بررسی الگوی بیان ژن‌ها با روش‌های مبتنی بر تجزیه به مقادیر منفرد

The pattern of expression of genes by methods based on Singular value decomposition (SVD)

سجاد طلایی

Talaei.s@arc-ordc.ir

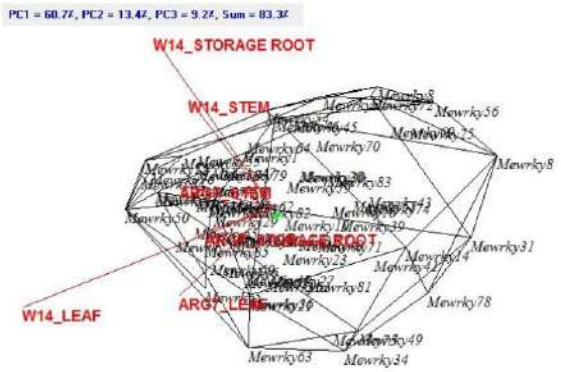
کارشناس ارشد اصلاح نباتات، مرکز تحقیقات کاربردی و تولید بذر، شرکت توسعه کشت دانه‌های روغنی

امروز ترکیب علم کامپیوتر، ریاضیات و آمار راهگشای خیلی از مسائل حوزه زیستی می‌باشد. تجزیه و تحلیل داده‌های مولکولی جدید نیاز به ابزار ریاضی دارد که با مقادیر داده زیاد سازگار بوده، و در عین حال با کاهش پیچیدگی اطلاعات، فهم آنها را تسهیل کند (Alter *et al.*, 2000). با استفاده از روش‌های آماری می‌توان مسائل بیولوژی را دسته بندی، توصیف و قابل فهم کرد. هدف از داده کاوی یافتن اطلاعات جدید از داده‌ها می‌باشد اما حجم و اطلاعات بسیار زیاد این کار را مشکل ساخته است. برای رفع این مشکل می‌توان از روش‌ها مبتنی بر تجزیه به مقادیر منفرد Singular value decomposition (SVD) از جمله تجزیه به مؤلفه‌های اصلی (PCA) استفاده نمود. تجزیه ماتریس دارای کاربردهای زیادی از قبیل تشخیص الگو، کاهش ابعاد، آنالیز بیان ژن و غیره می‌باشد. روش‌های مبتنی بر تجزیه به مقادیر منفرد جهت بررسی گروهی از تیمارهای همبسته مرتبط با یک یا چند حوزه مانند شاخص‌های وضعیت اقتصادی اجتماعی، رضایت شغلی، سلامت، اعتبار شخصی و وضعیت سیاسی، علوم زیستی، کشاورزی و غیره به کار می‌روند. مزیت عمده این روش‌ها کاهش ابعاد اطلاعات می‌باشد تا حدی که بتوان ساختار و توصیف داده‌ها را توجیه کرد. تکنیک‌های ریزآرایه و آنالیز پروتئوم که امکان بررسی همزمان بیان هزاران ژن را فراهم کرده است، حجم عظیمی از اطلاعات بیولوژیک را ایجاد می‌کند (Sturum *et al.*, 2002). این حجم بالای اطلاعات نیاز به دسته بندی ژن‌هایی که همزمان بیان و رونویسی می‌شوند را می‌رساند. تجزیه به مؤلفه‌های اصلی یک الگوریتم آماری است که بر اساس تجزیه به مقادیر منفرد ابعاد داده‌ها را کاهش می‌دهد بطوری که اطلاعاتی که واریانس بیشتری توجیه می‌کنند در مؤلفه‌های اول قرار می‌گیرند. با کاهش تعداد مؤلفه‌ها، هر نمونه می‌تواند به جای هزاران تیمار قرار گیرد. سپس نمونه‌ها را می‌توان بصورت گرافیکی نشان داد تا تفاوت‌ها و شباهت بهتر مشخص شوند. در واقع تجزیه به مقادیر منفرد یک چارچوب ریاضی برای پردازش و مدل‌سازی داده‌های بیانی را فراهم می‌کند (Alter *et al.*, 2000). یکی از مشکلات تجزیه و تحلیل داده‌های بیانی از آنجا ناشی می‌شود که معمولاً تعداد ژن‌ها خیلی بزرگتر از تعداد نمونه در آزمایشات با داده‌های بزرگ است. برای استفاده موثرتر از این داده‌ها باید یک پیش پردازش با کاهش ابعاد اطلاعات روی این داده‌ها انجام گیرد. در خیلی از این موارد از یک روش مبتنی بر تجزیه به مقادیر منفرد برای این کار استفاده می‌شود (Shi and Luo, 2010). از این مدل‌ها برای کاهش ابعاد داده‌های بیان ژن و تشخیص الگوی این داده‌ها با کاهش نویز استفاده می‌شود (Wall *et al.*, 2009). در این مدل‌ها از نمودارهای بای پلات برای نمایش و ترسیم همزمان نقاط و محورها استفاده می‌کنند. هر بعد از یک نمودار بای‌پلات مربوط به یک مؤلفه اصلی

می‌باشد. از روش گرافیکی GGEbiplot (مدلی مبتنی بر تجزیه به مقادیر منفرد) می‌توان برای پلات کردن مقادیر مؤلفه‌ها در برابر هم استفاده نمود که دارای مزایای بسیاری در تفسیر الگوی داده‌ها می‌باشد (Yan, 2002). بین تجزیه به مؤلفه‌های اصلی و مدل GGEbiplot با تجزیه به مقادیر منفرد بطور مستقیم (یا از طریق استاندارد کردن) با یک ماتریس واریانس-کواریانس مرتبط می‌شوند (Wall *et al.*, 2009). تئوری تجزیه به مقادیر منفرد (SVD) بصورت فرمول زیر است.

Anxp= Unxn Snxp VTpxp

که UTU = Inxn و VTV = Ipxp می‌باشند. ستون‌های ماتریس U بردارهای ویژه چپ (بردارهای ردیفی) و S (همان ابعاد ماتریس A را دارا می‌باشد) که حاوی مقادیر ویژه و ماتریس قطری می‌باشد. در ماتریس VT ردیف‌ها به عنوان بردارهای ویژه راست (بردارهای ستونی) در نظر گرفته می‌شوند. بردارهای ویژه PC1 و PC2 مستقیماً از طریق تقسیم‌بندی SVD نمی‌توانند در برابر هم پلات شوند.



نمایش سه مولفه اول الگوی بیان ژن‌ها با مدل GGEbiplot

منبع:

Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences, 97(18), 10101-10106.

Shi, J., & Luo, Z. (2010). Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. Computers in Biology and Medicine, 40(8), 723-732.

Sturn,A., J.Quackenbush, and Z.Trajanoski. 2002. Bioinformatics 18: 207-208.

Wall, M. E., Rechtsteiner, A., & Rocha, L. **M. (2009).** Singular value decomposition and principal component analysis. In A practical approach to microa ray data analysis (pp. 91-109). Springer US.

Yan, W. (2002). Singular-value partitioning in biplot analysis of multienvironment trial data. Agron. J. 94:990-996