کروموزومی حامل نشانگرهای مثبت را با استفاده از فرمول‌های زیر برآورد کرد (Mohammadi et al., 2002) :

$$a = \sum_{i=j}^{k-1} \sum_{j=i+1}^{k} \left(N_{ij} \left| X_{ii} - X_{jj} \right| / 2 \right) / \sum N_{ij}$$

$$d = \sum_{i=j}^{k-1} \sum_{j=i+1}^{k} \left(N_{ij} \left| X_{ij} - \left(X_{ii} + X_{jj} \right) / 2 \right| \right) / \sum N_{ij}$$

که در آنها، X_{ii}: میانگین ارزش صفت برای افراد هوموزیگوت برای آلل i ام جایگاه نشانگر، X_{jj}: میانگین ارزش صفت برای افراد هوموزیگوت برای الل jام جایگاه نشانگر، X_{ij}: میانگین ارزش صفت برای افراد هتروزیگوت برای آلل‌های i ام و jام جایگاه نشانگر، k: تعداد کل آلل‌ها در یک مکان نشانگر، N_{ij}: تعداد هیبریدهای هتروزیگوت برای الل i ام و jام است. درجه غالبیت برای هر جایگاه کروموزومی از فرمول $\frac{d}{a}$ قابل محاسبه است.

نتیجه‌گیری

توسعه تکنولوژی‌های نوین زیستی و تکنیک‌های مرتبط دریچه جدیدی را در مطالعات ژنتیکی، اصلاحی و زمینه‌های مرتبط ایجاد کرده است. با ابداع نشانگرهای DNA و امکان ارزیابی تفاوت‌های افراد در سطح مولکولی، حجم عظیمی از داده‌ها تولید می‌شود که تجزیه و تحلیل صحیح آنها به همان اندازه استفاده از این فن‌آوری‌ها مهم می‌باشد. در این راستا برنامه‌های کامپیوتری متعددی نیز طراحی شده‌اند. ولی عدم آشنایی اکثر کاربران تکنیک‌های مولکولی به اساس ژنتیکی و به‌خصوص آماری مدل‌ها و روشهای مورد استفاده سبب نتیجه‌گیری‌های غیرقابل اعتماد شده است.

مشکل اصلی در تجزیه ژنتیکی روابط بین افراد یا جمعیت‌ها با استفاده از نشانگرهای مولکولی، انتخاب نوع ضریب فاصله یا شباهت مورد استفاده است که در اکثر مطالعات انجام شده بدون توجه به ماهیت داده‌ها و نوع مواد ژنتیکی مورد ارزیابی، فقط براساس مطالعات قبلی انتخاب شده‌اند. هرچند که روش یا مدل جامع برای انتخاب یک ضریب خاص وجود ندارد، ولی آگاهی از اساس ژنتیکی و آماری ضرایب مختلف می‌تواند محققین را در انتخاب ضرایب مناسب کمک کند. مشکلات مشابه در کاربرد روشهای گروه‌بندی نیز وجود دارد که از جمله مهمترین آنها استفاده از الگوریتم‌های تجزیه خوشه‌ای در گروه‌بندی افراد یا جمعیت‌ها است. یکی از متداول‌ترین الگوریتم‌ها، UPGMA است که در اکثر مطالعات بدون درنظر گرفتن نوع ژرم‌پلاسم‌ها و ساختار آنها استفاده شده است. این الگوریتم در مواردی که ژرم‌پلاسم مورد مطالعه دارای ادغام باشد، دندروگرام‌هایی با اثر زنجیره‌ای تولید خواهد کرد که با استفاده از آنها امکان گروه‌بندی و تفسیر نتایج وجود نخواهد داشت.

منابع مورد استفاده

۱. محمدی، س. ا. ۱۳۸۱. روشهای آماری در ژنتیک. مجموعه مقالات ششمین کنفرانس بین‌المللی آمار ایران. ٤ تا ٦ شهریور ماه ۱۳۸۱، دانشگاه تربیت مدرس، تهران، صفحه ۳۹٤-۳۷۱.

2. Ajmone-Marsan, P., Castiglioni, P., Fusari, F., Kuiper, M. and Motto, M. 1998. Genetic diversity and its relationship to hybrid performance in maize as revealed by RFLP and AFLP markers. Theor. Appl. Genet. 96: 219-227.
3. Burstin, J., de Vienne, D., Dubreuil, P. and Damerval, C. 1994. Molecular markers and protein quantities as genetic descriptors in maize. I. Genetic diversity among 21 inbred lines. Theor. Appl. Genet. 89: 943-950.
4. Cockerham, C. C. 1973. Analysis of gene frequencies. Genetics 74: 679-700.
5. Excoffier, L., Smouse, P. E., and Quattro, J. M. 1992. Analysis of molecular variance inferred from metric distance among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131: 479-791.
6. Franco, J., Crossa, J., Villasenor, J., Taba, S., and Eberhart, S. A. 1997. Classifying Mexican maize accessions using hierarchical and density search methods. Crop Sci. 37: 972-980.
7. Johnson, A. R. and D. W. Wichern. 1992. Applied Multivariate Statistical Analysis. 3rd ed. Prentice-Hall, Englewood Cliffs, New Jersey.
8. Hartl, D. L., and Clark, A. G. 1997. Principles of population Genetics. 3rd ed. Sinauer Associates Inc. Sunderland, MA.
9. *Lee, M. 1998. DNA markers for detecting genetic relationships among germplasm and for establishing heterotic groups. Lecture during Maize Training Course, CIMMYT, Mexico.*
10. Karp, A., Kresovich, S., Bhat, K. V., Ayad, W. G. and Hodgkin, T. 1997. Molecular Tools in Plant Genetic Resources Conservation: A Guide to the Technologies. IPGRI, Rome, Italy.
11. Lombard, V., Baril, C. P., Dubreuil, P., Blouet, F., and Zhang, D. 2000. Genetic relationships and fingerprinting of rapeseed cultivars by AFLP: Consequences for varietal registration. Crop Sci. 40: 1417-1425.
12. Melchinger, A. E. 1993. Use of RFLP markers for analyses of genetic relationships among breeding materials and prediction of hybrid performance. p. 621-628. *In:* Buxton, D.R. (ed.) 1st Int. Crop Sci. Congress, Crop Science Society of America.
13. Melchinger, A. E., Boppenmaier, J., Dhillon, B. S., Pollmer, W. G. and Herrmann, R. G. 1992. Genetic diversity for RFLPs in European maize inbreds. II. Relation to performance of hybrids within versus between heterotic groups for forage traits. Theor. Appl. Genet. 84: 672–681.
14. Mohammadi, S. A. and Prasanna, B. M. 2003. Analysis of genetic diversity in crop plants: Salient statistical tools and considerations. Crop Sci. 43: 1235-1248.
15. Mohammadi, S. A., B. M. Prasanna, C. Cudan, and N. N. Singh. 2002. A microsatellite marker based study of chromosomal regions and gene effects on yield and yield components in maize. Cellular & Molecular Biology Letters 7: 599-606.
16. Mohammadi, S. A., B. M. Prasanna, C. Cudan, and N. N. Singh. 2003. SSR heterogenic patterns of parents and hybrid performance in maize. Abstracts Book of 12th Meeting of the EUCARPIA, Section of Biometrics in Plant Breeding: Integrated quantitative and molecular genetics in plant breeding. Sept. 3-5, A Coruna Spain. pp. 102.
17. Nei, M. 1973. Analysis of genetic diversity in subdivided populations Proc. Natl. Acad. Sci. USA 70: 3321-3323.
18. Pritchard, J. K., Stephens, M. and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

19. Reif, J. C. Melchinger, A. E.and Frisch, M. 2005. Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Sci. 45: 1-7.
20. Rincon, F., Johnson, B., Crossa, J., and Taba, S. 1996. Cluster analysis, an approach to sampling variability in maize accessions. Maydica 41: 307-316.
21. Rohlf, F. J. 1972. An empirical comparison of three ordination techniques in numerical taxonomy. Syst. Zool. 21: 271-280.
22. Saghai Maroof, M. A., Yang, G. P., Zhang, Q. and Gravios, K. A. 1997. Correlation between molecular marker distance and hybrid performance in U.S. Southern long rice. Crop Sci. 37: 145-150.
23. Thompson, J. A., Nelson, R. L. and Vodkin, L. O. 1998. Identification of diverse soybean germplasm using RAPD markers. Crop Sci. 38: 1348-1355.
24. Tivang, G., Nienhuis, J. and Smith, O. S. 1994. Estimation of sampling variance of molecular marker data using the bootstrap procedure. Theor. Appl. Genet. 89: 259-264.
25. Vuylsteke, M., R. Mank, B. Brugmans, P. Stam, and M. kuiper. 2000. Further characterization of AFLP data as tool in genetic diversity assessments among maize (*Zea mays* L.) inbred lines. Mol. Breed. 6: 265-276.
26. Wishart, D. 1987. CLUSTAN User Manual, 3rd edition, Program Library Unit, Univ. of Edinburgh, Edinburgh.
27. Zhang, Q., Y. J. Gao, S. H. Yang, R. Ragab, M. A. Saghai-Maroof, and Z. B. Li. 1994. A diallel analysis of heterosis in elite hybrid rice based on RFLPs and microsatellites. Theor. Appl. Genet. 89: 185-192.