نشریه زراعت و فناوری زعفران
جلد ۲، شماره ۳، پاییز ۱۳۹۳ ص: ۲۰۴-۱۹۹

Saffron Agronomy & Technology
Vol. 2, No. 3, Fall 2014, P. 199-204

دسته‌بندی و پیش‌بینی کلاله سه‌شاخه و چندشاخه زعفران با استفاده از ابزارهای آماری یادگیری ماشینی بدون‌نظارت

امیر حسین بیکی[1*]

تاریخ دریافت: ۱۳۹۳/۴/۲۴ تاریخ پذیرش: ۱۳۹۳/۶/۲۵

چکیده

زعفران یک گیاه تریپلوئید و عقیم است که در همه کشورها به‌عنوان یک ادویه و گیاه دارویی مورد استفاده قرار می‌گیرد. کلاله مهم‌ترین قسمت گیاه زعفران می‌باشد. تاکنون هیچ روش مطمئن مولکولی برای شناسایی و پیش‌بینی گونه‌های دارای کلاله سه و چندشاخه ارائه نشده است. در این بررسی بر اساس نشانگرهای مولکولی چندشکلی توالی مربوط تکثیر یافته و با استفاده از الگوریتم‌های بیوانفورماتیکی مختلف،روش‌های جدیدی برای پیش‌بینی کلاله زعفران ارائه شده است. پنج آلل M131400، M151200، M151100، M10850 و G6500 به‌عنوان مهم‌ترین دسته‌بندی کننده با دقت پیش‌بینی بالا بر اساس مدل‌های Attribute Weighting انتخاب شدند که دارای پتانسیل بالایی برای خوشه‌بندی و تشخیص کلاله سه‌شاخه ازچندشاخه هستند. دسته‌بندی بدون یادگیری بر اساس الگوریتم‌های K-Means و K-Medoids قادر به خوشه‌بندی صحیح کلاله زعفران هستند. نتایج نشان می‌دهد که برای اولین بار، روش‌های داده‌کاوی می‌توانند شیوه‌ای بسیار مؤثر، با دقت و صحت بالای ۹۰ درصد برای تمایز ژنتیکی کلاله سه‌شاخه از چندشاخه مورد استفاده قرار گیرد. این روش‌ها می‌توانند در مکان‌یابی ژنی و انتخاب به کمک بیومارکرها مورد استفاده قرار گیرند.

کلمات کلیدی: چندشکلی توالی مربوط تکثیر یافته، مارکر مولکولی، یادگیری ماشینی.

مقدمه

زعفران زراعی (*Crocus sativus* L.) یکی از اعضای خانواده زنبق (Iridaceae) می‌باشد. شناسایی ژرم‌پلاسم زعفران زراعی و گونه‌های وحشی جنس زعفران و بررسی روابط خویشاوندی بین آن‌ها می‌تواند در حفظ ذخایر ژنتیکی و شناسایی گونه‌های وحشی وابسته جهت افزایش مواد مؤثره مورد استفاده قرار گیرد. ویژگی‌های مورفولوژیکی، مثل طول و ضخامت کلاله، به‌واسطه نقش آن در خصوصیات ادویه‌ای، اهمیت زیادی در تجارت زعفران دارد. به همین خاطر در ارزیابی تنوع و طبقه‌بندی آن مورد استفاده قرار می‌گیرد (Caiola and Canini, 2010). مختلف را می‌توان بر اساس طول و ضخامت کلاله طبقه بندی کرد (Neghbi, 2003).

گل‌های زعفران دوجنسی هستند. حلقه اول شامل ۳ کاسبرگ، حلقه دوم شامل ۳ گلبرگ و حلقه‌های بعدی به ترتیب شامل ۳ پرچم و یک مادگی با کلاله سه‌شاخه است (Tsaftaris et al., 2007). موتانت‌های مختلف فنوتیپی در گل‌های زعفران گزارش شده است (Caiola et al., 2004). در سال‌های اخیر، کلاله چندشاخه نیز در گونه‌های زراعی گزارش شده است (Beiki et al., 2011). با توجه به ارزش اقتصادی بسیار بالای کلاله زعفران، درک تکوین گل‌های زعفران می‌تواند راه‌های افزایش عملکرد را آشکار نماید (Tsaftaris

۱- استادیارگروه زیست شناسی، دانشکده علوم پایه، دانشگاه قم.
(*-نویسنده مسئول: (Email: amirbeiki@gmail.com

et al., 2005). ژن‌های مختلفی در فرآیند تکوینی گلدهی دخالت دارند. بر اساس تفاوت الگوی بیان ژن‌های MADS-box در برگ و قسمت‌های مختلف گل، می‌توان ژن‌هایی که در تشکیل قسمت‌های مختلف گل نقش دارند، تعیین نمود (Chang et al., 2010). ژن‌های MADS-box، مجموعه‌ای از فاکتورهای رونویسی را رمز می‌کنند که به فرآیندهای فوق کمک می‌کنند (Tsaftaris et al., 2006). تقریباً تمام ژن‌های مرتبط با سیستم ABC به مجموعه MADS-box تعلق دارند (Tsaftaris et al., 2007). مطالعات زیادی روی گیاه زعفران و در ارتباط با بیان ژن‌های MADS-box انجام گرفته است. این بررسی‌ها نشان می‌دهد الگوی بیان این ژن‌ها بسیار مشابه با تک‌لپه‌ای‌ها است (Chang et al., 2007; 2010; 2011; 2012; Tsaftaris et al., 2005).

تاکنون هیچ روش مولکولی مطمئنی برای شناسایی گونه‌های سه و چندشاخه ارائه نشده است (Keify and Beiki, 2012). امروزه رویکردهای مولکولی مختلف مبتنی بر واکنش زنجیره‌ای پلیمراز به خاطر سهولت و سادگی، کاربرد زیادی یافته است (Bernardo, 2008). مارکر SRAP ساده، قابل‌اعتماد و با توان عملکردی بالا است (Li & Quiros, 2001). به علت هم بارز بودن و تولید باندهایی با وضوح بالا برای استخراج از ژل به‌منظور توالی یابی مناسب است (Sunet al., 2006). از این مارکر برای بررسی تنوع ژنتیکی زعفران استفاده شده است (Keify & Beiki, 2012).

ابزارها و روش‌های محاسباتی و الگوریتم‌های مختلفی برای تجزیه‌وتحلیل اطلاعات و به تصویر کشیدن یافته‌های بیولوژیکی ارائه شده است (Hall & Holmes, 2003). روش‌های تئوریکی، تحلیلی، مدلینگ ریاضی و شبیه‌سازی کامپیوتری زیادی وجود دارد که به‌منظور مطالعه سیستم‌های بیولوژیکی و مولکولی ایجادشده است (Bishop, 2006; Ornella et al., 2012). نکته کلیدی در یادگیری ماشینی، شناسایی ویژگی‌هایی است که از آن‌ها می‌توان برای ساخت یک مدل طبقه‌بندی استفاده کرد (Mitra and Acharya, 2005). روش فوق یک شیوه جدید علمی در حال ظهور با تأثیر انقلابی در زمینه‌های مختلف است (Duda et al., 1999). این شیوه آماری امکان تشخیص الگوهای پیچیده بر اساس داده‌های موجود را فراهم می‌آورد (Steinfath et al., 2010). به علت قابلیت تعمیم بسیار بالای آن و عدم وابستگی به نوع توزیع آماری، می‌تواند به‌عنوان یک جایگزین بسیار باارزش برای روش‌های سنتی آماری محسوب شود (Maenhout et al., 2007). از این روش برای مدل‌سازی و تخمین میزان ارزش اصلاحی ژنوتیپ گیاهان گندم، جو، ذرت (Jannink et al., 2010)، پیش‌بینی بیوماس هیبریدهای آرابیدوبسیس مبتنی برSNP و مارکرهای متابولیکی مادری (Steinfath et al., 2010)، پیش‌بینی قدرت هیبریدهای ذرت (Maenhout et al., 2008)، پیش‌بینی و دسته‌بندی ارقام مقاوم به مگس زیتون (Beiki et al., 2012)، استفاده شده است. در این بررسی برای اولین بار از شیوه‌های مختلف داده‌کاوی و یادگیری ماشینی برای تعیین طبقه‌بندی کننده‌ای که می‌تواند زعفران سه‌شاخه و چندشاخه را بر اساس داده‌های به‌دست آمده از مارکرهای مولکولی SRAP متمایز نماید، استفاده شده است.

جدول ۱- توالی آغازگرهای SRAP

Table 1- Sequence of SRAP primers

	Forward پیش رو	Revers پس رو
1	TGAGTCCAAACCGGAAG	GACTGCGTACGAATTAAT
2	TGAGTCCAAACCGGATA	GACTGCGTACGAATTGCA
3	TGAGTCCAAACCGGACC	GACTGCGTACGAATTGCA
4	TGAGTCCAAACCGGAAG	GACTGCGTACGAATTAAC
5	TGAGTCCAAACCGGACC	GACTGCGTACGAATTTGA
6	TGAGTCCAAACCGGAGC	GACTGCGTACGAATTAAC
7	TGAGTCCAAACCGGAGC	GACTGCGTACGAATTGCA
8	TGAGTCCAAACCGGATA	GACTGCGTACGAATTTGA
9	TGAGTCCAAACCGGATA	GACTGCGTACGAATTTGC
10	TGAGTCCAAACCGGATA	GACTGCGTACGAATTAAT

مواد و روش‌ها

نمونه‌های گیاهی مورداستفاده از مناطق مختلف ایران از طریق بانک ژن گیاهی ملی ایران و موسسه تحقیقات بیوتکنولوژی کشاورزی کرج جمع‌آوری شد که شامل ۳۰ ژنوتیپ مختلف زعفران

زراعی و وحشی بود. استخراج DNA از بنه انجام گرفت (Beiki et al., 2011). آغازگرهای SRAP استفاده شده در این بررسی همان آغازگرهای استفاده شده برای جنس *Brassica* است (Li & Quiros, 2001) که توالی آن‌ها در جدول ۱ آمده است.

واکنش زنجیره‌ای پلیمراز بااستفاده از دستگاه ترموسایکلر در حجم ۱۵ میکرولیتر انجام شد. هر واکنش دارای ۱۰۰ نانوگرم DNA ژنومی، ۰/۲ میلی‌مولار dNTP، ۱/۵ میکرولیتر از بافر ۱۰ برابر غلظت، ۱۵ پیکومولار آغازگر، ۲/۵ میلی‌مولار MgCl2 و یک واحد Taqپلیمراز بود. پس از پایان الکتروفورز، رنگ‌آمیزی ژل با روش نیترات نقره انجام شد.

بعد از مشاهده قطعات حاصل از واکنش زنجیره‌ای پلیمراز، داده‌ها شامل ۲۴۷ آلل مختلف به‌دست آمده از بررسی مارکرهای مولکولی SRAP در ۳۰ نمونه مختلف بود. داده‌ها بر اساس وجود باند (۱) و عدم وجود باند (۰) امتیازدهی گردیدند. برای داده‌کاوی از نرم‌افزار Rapid Miner استفاده گردید (Mierswa, 2009). داده‌های تکراری بر اساس انتخاب اختصاصی مشخصه‌ها حذف شدند. همچنین مشخصه‌های زائد و غیرمفید حذف گردیدند.

برای تعیین و شناسایی آلل‌های مهم و یافتن الگوهای ممکن از ۱۰ الگوریتم مختلف وزن‌دهی استفاده گردید (Hall and Holmes, 2003). بعد از اجرای مدل‌های مختلف وزن‌دهی، آلل‌ها امتیازدهی شده و اهمیت آن‌ها در ویژگی‌ها با توجه به نشان‌دار کردن ارقام تعیین شدند. تمام الگوهای با وزن بالای ۰/۵ انتخاب شدند (Beiki et al., 2012). از داده‌های ایجادشده جدید به‌منظور آموزش متغیر هدف استفاده گردید تا به دسته‌بندی داده‌ها و پیش‌بینی آن‌ها بپردازد. الگوریتم‌های یادگیری بدون نظارت (EM،K-Medoids،K-Means وSMV) روی ۱۰ داده ایجادشده جدید و داده اولیه (FCdb) اجرا گردید. آموزش سیستم یادگیری بدون نظارت بدون حضور متغیر هدف انجام می‌گیرد (Webb, 2003) و دسته‌بندی بدون هیچ دسته از پیش تعیین‌شده‌ای انجام می‌گیرد (Fukunaga, 1990). فرضیه اصلی این است که مجموعه‌ای از ویژگی‌های مولکولی خوب، شامل ویژگی‌هایی است که مجزا با یکدیگر ولی به‌شدت با سه‌شاخه یا چندشاخه بودن کلاله هم‌بسته هستند.

جدول ۲- مهم‌ترین آلل‌های انتخاب‌شده بر اساس مدل‌های مختلف

Table 2- The most important alleles selected based on different models

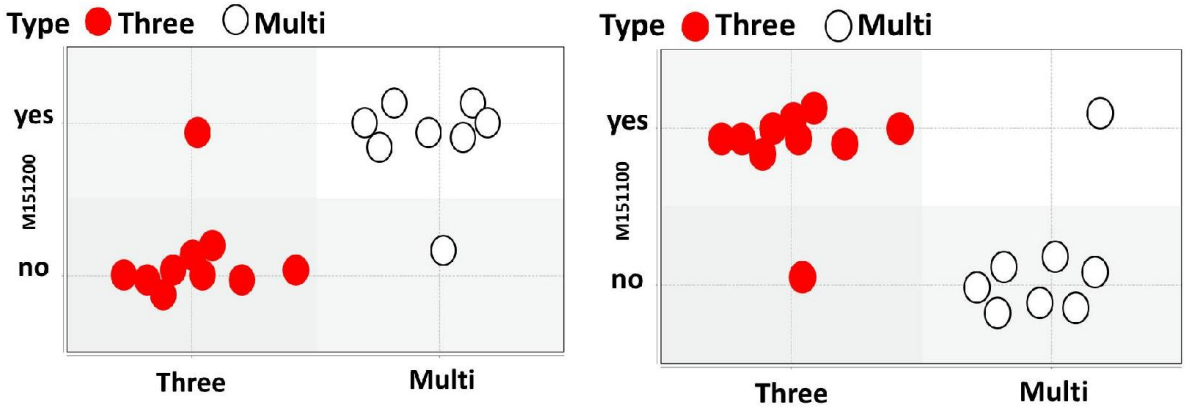
آلل Allele	درصد مدل‌هایی که آلل فوق را مناسب پیش‌بینی می‌کنند The percentage of models that predict favorable allele	میانگین وزن‌هایی که در مدل‌های مختلف کسب‌شده است Meanweights were obtained in different models
M131400	0.97	10
M151200	0.77	10
M152000	0.60	10
M13780	0.60	10
M15850	0.61	9
G121200	0.54	9
G6500	0.71	8
M121200	0.55	8
M151000	0.54	8
M151100	0.75	6
M10850	0.71	6
K61100	0.52	6

نتایج و بحث

در الگوریتم‌های یادگیری ماشینی روی ویژگی‌هایی متمرکز می‌شود که برای تجزیه‌وتحلیل و پیش‌بینی داده‌ها مفید هستند (Gentleman & Carey, 2008). سودمندی تعیین و شناسایی ویژگی‌ها برای یادگیری ماشینی می‌تواند سبب کاهش تعداد داده‌هایی باشد که برای یادگیری نیاز است (Kohavi & John, 1997) و در نتیجه موجب بهبود صحت و کاهش زمان اجرای تجزیه‌وتحلیل داده‌ها می‌شود که در موضوع داده‌کاوی بسیار مهم می‌باشند. بعد از

حذف داده‌های پرت و زائد و هم‌راستا از ۲۴۸ آلل موجود در داده‌های اولیه تعداد ۱۹۷آلل باقی ماند. نتایج حاصل از اجرای الگوریتم‌های مختلف وزن‌دهی روی داده‌های فوق نشان می‌دهد ۱۲ آلل به نام‌های M131400، M151200، M152000، M13780، M15850، G121200، G6500، M121200، M151000، M151100، M10850 و K61100 در تمام الگوریتم‌ها دارای وزنی بیش از ۰/۵ هستند و آلل M131400T در ۸۰ درصد الگوریتم‌ها دارای وزنی بیش از ۰/۵ می‌باشد (جدول ۲). میانگین وزن آلل‌های فوق همگی از ۰/۵ بیشتر است. این اطلاعات نشان‌دهنده این موضوع است که از آلل‌های فوق می‌توان در تمایز زعفران بر حسب نحوه نشان‌دار کردن آن‌ها، سه‌شاخه یا چندشاخه بودن، استفاده نمود. بر اساس الگوریتم‌های ۱۰ گانه وزن‌دهی، آلل‌های M131400T، M151200، M151100، M10850 و G6500 دارای پتانسیل بالاتری نسبت به آلل‌های دیگر برای پیش‌بینی و تشخیص کلاله سه یا چندشاخه هستند؛ زیرا در ۶۰ درصد از الگوریتم‌ها دارای حداکثر وزن، در ۸۰ درصد موارد وزنی بیش از ۰/۵ و میانگین وزنی بالاتر از ۰/۶ هستند.

برتری آلل‌های فوق را در مدل‌های مختلف می‌توان مشاهده نمود. برای مثال، همان‌طور که در نمودار ۱ سمت چپ مشاهده می‌شود به وضوح نشان می‌دهد که آلل‌های P151100 و P151200 در تمام الگوریتم‌های یادگیری بدون نظارت، پتانسیل بالایی برای پیش‌بینی نمونه‌های سه‌شاخه و چندشاخه را دارند. وجود آلل اول پیش‌بینی کننده کلاله سه‌شاخه و آلل دوم پیش‌بینی کننده کلاله چندشاخه است.



شکل ۱- دسته‌بندی کلاله سه‌شاخه با آلل M151100 و کلاله چند شاخه با آلل M151100

Figure 1- Classification of three branches stigma with M151100 fragment and multiple branches stigma with M151200.

از ۴ الگوریتم بکار رفته در یادگیری بدون نظارت، الگوریتم‌های EM و SVM قادر به دسته‌بندی صحیح زعفران سه‌شاخه از چندشاخه نیستند و هیچ الگوی مناسبی ایجاد نمی‌کنند. درصورتی‌که الگوریتم‌های K-Means و K-Medoids فقط در ترکیب با مدل‌های Deviation، PCA و SV قادر به ایجاد مدل مناسب برای دسته‌بندی زعفران بر اساس کلاله سه‌شاخه و چندشاخه می‌باشند (جدول ۳). الگوریتم K-Means در ترکیب با مدل‌های Deviation، PCA و SVM به‌ترتیب قادر به دسته‌بندی با دقت ۸۰، ۶۰ و ۲۰ درصدی نمونه‌های زعفران با کلاله سه‌شاخه و دقت ۳۸، ۶۲ و ۶۳ درصدی نمونه‌های زعفران با کلاله چندشاخه می‌باشند. درصورتی‌که الگوریتم K-Medoids در ترکیب با مدل‌های Deviation، PCA و SVM به ترتیب قادر به دسته‌بندی با دقت ۱۰۰، ۷۵ و ۷۵ درصدی کلاله چندشاخه و دقت ۷۰ درصدی کلاله سه‌شاخه در ترکیب با هر سه مدل است. به عبارت بهتر، الگوریتم K-Medoids در ترکیب با مدل PCA بهترین گزینه برای انتخاب کلاله چندشاخه و الگوریتم K-Means در ترکیب با مدل Deviation بهترین شیوه برای انتخاب کلاله سه‌شاخه می‌باشد.

علی‌رغم انتشار گزارشات و مقالات مختلف در ارتباط با مارکرهای

مولکولی و بیومارکرهای مختلف، کاربرد عملی این نتایج در اصلاح گیاهان بسیار اندک است (Benešová et al., 2012) و نتایج رضایت‌بخش نیست و بنابراین تغییر روش ضروری به نظر می‌آید (Ornella et al., 2012). این تغییر با کمک ابزارها و الگوریتم‌های مختلف داده‌کاوی امکان‌پذیر است.

جدول ۳- تعداد نمونه‌های با دسته‌بندی صحیح

Table 3- The number of samples correctly classified

الگوریتم Algorithm	مدل Model	چند شاخه Multiple branches	سه‌شاخه Three branches
K-Means	SVM	5	2
	Deviation	3	8
	PCA	5	6
K-medoids	SVM	6	7
	Deviation	6	7
	PCA	8	7

هدف اصلی در یادگیری ماشینی بدون نظارت، گروه‌بندی ژنوتیپ‌های مختلف زعفران می‌باشد که به‌شدت به کیفیت اطلاعات بکار رفته برای شیوه آموزش ماشینی بستگی دارد. اگر داده‌ها ناکافی و بی‌ربط باشند، در آموزش بدون نظارت بازتاب یافته و دسته‌بندی ژنوتیپ‌ها نامناسب خواهد بود (Pang et al., 2002). نتایج این بررسی نشان می‌دهد با استفاده از ابزارهای جدید بیوانفورماتیکی و الگوریتم‌های مختلف داده‌کاوی می‌توان الگوهای مختلفی را برای دسته‌بندی صفات مختلف نشان‌دار شده مثل کلاله سه‌شاخه و چندشاخه به کاربرد. با کمک ابزارهای فوق و بر اساس داده‌های به‌دست آمده از مارکر مولکولی SRAP می توان سه یا چندشاخه بودن کلاله زعفران را دسته‌بندی و پیش‌بینی نمود و بر اساس این دسته‌بندی به اهداف مختلف اصلاحی دست یافت. با داشتن تعداد اندکی از مارکرهای مولکولی SRAP و تکثیر قطعه مورد نظر می‌توان کلاله چندشاخه یا سه‌شاخه را با استفاده از کورم زعفران بررسی نمود و زمینه مساعدی برای اصلاح این گیاه مهم اقتصادی فراهم نمود. این روش می‌تواند در مکان‌یابی ژنی و انتخاب به کمک بیومارکرها کمک شایانی کند.

منابع

Beiki, A., Keify, F., and Mozafari, J. 2011. Rapid genomic DNA isolation from corm of *Crocus* species for genetic diversity analysis. Journal of Medicinal Plants Research 5: 4596-4600.

Beiki, A.H., Saboor, S., and Ebrahimi, M. 2012. A new avenue for classification and prediction of olive cultivars using supervised and unsupervised algorithms. PloS ONE 7: e44164.

Benešová, M., Holá, D., Fischer, L., Jedelský, P.L., Hnilička, F., Wilhelmová, N., Rothová, O., Kočová, M., Procházková, D., and Honnerová, J. 2012. The physiology and proteomics of drought tolerance in maize: early stomatal closure as a cause of lower tolerance to short-term dehydration? PloS ONE 7: e38017.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Science 48: 1649-1664.

Bishop, C.M. 2006. Pattern recognition and machine learning.Springer New York, 738 pages.

Caiola, M.G., and Canini, A. 2010. Looking for saffron's (*Crocus sativus* L.) parents. Saffron. Functional Plant Science and Biotechnology 4: 1-14.

Caiola, M.G., Caputo, P., and Zanier, R. 2004. RAPD analysis in *Crocus sativus* L. accessions and related *Crocus* species. Biologia Plantarum 48: 375-380.

Chang, Y.-Y., Kao, N.-H., Li, J.-Y., Hsu, W.-H., Liang, Y.-L., Wu, J.-W., and Yang, C.-H. 2010. Characterization of the possible roles for B class MADS box genes in regulation of perianth formation in orchid. Plant Physiology 152: 837-853.

Duda, R.O., Hart, P.E.,and Stork, D.G. 1999. Pattern classification. John Wiley & Sons, 680 pages.

Fukunaga, K. 1990. Introduction to statistical pattern recognition. Academic press, 592 pages.

Gentleman, R., and Carey, V. 2008. Unsupervised machine learning. Bioconductor Case Studies. Springer, pp. 137-157.

Hall, M.A., and Holmes, G. 2003. Benchmarking attribute selection techniques for discrete class data mining. Knowledge and Data Engineering, IEEE Transactions on 15: 1437-1447.

Jannink, J.-L., Lorenz, A.J., and Iwata, H. 2010. Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics 9: 166-177.

Keify, F., and Beiki, A.H. 2012. Exploitation of random amplified polymorphic DNA (RAPD) and sequence-related amplified polymorphism (SRAP) markers for genetic diversity of saffron collection. Journal of Medicinal Plants Research 6: 2761-2768.

Kohavi, R., and John, G.H. 1997. Wrappers for feature subset selection. Artificial Intelligence 97: 273-324.

Li, G., and Quiros, C.F. 2001. Sequence-related amplified polymorphism (SRAP), a new marker system based on a simple PCR reaction: its application to mapping and gene tagging in *Brassica*. Theoretical and Applied Genetics 103: 455-461.

Maenhout, S., De Baets, B., Haesaert, G., and Van Bockstaele, E. 2007. Support vector machine regression for the prediction of maize hybrid performance. Theoretical and Applied Genetics 115: 1003-1013.

Maenhout, S., De Baets, B., Haesaert, G., and Van Bockstaele, E. 2008. Marker-based screening of maize inbred lines using support vector machine regression. Euphytica 161: 123-131.

Mierswa, I. 2009. Open Source data mining Rapid Miner. KI 23: 62-63.

Mitra, S., and Acharya, T. 2005. Data mining: multimedia, soft computing, and bioinformatics. John Wiley & Sons, 424 pages.

Negbi, M. 2003. Saffron: *Crocus sativus* L. CRC Press, 148 pages.

Ornella, L., Cervigni, G., and Tapia, E. 2012. Applications of Machine Learning in Breeding for Stress Tolerance in Maize. Crop Stress and its Management: Perspectives and Strategies. Springer, pp. 163-192.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on EmpiricalMethods in Natural Language Processing-Volume 10. Association for Computational Linguistics, pp. 79-86.

Steinfath, M., Gärtner, T., Lisec, J., Meyer, R.C., Altmann, T., Willmitzer, L., and Selbig, J. 2010. Prediction of hybrid biomass in *Arabidopsis thaliana* by selected parental SNP and metabolic markers. Theoretical and Applied Genetics 120: 239-247.

Sun, S.-J., Gao, W., Lin, S.-Q., Zhu, J., Xie, B.-G., and Lin, Z.-B. 2006. Analysis of genetic diversity in Ganoderma population with a novel molecular marker SRAP. Applied Microbiology and Biotechnology 72: 537-543.

Tsaftaris, A., Pasentsis, K., Kalivas, A., Michailidou, S., Madesis, P., and Argiriou, A. 2012. Isolation of a CENTRORADIALIS/TERMINAL FLOWER1 homolog in saffron (*Crocus sativus* L.): characterization and expression analysis. Molecular Biology Reports 39: 7899-7910.

Tsaftaris, A., Pasentsis, K., Makris, A., Darzentas, N., Polidoros, A., Kalivas, A., and Argiriou, A. 2011. The study of the E-class SEPALLATA3-like MADS-box genes in wild-type and mutant flowers of cultivated saffron crocus (*Crocus sativus* L.) and its putative progenitors. Journal of Plant Physiology 168: 1675-1684.

Tsaftaris, A., Pasentsis, K.,and Polidoros, A. 2005. Isolation of a differentially spliced C-type flower specific AG-like MADS-box gene from *Crocus sativus* and characterization of its expression. Biologia Plantarum 49: 499-504.

Tsaftaris, A.S., Polidoros, A.N., Pasentsis, K., and Kalivas, A. 2006. Tepal formation and expression pattern of B-class paleo AP3-like MADS-box genes in crocus (*Crocus sativus* L.). Plant Science 170: 238-246.

Tsaftaris, A.S., Polidoros, A.N., Pasentsis, K., and Kalivas, A. 2007. Cloning, structural characterization, and phylogenetic analysis of flower MADS-box genes from crocus (*Crocus sativus* L.). The Scientific World Journal 7: 1047-1062.

Webb, A.R., 2003. Statistical pattern recognition. John Wiley & Sons, 672 pages.

Classification and prediction of three and multi stigma in saffron byStatistical, unsupervised machine learning Tools

Amir H. Beiki[1]*

Received: 13 July, 2014 **Accepted:** 14 September, 2014

Abstract

Saffron is a triploid, sterile plant, used as a spice and medicinalplant in all countries. Stigma is the most important part of saffron. So far no reliable molecular methods were provided to identify and prediction of the three/multi branches species. In this study, using different bioinformatics algorithms, new tools for prediction based on Sequence-Related Amplified Polymorphismmolecular markers is presented. Five alleles M1311400, M151200, M12100 and M10850 selected as the most important classifier by Attribute Weighting models which has the potential to cluster and recognize the three from multi branches stigma. K-Means and K-Medoids unsupervised clustering algorithms were fully able to cluster each genotype to the right classes. Our results showed that for the first time, data mining techniques can be effectively used to genetic differentiation between three and multi stigma with above 90 percent the accuracy andprecision. These methods can use in gene mapping and selection by biomarker.

Keywords: Classifier, Machine learning, Molecular marker, Sequence-Related Amplified Polymorphism

1-Assistant Professor,Department of Biology, Faculty of Science, Qom University.
(*- Corresponding author Email: amirbeiki@gmail.com)