Turkish Journal of Agriculture - Food Science and Technology, 8(2): 279-287, 2020
DOI: https://doi.org/10.24925/turjaf.v8i2.279-287.2904

Turkish Journal of Agriculture - Food Science and Technology

Available online, ISSN: 2148-127X | www.agrifoodscience.com | Turkish Science and Technology

Application of Principal Component Analysis for Gene Sequences (cDNA microarrays)

Yalçın Tahtalı[1,a,*], Zeynel Cebeci[2,b]

[1]*Department of Animal Science, Faculty of Agriculture, Tokat Gaziosmanpasa University, 60250 Tokat, Turkey*
[2]*Department of Animal Science, Faculty of Agriculture, Cukurova University, 01330 Adana, Turkey*
[*]*Corresponding author*

ARTICLE INFO

Research Article

Received : 20/08/2019
Accepted : 28/01/2020

Keywords:
cDNA microarrays
Gene expression
Principal Component Analysis
Gene chip
Microarray

ABSTRACT

In this study, principal component analysis has been applied on data comprising of 6675 gene and 20 sequence collected by using cDNA microarray technology from livers of mice used in toxicology studies in certain time periods. Forming of gene groups from similar expression profiles and description of related genes which are implemented by similar component loads among the groups have been explained by using this cDNA technology. Besides that, interpretation and decomposition of factors (components) from correlation matrix which belongs to same data group have been explained. Some of the methods developed for minimizing the data set to fewer components which can explain the whole data structure have been evaluated. According to methods, if we assume that the first 9 eigen values are enough to describe the whole variance, then in this case, it is thought that it is good enough to describe the whole variance by using 9 eigen values with a variance loss of 20,79% instead of describing the whole variance by using 20 eigen values.

Türk Tarım – Gıda Bilim ve Teknoloji Dergisi, 8(2): 279-287, 2020

Gen Dizilerinde (cDNA Mikroarray) Temel Bileşenler Analizinin Uygulanması

MAKALE BİLGİSİ

Araştırma Makalesi

Geliş : 20/08/2019
Kabul : 28/01/2020

Anahtar Kelimeler:
cDNA mikrodizi
Gen ifadesi
Temel bileşenler analizi
Gen çip
Mikroarray

ÖZ

Bu çalışmada, farelerin karaciğerleri üzerine belirli zaman periyotlarında uygulanmış olan, toksikolojik çalışmalardan alınan ve cDNA mikrodizi teknolojisi kullanılarak elde edilen 6675 gen ve 20 dizi içeren verilere temel bileşenler analizi uygulanmıştır. cDNA teknolojisi kullanılarak, birbirine benzer ifade profilleri ile gen gruplarının oluşturulması ve gruplar içerisindeki benzer bileşen (component) yükleri vasıtasıyla birbirleriyle ilişkili genlerin tanımlanması açıklanmıştır. Bunun yanı sıra aynı veri kümesine ait korelasyon matrisinden faktörlerin ayrıştırılması ve yorumu hakkında bilgiler verilmiştir. Kullanılan veri seti içinde, bütün veri yapısın izah edebilecek daha az sayıda bileşene indirgemek için temel bileşen sayısına karar verme yöntemlerinden birkaçı değerlendirilmiştir. Bu yöntemlere göre ilk 9 temel bileşenin bütün yapının varyansını açıklamaya yeterli olduğu düşünülürse bu durumda %20,79 oranında bir varyans kaybı ile 20 temel bileşen yerine 9 temel bileşen ile açıklamanın yeterli olduğu düşünülmektedir.

[a] *yalcin.tahtali@gop.edu.tr* *https://orcid.org/0000-0003-0012-0611* | [b] *zcebeci@cu.edu.tr* *https://orcid.org/0000-0002-7641-7094*

Giriş

Bilimsel çalışmalarda, incelenen olayların genellikle birçok etkenin altında olması ve incelemeye konu olan nesnelerin özelliklerinin birbirleriyle ilişkili olması, çok sayıda değişkenle karşılaşılmasına neden olmaktadır. Tek değişkenli yöntemlerin kısıtlayıcı varsayımlar altında gerçekleştiği düşünülecek olursa bu yöntemlerin yeterli olamayacağı açıktır. Aynı anda birden fazla özelliği incelemek üzere "çok değişkenli istatistik" adı altında birçok yöntem geliştirilmiştir.

İki veya daha fazla bağımlı ya da bağımsız gruplarda çok değişkenli normal dağılıma dayalı hipotezlerin test edilmesinde yararlanılan bir yöntem olan çok değişkenli varyans analizi (Multivariate Analysis of Variance: MANOVA) en sık kullanılan analiz yöntemlerinden biridir. Bunun yanında, kümeleme analizi, faktör analizi, çok boyutlu ölçeklendirme analizi, kanonik korelasyon analizi, çok değişkenli regresyon analizi ve uyum analizi çok değişkenli istatistik analiz olarak sayılabilir (Özdamar, 2002). Çok değişkenli analizlerde, kendilerinden bilgi toplanan deney birimlerinden tespit edilen birçok özelliğin, bir arada ele alınması, özelliklere ait korelasyon veya varyans-kovaryans matrisinin yapısının analizi ile mümkündür. Temel bileşenler analizi orijinal varyans-kovaryans veya korelasyon matrisinin, fazla bir bilgi kaybı olmadan indirgenmesini sağlayarak yeni değişkenlerin belirlenmesi esasına dayanmaktadır (Tatlıdil, 1996). Temel bileşenler analizi, vektörleri varyanslarına göre sıralayan veya başka bir ifadeyle ilk kombinasyon tipi olarak, en büyük varyansı veren kombinasyon seçilmesini sağlayan bir analizdir. Temel bileşenler vasıtasıyla, orijinal değişkenlerden elde edilecek olan sonuçlardan fazla bir bilgi kaybı olmaksızın veri özetlemesi yapılırken, temel bileşenlerin sayısının orijinal değişken sayısından az olması istenir (Cooley ve Lohnes, 1971).

Bu analiz yöntemi, ekonomi, ziraat, kimya, biyoloji, sosyoloji, jeoloji gibi bilimin birçok dalında sıkça kullanılmasının yanı sıra, genomik araştırmalarda da yaygın olarak kullanılmaktadır. Biyolojik sistemler ve olaylar hakkında toplanan bilgilerin analizinde biyoloji yanında biyokimya, kimya ve tıp ile bilişim bilimleri, matematik ve istatistiğin etkileşimiyle yeni ve interdisipliner bir bilim dalı olarak biyoinformatik doğmuştur (Collins ve ark., 2003). Biyoinformatik, geniş veri tabanları arasındaki ilişkileri değerlendirmek için yeni algoritma ve istatistiklerin geliştirilmesi, nükleotid ve amino asit dizilişlerini, protein bölgeleri ve yapılarını kapsayan farklı tipteki verilerin analizi için yeni araçların geliştirilmesi ile ilgilidir (Gardeux ve ark., 2013).

Moleküler genetikte son yıllarda çok hızlı gelişmeler gözlenmektedir. Bu alanda yapılan çalışmaların yoğunlaştığı gen mikrodizi (microarray) teknolojilerindeki ilerlemeler, binlerce genin ifade şekillerinin (biçimlerinin) eş zamanlı olarak izlenebilmesi için güçlü araçların geliştirilmesine olanak sağlamıştır (Jordan, 2001).

Gen ifade dizileri üzerine yapılan çalışmalara paralel olarak, binlerce genin, farklı deneme ya da örneklerden elde edilen sonuçlarını analiz etmek amacıyla geliştirilen birkaç farklı mikrodizi sistemi bulunmakla beraber, Lipshutz ve ark. (1999)'a göre mikrodizileri iki grup halinde sınıflandırmak mümkündür. Bunlar; cDNA mikrodiziler ve oligonükleotid mikrodizilerdir (DNA çip'leri). cDNA mikrodiziler de kendi aralarında iki gruba ayrılmaktadır. Bunlar, zar üzerine sentezlenen cDNA mikrodizi'ler ve cam üzerine sentezlenen cDNA mikrodizi'lerdir (Cox, 2001; Barash ve ark., 2004).

Genetik analiz çalışmalarında elde edilen veriler ve/veya gözlemleri etkileyen çok sayıda varyasyon kaynağının olması, çalışılan boyut sayısının yüksek olması ve az sayıda tekerrür gibi sorunlar nedeniyle çeşitli istatistiksel analiz güçlükleriyle yüz yüze kalınmaktadır (Kerr ve Chuchill, 2001). Bu nedenle bu tür problemlerin çözümüne yönelik yeni istatistiksel yöntemlerin geliştirilmesi ve bu yöntemlerin gen ifade dizi verileri üzerinde yapılan çalışmalardan elde edilen verilere uygulanması büyük önem taşımaktadır.

Bu çalışmanın amacı, temel bileşenler analizinin matematiksel metodunu temel alarak, binlerce geni aynı anda izleme olanağı veren, mikrodizi teknolojisi ile elde edilen cDNA mikrodizi (cDNA microarray) verilerine temel bileşenler analizi uygulamak, değişkenler arasındaki bağımlılık yapısını yok ederek binlerce gen verisi içerisinde bu yapıyı açıklayabilecek daha az sayıda gen için veri indirgemesi yapmak ve daha sonra yapılacak istatistik analizlere veri hazırlamak amaçlanmıştır.

Materyal ve Metot

Materyal

Bu çalışmanın materyalini, Germantown, NewYork'ta Taconic Laboratuarı'nda Heinloth ve arkadaşları tarafından 2004 yılında, erkek fareler üzerine yapılan toksikolojik çalışmalar oluşturmuştur. 6675 gen ve 20 dizi (array) ile yapılan çalışmada, 36±3 günlük yaştaki 344 adet fare alınmış ve 89±3 günlük yaşa kadar 21.6ºC (71ºF) ile 23,8ºC (75ºF) sıcaklıkta ve %36 ile %48 nem ortamında toksik doz uygulaması yapılmıştır. Gen ifadesi analizi için, farelerin karaciğerlerinden alınan parçalara ait genler belirlenmiştir. PCR ve cDNA mikrodizi teknolojisi kullanılarak genlere ait mikrodizi görüntüsü elde edilmiş ve Scanalytics yazılımı ile görüntü analizi yapılmış, floresan yoğunlukları ölçülen ve logaritmik değerleri alınan veriler normalizasyona tabi tutularak veriler elde edilmiştir (Heinloth ve ark., 2004).

Metot

Bu çalışmada, birbirine benzer ifade profillerinden gen gruplarının oluşturulması ve gruplar içerisindeki benzer bileşen (component) yükleri vasıtasıyla birbirleriyle ilişkili genlerin tanımlanması amacıyla veri kümesine ait korelasyon matrisinden faktörlerin ayrıştırılması ve indirgenmesinin yorumu izah edilmiştir. cDNA mikrodizi (cDNA microarray) teknolojisi ile bir organizma ya da bir hücredeki binlerce genden elde edilen, mikrodizi görüntülerine, görüntü işleme analizi uygulanarak görüntü sayısal değerlere dönüştürülmüştür. Deneme materyaline, veri indirgemek ve başka istatistik analizlere veri hazırlamak amacıyla çok değişkenli istatistik analiz tekniklerinden temel bileşenler analizi uygulanarak boyut ölçeklendirmesi (S-PLUS 2000 ve MINITAB 13.0) yapılmıştır.

***cDNA Mikrodizi (cDNA Microarray*)**

İşlevsel genomik çalışmalar, insan ve model organizmalara ait genlerin baz dizisini taşıyan tam uzunlukta cDNA yapılarının belirlenmesini, protein kodlamayan dizilerin işlevlerinin araştırılmasını, gen ifadesi ve protein analizlerini, ayrıca bu konularda yeni ve daha hızlı teknolojilerin geliştirilmesini kapsamaktadır (Shoemaker ve ark., 2001).

cDNA mikrodizi (cDNA microarray), tek bir dizi (array) üzerinde tüm genomu inceleyebilme özelliğindedir. Araştırıcılara aynı anda yüzlerce farklı koşullar altında binlerce genin ifade düzeylerini ve birbirleriyle olan ilişkisini öğrenebilme olanağını sağlamaktadır (Eisen ve ark., 1998). Bu teknolojiyi tanımlamak üzere çeşitli kaynaklarda, DNA yongası (DNA chip), DNA mikrodizi (DNA microarray), gen dizisi (gene-array), biyo yonga (bio chip) gibi değişik terimler kullanılabilmektedir (Brown ve Bolstein, 1999; Kamberova ve Shah, 2002).

Mikrodizi teknolojisinin kullanıldığı belli başlı alanlar; kompleks genetik hastalıkların araştırılması, ilaç geliştirme ve toksikoloji çalışmaları, genetik testler, hastalıklara, böceklere ve kuraklığa dayanıklı tohumların geliştirilmesi, sağlıklı ve daha verimli, hastalıklara dirençli çiftlik hayvanlarının geliştirilmesi, besin değeri yüksek ürünlerin geliştirilmesi olarak belirtmek mümkündür (Brown ve ark., 2000; Gardeux ve ark., 2013).

Bu tekniğin temelinde bir cam lam veya naylon zar yüzeyi üzerinde birbirine çok yakın olarak dizilmiş kısa ve RNA sentezinde kullanılan yüzlerce/binlerce DNA hedef dizilimlerinin (3'expressed sequence tags) yerleştirilmesi ve incelenecek olan cDNA'nın floresan/radyoaktif işaretlemenin ardından yerleştirilmiş DNA ile bir araya getirilerek işleme alınması bulunmaktadır (Schena, 2001).

Gen ifade dizileri için öncelikli olarak hücrelere ait mRNA'lardan cDNA'lar elde edilmekte ve bunlar floresan tekniği kullanılarak birisi yeşil (örneğin sağlam bir hücreden alınan mRNA örneği) "Cy5" ile, diğeri kırmızı (hastalıklı hücreden alınan mRNA örneği) "Cy3" ile boyanmaktadır. Bunların birlikte karışımı probları belli olan ve DNA nükleotidleri yerleştirilen cam ya da naylon membran üzerine uygulanmaktadır (Şekil 1). Bu işlem sonucunda beneklerden (spot) oluşan ve her birinin bir geni ifade ettiği cDNA mikrodizileri elde edilmektedir (Brown ve ark., 2000; Leung ve Cavalieri, 2003; Gonzalo ve Sanchez., 2018).

Beneklerde oluşan parlak kırmızı renk, hastalıklı hücreden alınan örneğin, parlak yeşil renk ise sağlam hücreden alınan örneğin ifade düzeyinin yüksek olduğunu göstermektedir. Kırmızı ve yeşil rengin tonları ise yine sırasıyla hastalıklı hücrelerin farklı ifade düzeylerini ve sağlam hücrelerin farklı ifade düzeylerini göstermektedir. Denemelerde, oluşan gri renk tonları, kayıp veya ölçülememiş değerleri ifade etmektedir (Draghici, 2003; Ocampo ve ark., 2016).

Hücre gruplarına ait cDNA örnekleri bir plaka üzerine yerleştirilmekte ve mikrodizi olarak adlandırılan plaka, karanlık bir ortamda lazer taramasından geçirilerek renk sinyalleri kaydedilmektedir (Shao ve ark., 2019). Elde edilen değerler (renk yoğunlukları) bir veri tabanında tutulmaktadır. Daha sonra her bir örneğin her bir genine ait renk yoğunluk verilerinin logaritmaları alınıp, değerler dönüştürme işlemine tabi tutulduktan sonra bir veri matrisine atanarak, istatistiksel sonuç çıkarılmaktadır (Schuchhardt ve ark., 2000; Taguchi, 2017). Genellikle veri bankalarından alınan cDNA mikrodizi verileri, logaritması alınmış değerler olup;

$$y_{ij} = \log_2(x_{ij})$$

şeklinde hesaplanır ve renk yoğunlukları dikkate alınarak $x_{ij}=Cy5_{ij}/_{Cy}3_{ij}$ şeklinde bulunur.

Mikrodizi görüntülerinin temel yapısı satır ve sütunlardan oluşan ve beneklerin oluşturduğu ızgaraları içermektedir. Bu nedenle beneklerin adreslenmesi için bazı parametrelerin belirlenmesi gerekir. Bunlar; Izgaralardaki satır ve sütunlar arasındaki ayrışmaların tespiti, her bir ızgaradaki satır ve sütunlardaki benekler arasındaki ayrışmaların belirlenmesi, beneklerin özel olarak işaretlenmesi ve görüntülerdeki dizilerin belirlenmesi olarak sıralanabilir (Draghici, 2003; Shao ve ark., 2019).

Her gen için ifade farklılığının ortaya konulmasında, dizi üzerinde oluşan bir beneğin gerçek floresan yoğunluğunun ve arka plan yoğunluğunun miktarlarının ve özelliklerinin doğru olarak tespit edilmesi gerekmektedir (Newton ve ark., 2001). Bu benekler tespit edilirken;

$\bar{r}_i^m$ = ortalama yoğunluğu kırmızı ile boyanmışlar,

$\bar{g}_i^m$ = yeşil ile boyanmışlar,

$\bar{r}_i$ ve $\bar{g}_i$ = floresan tekniği ile cDNA'ların hibridizasyonundan meydana gelen beneklerin ortalama yoğunluğunu,

$\bar{r}_i^b$ ve $\bar{g}_i^b$ = cam veya naylon membran üzerinde meydana gelen ancak önem arz etmeyen beneklerin ortalama yoğunluğunu,

r_i^e ve g_i^e = kullanılan elektronik teknik ya da kullanılan materyalden kaynaklanan hatadan oluşan benekleri göstermektedir.

Burada i. benek için:

$$\bar{r}_i^m = \bar{r}_i + \bar{r}_i^b \pm r_i^e,\ \bar{g}_i^m = \bar{g}_i + \bar{g}_i^b \pm g_i^e$$

şeklinde gösterilebilir. Burada,

$$\bar{r}_i^m = \frac{1}{j_i}\sum_{k=1}^{j_i} r_{ik}^m$$

j, ölçüm sonucu elde edilen her bir benekteki piksel sayısıdır. i. genin ifade oranını Z ile gösterdiğimizde,

$$Z_i = c\,\frac{\bar{r}_i^m - \bar{r}_i^b \pm r_i^e}{\bar{g}_i^m - \bar{g}_i^b \pm g_i^e}$$ elde edilir.

$\bar{r}_i^m$ ve $\bar{g}_i^m$ için dağılış, yaklaşık normal dağılış olup i. genin ifade oranına ait dağılış,

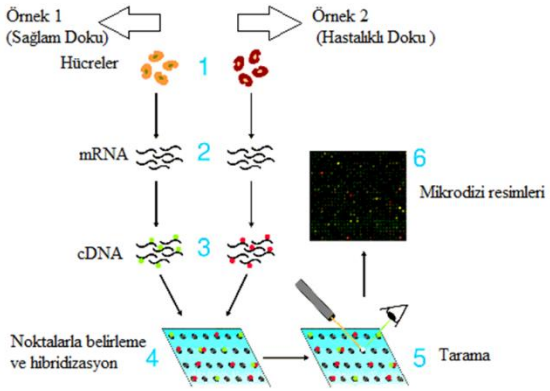
$$\sigma_{Z_i}^2 \approx c\left[\sigma_{g_i}^2 \frac{\bar{r}_i^{m^2}}{\bar{g}_i^{m^4}} + \frac{\sigma_{r_i}^2}{\bar{g}_i^{m^2}} - 2\sigma_{rg_i}\frac{\bar{r}_i^m}{\bar{g}_i^{m^3}}\right]$$

şeklinde gösterilir.

Burada

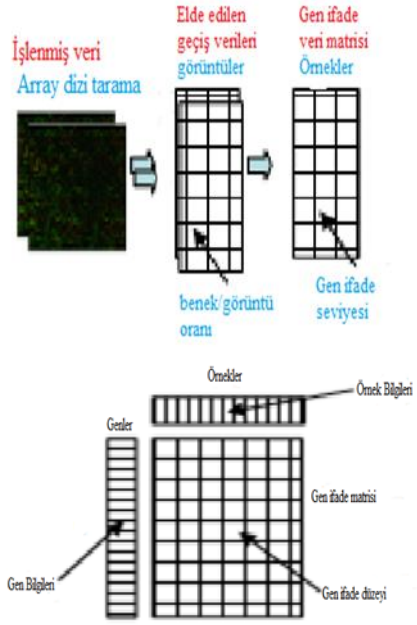
σ_{ri} ve σ_{gi}, $\bar{r}_i^m$ ve $\bar{g}_i^m$ ile ilgili standart sapmayı,

σ_{rg} kırmızı renk ve yeşil renk kovaryansını ifade etmektedir. *c* ise kırmızı ve yeşil bölgeleri elde ederken oluşan farklılıkları düzelten bir sabittir (Brown ve ark., 2000).



Şekil 1. cDNA Mikrodizilerinin elde edilmesi (Herrero ve ark., 2003)

Figure 1. Obtaining cDNA microarrays



Şekil 2. Ham verilerin gen ifade veri matrisi haline getirilmesi

Figure 2. Making raw data into gene expression data matrix

Mikrodizi Görüntülerinin Veri Matrisine Aktarılması

Mikrodizi denemelerinde elde edilen görüntü ham veri olup, bu görüntülerdeki floresan renk yoğunlukları sayısal değerlere dönüştürüldükten sonra logaritmik dönüşümü yapılmaktadır. Şekil 2'de, görülebileceği gibi gen ifade matrisinde, sütunlar farklı birey ve dokuları veya farklı zamanlarda aynı birey veya dokulardan alınmış örnekleri, satırlar ise genleri ve genlerin bütün dizilerdeki ölçümlerini göstermektedir (Lin ve Johnson, 2002; Causton ve ark.,2003).

Birçok deneme sonunda aynı çalışmaya ait denemelerden elde edilen tekrarlanmış cDNA mikrodizileri bir araya getirilerek, deneme hatası minumum seviyeye indirilmeye çalışılır. Satır ve sütunların kesiştiği nokta olan beneklerin, ortalama, medyan, piksel yoğunluğu ve her bir beneğin arka plan değerleri olmak üzere bazı nicel özellikleri vardır. Tekrarlanan denemeler kombine edilerek gen ifade matrisleri oluşturulur.

Temel bileşenler analizi, bu şekildeki değişkenler setinin varyans - kovaryans yapısını, değişkenlerin doğrusal birleşimleri vasıtasıyla açıklayarak, veri indirgenmesi ve yorumlanmasını sağlamaktadır. Yani mümkün olduğu kadar az sayıdaki yeni değişkenin, orijinal değişken yerini alacak şekilde oluşturmaktadır. Bu yöntemde, karşılıklı bağımlılık yapısı gösteren, ölçüm sayısı *n* olan *p* adet değişken; doğrusal, ortogonal ve birbirinden bağımsız olma özelliklerini taşıyan k(k≤p) tane yeni değişkene dönüştürülmektedir.

Toplam değişkenliğin önemli bir kısmı k(k≤p) bileşen tarafından açıklanabildiği durumlarda, *k* bileşen orijinal *p* değişkeni temsil edebilmektedir (Peterson, 2003; Raychaudhuri ve ark., 2000).

Matematiksel olarak temel bileşenler

$X_1, X_2, ..., X_p$

gibi p tane değişkenin doğrusal birleşimleridir. Temel bileşenler,

$X' = [\, x_1, x_2, .., x_p \,]$

şansa bağlı vektörü ve bu vektörün kovaryans matrisi Σ, öz değerleri de $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$ ise ℓp elemanlı sütun vektör olmak üzere, değişken sayısı artırıldığında,

$X_1, X_2, \ldots X_p$

gibi *p* tane tesadüfi değişkenine sahip X vektörünün lineer dönüşümlerini,

$$\begin{aligned} y_1 &= \ell`_1 X = \ell_{11}X_1 + \ell_{21}X_2 + ... + \ell_{p1}X_p \\ y_2 &= \ell`_2 X = \ell_{12}X_1 + \ell_{22}X_2 + ... + \ell_{p2}X_p \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ y_p &= \ell`_p X = \ell_{1p}X_1 + \ell_{2p}X_2 + ... + \ell_{pp}X_p \end{aligned} \qquad (1)$$

şeklinde yazmak mümkündür. Veya matris formunda,

$$Y = \ell' X \qquad (2)$$

şeklinde yazılabilir. Buradaki amaç, X vektörünün doğrusal dönüşümü olan (1) şansa bağlı değişkenlerini $\ell'\ell = 1$ kısıtlaması altında maksimum varyansa sahip olacak şekilde bulmaktadır. Burada ℓ'nin her bir sütunu bir tek temel bileşen katsayılarını göstermektedir. (2) eşitliğinin varyansı

$$E(\ell' X)^2 = E(\ell' X X' \ell) = \ell' \sum \ell \qquad (3)$$

olmak üzere Lagrange fonksiyonu

$$\Phi(\ell, \lambda) = \ell' \sum \ell - \lambda(\ell'\ell - 1) \qquad (4)$$

olsun.

Burada λ Lagrange çarpanı ve $\ell'\ell = 1$ dir. ℓ ye göre kısmi türevi alınıp sıfıra eşitlenirse:

$$\frac{\partial \Phi}{\partial \ell} = 2\Sigma\ell - 2\lambda\ell = 0 \quad (5)$$

sonucuna ulaşır. Buradan,

$$(\Sigma - \lambda I)\ell = 0 \quad (6)$$

elde edilir. Eşitlik (6) $\ell'\ell = 1$ kısıtlaması altında ℓ=0'dan başka çözümünün olması için

$$|\Sigma - \lambda I| = 0 \quad (7)$$

olmalıdır. Eşitlik (7), λ'nın *p*. dereceden bir polinomudur. Dolayısıyla (7) ifadesinin *p* tane kökü vardır ve bu kökler $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ dır. p tane öz değer kullanılarak her birine karşılık gelen *p* tane öz vektör elde edilir. (7) eşitliğini soldan ℓ' ile çarparsak:

$$\ell'\Sigma\ell = \lambda\ell'\ell = \lambda \quad (8)$$

elde edilir. Buradan $y_1 = \ell_1' X$ ifadesinin varyansı λ_1 olur. Böylece maksimum varyans için;

λ_1 kökünü kullanacak olursak,

ℓ_1, $(\Sigma - \lambda_1 I)\ell = 0$'ın

λ_1'e karşılık gelen bir çözümü olduğundan

$y_1 = \ell_1' X$ maksimum varyanslı lineer bir bileşen olur (Anderson,1974).

Eşitlik (3)'den faydalanarak,

$$Var(y_i) = \ell_i' \Sigma \ell_i \text{, i=1,2,…,p} \quad (9)$$

$$Kov(y_i) = \ell_i' \Sigma \ell_k \quad \text{i,k = 1,2, …,p} \quad (10)$$

eşitliklerini yazabiliriz. Temel bileşenler varyansları mümkün olduğu kadar büyük olan ve birbirleri ile ilişkili olmayan $y_1, y_2, \ldots, y_p$ olduğundan ilk temel bileşen olan y_1 maksimum varyanslı doğrusal bileşendir.

Yani

$Var(y_1) = \ell_1' \Sigma \ell_1$

maksimum değere sahip doğrusal bileşendir. Sonuç olarak,

Σ, $X' = \left[X_1, X_2, \ldots, X_p\right]$

şansa bağlı vektörünün kovaryans matrisi ve $(\lambda_1, e_1), (\lambda_2, e_2), \ldots (\lambda_p, e_p)$ özdeğer ve özvektör ikililerine sahip olsun,

Burada, $\lambda_1 \geq \lambda_2, \ldots \geq \lambda_p \geq 0$ *i*.temel bileşen

$$Var(y_i) = e_i' \Sigma e_i = \lambda_i \quad (11)$$

$$Kov(y_i, y_k) = e_i' \Sigma e_k = 0 \qquad \text{i≠k} \quad (12)$$

eşitlikleri yardımı ile

$$y_i = e_i' X = e_{1i} X_1 + e_{2i} X_2 + \ldots + e_{pi} X_p \quad (13)$$

şeklinde yazabiliriz. Eğer bazı λ_i'ler eşit ise karşılık gelen katsayılar vektörleri olan e_i'lerin seçimi y_i'den dolayı tek değildir (Johnson, 1982).

Yukarıda da belirtildiği gibi temel bileşenler analizinde bir bileşenin uzunluğunu bir özdeğer, yönünü ise bu özdeğere bağlı olarak bulunan bir özvektör tayin etmektedir. Dolayısıyla temel bileşenler analizinde en önemli nokta, bir çok özellikten oluşan bir tesadüf vektörünün varyans-kovaryans veya korelasyon matrisinden, özdeğer ve özvektör çiftlerinin elde edilmesidir.

Toplam populasyon varyansı

(σ_T), $\sigma_T = \sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp}$

$= \lambda_1 + \lambda_2 + \ldots + \lambda_p$

şeklinde olup, k. temel bileşenin toplam varyansa oranı,

$$= \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}, \text{ k=1,2,…,p}$$

şeklindedir.

Ayrıca temel bileşenler analizinin uygulanıp, uygulanamayacağına küresellik testi ile karar verilir. Küresellik testi olarak bilinen test için; H_0: ∑=I;H_1: ∑≠I hipotezleri kullanılmakta ve H_0 hipotezinin red edilmesi durumunda temel bileşenler analizinin kullanılması önerilmektedir (Johnson ve Wichern, 1982).

Eğer toplam varyansın büyük bir kısmı (%80, %90), p'nin büyük olması durumunda, birinci, ikinci veya üçüncü temel bileşen diye adlandırılan temel bileşenler tarafından açıklanıyorsa bu temel bileşenler çok fazla bilgi kaybı olmaksızın orijinal *p* değişkeninin yerini alabilmektedir (Ocampo ve ark., 2016).

Temel bileşenlerin hesaplanması sırasında öz değerlerin bulunmasından sonra, öz değer sayısına karar vermek önemlidir. Bu amaçla birçok yöntem vardır.

Bunlardan birisi, varyans-kovaryans matrisinden elde edilen öz değerlerin ortalamasının alınmasını ve bu ortalama değerden daha büyük olan öz değerlerin modele dahil edilmesini teklif etmiştir. Ayrıca, temel bileşen sayısının belirlenmesinde, korelasyon matrisinin, birden büyük öz değerlerinin dikkate alınması ve bu öz değerlere karşılık gelen bileşenlerin modele dahil edilmesi ise bir başka kriterdir. Cattle (1966), tarafından önerilen başka bir yöntem de scree grafik testidir (Jollifee, 2002). Bu yöntem de Y eksenindeki öz değerler, X ekseninde ait oldukları öz değerin sayısına göre eğrinin ilk kırılım noktasının belirlenerek temel bileşen sayısına karar verilmesidir. Uygulamada kullanılan bir başka kriter, incelenen değişkenlere ait toplam varyansın, %75-80 oranında açıklanabilmesidir. Genlere ait matris yapısı ise şu şekilde tanımlanır; *p* diziye ait bir mikrodizi denemesinde, *p* dizi sayısını, *n* ise bir diziye ait genleri ifade etmektedir. *Y* veri matrisinin örnek büyüklüğünü genler ve bu genlerin yer aldığı deneme veya örnekler oluşturur (n(gen) × p(dizi)). Matrisin her bir elemanı y_{ij} ile gösterilecek olursa, i. satır

ve j. sütunda 1≤ i ≤n; 1≤ j ≤p aralığında yer alan gen ifade değerleri olup *Y* matrisini aşağıdaki şekilde göstermek mümkündür (Peterson, 2001).

$$Y_{nxp} = \begin{bmatrix} y_{11} & y_{12} & \cdots y_{1p} \\ y_{21} & y_{22} & \cdots y_{2p} \\ . & . & . \\ . & . & . \\ y_{n1} & y_{n2} & y_{np} \end{bmatrix}$$

Burada *Y*'nin her elemanı, y_{ij}, i. genin ve j. dizi üzerindeki ifade oranın logaritmasını (LER) tanımlamakta olup aşağıdaki şekilde göstermek mümkündür. LER (ifade oranının logaritması)

$\log_2\left(\frac{I_{ij,\,(KIRMIZI)}}{I_{ij,(YESIL)}}\right)$ dir. Burada, $I_{ij,KIRMIZI}$ ve $I_{ij,YESIL}$

kırmızı ve yeşil lazer ile taraması sonucunda, kırmızı ve yeşil sinyal yoğunluklarını göstermektedir (Peterson, 2003). Daha açık bir ifade ile, Y matrisinin i. satır vektörü, farklı örneklerden ya da farklı denemelerden, dizilerden alınmış belli bir gen için floresan yoğunluk oranlarını göstermekteyken j.sütun vektör ise j.dizi ya da örneğe ait farklı genlerin floresan yoğunluklarını göstermektedir. Burada, y_{ij}, pozitif ise, kontrol grubu ile ilişkili olarak j. deneme ya da örnekteki i. genin yüksek ifade oranına sahip olduğu; negatif ise, j. deneme ya da örnekteki i. genin düşük ifade oranına sahip olduğu anlaşılmaktadır. *k* ve *l* gibi iki gen ele alındığında, bu genler arasındaki ikili korelasyon ise

$$r_{kl} = \frac{\sum_{m=1}^{p}((z_{km} - \bar{z}_k)/s_k)((z_{lm} - \bar{z}_l)/s_l}{p}$$

olarak bulunur (Raychaurhuri ve ark., 2000). Buradan elde edilecek sonuçlar ile *nxn* (*gen x gen*) korelasyon matrisi ise;

$$R_{nxn} = \begin{bmatrix} r_{11} & r_{12} & \ldots r_{1n} \\ r_{21} & r_{22} & \ldots r_{2n} \\ . & . & . \\ . & . & . \\ r_{n1} & r_{n2} & r_{nn} \end{bmatrix}$$ şeklinde gösterilebilir.

Bulgular ve Tartışma

Bu çalışmada, Germantown, NewYork'ta Taconic Laboratuarı'nda 2004 yılında, erkek fareler üzerine yapılan toksikolojik çalışmalardan alınmış 6675 gen ve 20 diziye ait cDNA mikrodizi (mikroarray) verileri kullanılmıştır. Genlere ait mikrodizi görüntüsü, görüntü işleme yazılımları yardımı ile yeşil ve kırmızı floresan boya değerlerinin oranlarına göre sayısal değerlere dönüştürülmüş, daha sonra elde edilen değerler logaritmik dönüşüme tabi tutulmuş ve kullanılan sayısal veriler elde edilmiştir. Elde edilen veriler standartlaştırılmış veri matris olduğundan ayrıca dizilerin farklı zamanlarda yapılmış denemelere ait olabileceği PCR uygulamalarında, bir standarda uyulmamış olunabileceğinden dolayı korelasyon matrisinden yararlanılmıştır. Korelasyon matrisinden hesaplanan öz değerler Tablo 1'de verilmiştir. Tablo 1'den de görüleceği gibi,

$$\sum_{i=1}^{20}\lambda_i = 4{,}53336 + 3{,}25776 + 1{,}86271 + 1{,}33514 + 1{,}17722 + 1{,}05621 + \ldots + 0{,}13465 = 20$$

dir ve

$$tr(R) = \sum_{i=1}^{20}\lambda_i = 20$$

dir. Ayrıca,

$$|R| = \prod_{i=1}^{20}\lambda_i = 0{,}000101$$

olarak bulunmuştur.

Bulunan bu değerler, verilerin temel bileşenler analizi için uygun olup olmadığını belirlemek üzere küresellik testinde kullanılmıştır. Anderson (1974) tarafından önerilen küresellik testi için kurulan hipotez testine göre, H_0: R=I, H_1: R≠I olup analiz sonucunda H_0 hipotezi red edilmiştir. Yani temel bileşenler analizinin uygulanması uygundur. Ayrıca bileşenler tarafından açıklanan varyans oranları Tablo 2'de verilmiştir.

Birinci bileşen, veri matrisindeki değişkenliğin %22,94'ünü açıklar. Bu çalışmada kullanılan deneme materyaline göre ilk 9 öz değer, veri matrisindeki değişkenliğin toplam olarak %79,21'ini açıklamaktadır. Bu değer, temel bileşen sayısına karar verilirken göz önünde bulundurulacak kriterlerden biri olan, Jollifee (2002)'nin bildirdiği, bileşen sayısına karar verilecek kümülatif varyansın %75-80 arasında olması kriterine uymaktadır.

Ayrıca %75-80'lik varyanstan sonra yapıya girecek bir bileşen varyansı açıklama oranına %5'den daha az oranda katkıda bulunuyorsa bu temel bileşenlerin önemsiz kabul edileceği çeşitli kaynaklarda yer almaktadır. Bilgi kaybı amaca göre kabul edilebilir düzeyde ise daha az sayıda bileşenle analize devam edilebilmektedir. İlk 9 temel bileşenin varyasyonun tamamını izah etmeye yeterli olduğu düşünülürse bu durumda %20,79 bir varyans kaybı ile 20 temel bileşen yerine 9 temel bileşen ile çalışmanın yeterli olduğu düşünülür. cDNA microarray verileri ile ilgili olarak literatürlerde benzer çalışmalar yapılmıştır. Wagner ve ark. (2015), farelerin hematopoetik hücrelerinden elde edilen cDNA doku örnekleri ile yaptığı çalışmada, temel bileşenler analizi ile 4 bileşen ve %79,56 izah oranı ile ifade edebileceğini bildirmiştir. Ocampo ve ark. (2016) yaptıkları çalışmada, Leukemia genlerini kullanarak 7,129 gen ve 72 örneğe temel bileşenler analizini uygulamışlar, bu verileri 4 temel bileşen ve %88,24 açıklama oranıyla izah edebileceklerini bildirmişlerdir. Taguchi (2017), 17,112 Dengue geni ile yaptığı çalışmada, 3 temel bileşen ile %89,5 izah elde etmiştir. Ayrıca sınıflama metodlarını uygulayarak genleri birbirleriyle ilişkili iki sınıfa ayırmıştır. Rao ve ark. (2019) fareler ile yaptıkları toksikogenomik çalışmada, 32,633 gen ve 26 dizi için, 10 temel bileşen ile veri yapısını açıklayabildiklerini bildirmişlerdir. Bu sonuçlar, yaptığımız çalışma ile benzerlik göstermektedir.

Tablo 1 Yirmi adet diziye ait öz değerler tablosu
Table 1 Eigenvalues table of twenty arrays

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
4,58886	3,25776	1,86271	1,33514	1,17722	1,05621	0,90364	0,86766	0,79236	0,68415
λ_{11}	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}	λ_{17}	λ_{18}	λ_{19}	λ_{20}
0,65972	0,58719	0,55955	0,47211	0,28422	0,25829	0,19176	0,17294	0,15385	0,13465

Tablo 2. Açıklanan varyans payları ve birikimli paylar
Table 2. Explained variance shares and cumulative shares

Fare gen no	Öz değerler (λ)	Açıklanan Varyans Payları (%)	Birikimli açıklanan Pay (%)
Fare.3125	4,588	22,94	22,94
Fare.3126	3,257	16,29	39,23
Fare.3127	1,862	9,31	48,54
Fare.3149	1,335	6,67	55,21
Fare.3151	1,177	5,90	61,11
Fare.3083	1,056	5,28	66,39
Fare.3069	0,903	4,52	70,91
Fare.3070	0,867	4,34	75,25
Fare.3074	0,792	3,96	79,21
Fare.3057	0,684	3,42	82,63
Fare.3058	0,659	3,29	85,92
Fare.3030	0,587	2,96	88,87
Fare.3031	0,559	2,80	91,67
Fare.3032	0,472	2,36	94,03
Fare.3018	0,284	1,42	95,45
Fare.3019	0,258	1,29	96,74
Fare.3006	0,191	0,96	97,70
Fare.3007	0,172	0,86	98,56
Fare.3008	0,153	0,77	99,33
Fare.3026	0,134	0,67	100,000

Tablo 3. İlk yedi bileşen ve ilk 10 fare için faktör yapısı
Table 3. Factor structure for the first seven components and the first 10 mice

	Faktör 1	Faktör 2	Faktör 3	Faktör 4	Faktör 5	Faktör 6	Faktör 7
Fare.3125	-0,0734	0,3066	-0,0754	0,1780	-0,0534	-0,7364	-0,2524
Fare.3126	-0,2106	-0,3221	-0,2110	-0,2804	0,2933	0,0043	-0,1109
Fare.3127	-0,2625	0,0656	-0,4322	0,2885	-0,0270	0,1118	-0,1305
Fare.3149	-0,1287	-0,1589	0,0183	0,5362	0,1596	0,4813	-0,2404
Fare.3151	-0,2996	0,0039	-0,2850	0,5278	-0,0122	-0,1319	0,0494
Fare.3083	-0,0063	0,5263	0,0427	0,0664	-0,4241	0,2509	0,2559
Fare.3069	-0,2864	-0,2621	-0,4843	-0,2353	-0,0734	-0,0949	-0,1040
Fare.3070	-0,0523	0,1204	-0,4256	-0,3381	-0,2548	0,2499	0,0828
Fare.3074	0,0535	0,5718	-0,2726	-0,0079	-0,3026	0,0680	-0,1777
Fare.3057	0,2362	0,0608	-0,1300	-0,1270	0,5440	0,1709	-0,2477
Fare.3058	-0,0347	-0,1265	-0,3284	0,1539	0,3048	-0,1095	0,7303
Fare.3030	-0,0872	0,8179	0,0052	-0,0702	0,2070	0,0865	0,0846
Fare.3031	0,1207	0,8367	0,0666	-0,0198	0,1763	-0,0448	0,1250
Fare.3032	-0,1087	0,8032	-0,0181	-0,0311	0,2640	-0,0071	0,0465
Fare.3018	-0,8375	0,1867	-0,0290	-0,0642	0,0101	0,0024	-0,1754
Fare.3019	-0,7860	0,4121	0,0951	-0,0015	0,0106	0,0347	-0,1143
Fare.3006	-0,8621	-0,1518	0,1276	0,0129	-0,0063	0,0183	0,0727
Fare.3007	-0,8244	-0,1805	0,1404	-0,0145	-0,0558	0,0331	0,1476
Fare.3008	-0,8077	-0,2057	0,1169	-0,0692	0,0080	-0,0548	0,1858
Fare.3026	-0,8889	0,0788	0,0136	-0,1002	0,0831	0,0240	-0,0812

Golub (1999), hematolojik malignansilerle yapılan bir çalışmada 6817 gen bulunduran bir dizi üzerinden seçilen 50 gen analiz edilmiş ve lösemiler ALL ve AML olarak sınıflandırılmıştır. Raychaudhuri ve ark. (2000), *Saccharomyces cerevisiae*'dan alınan 6118 gene ait ifade oranlarını içeren değerleri kullanmışlarıdır. Cattle'in scree grafik yöntemini kullanarak, yedi temel bileşenden ilk iki temel bileşenin, toplam varyansın %90,4'lük kısmını izah ettiğini, bunun da diğer bileşenleri temsil edebileceğini göstermişlerdir.

Landgrebe ve ark (2002), fare beyni gen ifade profillerinin kullanımında antidepresan etkilerin moleküler karakterizasyonu üzerine bir çalışma yapmışlardır. Dizi ve genlere ait ifade oranlarıyla yapılan analizde seçilmiş olan iki temel bileşenin, toplam varyansın %72,2'sini izah edebildiğini göstermişlerdir.

Temel bileşenlerin yapılarının daha iyi tanımlanabilmesi için faktörler ile değişkenler arasındaki korelasyon katsayılarını veren faktör yapısını incelemek gerekir. Bu değerler temel bileşenin, hangi faktörlerin, hangi oranda katkıda bulunduklarını göstermektedir. İlk yedi temel bileşene ait faktör yapısı Tablo 3'de verilmiştir.

Tablo 3. incelendiğinde görüleceği gibi birinci temel bileşen ile farklı fare gruplarının özetlenmesinde çok fazla bilgi kaybına yol açmıştır. Ayrıca bazı katsayıların eksi işaretli olması, bu tip faktörlerin iki kutuplu faktör olduğunu göstermektedir.

Tablo 3. incelendiğinde, yedi faktör yapısı içinde, birincisinde, en yüksek korelasyona sahip değişkenler; 3026 nolu fare, 3006, 3008 ve 3018 nolu farelerdir. Yani birinci bileşen birinci faktörün ve ağırlıklı olarak ta bu farelere ait genlerin katkılarıyla oluşmuştur. İkinci faktörde ise en yüksek korelasyona sahip değişkenler; 3031, 3030, 3032 ve 3083 nolu fare isimleri ile belirtilen değişkenlerdir.

Önemli temel bileşenlerin belirlenmesinde, faktörler ile farelere ait gen ekspresyon ölçüleri arasındaki yukarıda anlatılan ilişkilerden anlaşıldığı üzere, 3026, 3006, 3008 ve 3018 nolu farelere ait değerlerin birinci faktör etrafında, 3031, 3030, 3032 ve 3083 nolu fareler ait değerlerin ikinci faktör etrafında yoğunlaştığı söylenebilir. Diğer faktörlere ait değerlerde aynı şekilde izah edilebilmektedir. Örneğin toplam varyansın %48,54 ünü açıklayan ilk üç faktör, ele alınan 6675 gen ve 20 dizi yapısındaki cDNA mikrodizi verileri içinde gen ifade profilini tanımlama bakımından, sırası ile 3026, 3006, 3008, 3031, 3030, 3032, 3083 nolu değişkenler ile yorumlanabilir.

Bu analiz sonuçlarına göre, 6675 gen ve 20 dizi yapısında, cDNA mikrodizi veri seti içinde gen ifade profilini belirleyen 20 fareye ait dizi, 9 bileşen etrafında yoğunlaşmıştır. Veri matrisinin yapısına ve problemin özelliğine göre karar verme yöntemlerini birlikte ele alarak ana bileşen sayısına karar verilmelidir. Bu temel bileşenler diğer başka analizlere başlamadan (örneğin kümeleme analizi ile gen grupları tanımlanmaya başlamadan) değişkenler hakkında bir fikir verebileceği gibi çok yüksek sayıda deneme materyali ile çalışmak yerine daha az varyans kaybıyla ancak bütün yapıyı izah edebilecek deneme materyalini içeren veri ile analiz yapma konusunda ışık tutmaktadır.

Sonuç

Temel bileşenler analizi, değişkenler arasında bağımlılığın olması ve boyutlarının fazla olması durumunda, çok değişkenli analizin diğer yöntemleriyle birlikte çok daha başarılı sonuçlar vermektedir. Ancak temel bileşenler analizine bazı eleştiriler de vardır. Bunlar; Bu analizin değişkenler arasında doğrusal bir ilişki olduğu varsayımına dayanmasıdır. Ayrıca, bu yöntemin öncelikli amacı boyut indirgeme olduğundan, bilginin tümü kullanılmamaktadır. Temel bileşenler analizi, ham değerlere veya standartlaştırılmış değerlere uygulanabildiğinden veri yapısına göre korelasyon matrisinden ya da kovaryans matrisinden faydalanarak analizler yapılabilmektedir. Bu durumda, farklı temel bileşen değerlerinin ortaya çıktığı bilinmektedir.

Bu çalışma da 2000'li yıllarda ortaya çıkan ve çalışma alanı genişleyen cDNA mikrodizi teknolojisi ve bu teknoloji ile beraber binlerce geni içeren yapıların çok değişkenli istatistik analiz yöntemlerinden biri olan temel bileşenler analizi ile veri yapısının ortaya konulması tartışılmıştır.

Sonuç olarak, temel bileşenler analizi, değişkenler arasında bağımlılığın olması ve boyutlarının fazla olması durumunda, çok değişkenli analizin diğer yöntemleriyle birlikte uygulandığında çok daha başarılı sonuçlar vermektedir. Bunun yanında yeni ve ümit verici olan mikrodizi teknolojisinin önümüzdeki yıllarda yaygın olarak kullanılacağı ve genlerin fonksiyonlarına ilişkin önemli bilgiler sağlayacağı beklenmektedir. Elde edilen genetik bilgilerin, hastalıkların teşhisi ya da önlenmesi ve kişilerde mutasyona neden olan kimyasal ve fiziksel ajanlara duyarlılığın belirlenmesi (risk tayini) için kullanılması hedeflenmektedir.

Bu amaçla genetik testler, ön plana çıkmakta olup, taşıyıcı bireylerin belirlenmesi için rutin taramalar doğum öncesi teşhis yetişkinlikte başlayan hastalıkların, belirtiler ortaya çıkmadan önce teşhisi (örneğin Huntington hastalığı) ve risk tayini (örneğin çeşitli kanser türleri, Alzhemier hastalığı) hastalık belirtileri sonrası kesin tanı gibi konularda faydası olacağı düşünülmektedir.

cDNA mikroarray teknolojisi kullanılarak elde edilen bilgiler, daha güçlü, iklim koşullarına ve hastalıklara karşı daha dayanıklı, besin değeri daha yüksek bitki ve hayvanlar yetiştirilmesinde kullanılabilecek, bunların yanı sıra gen aktarımı yoluyla zararlı böceklere karşı kendi biyopestisidini üreten ve bu bakımdan toksik pestisitlerin tarımda kullanımına son verecek bitkilerin, yenilebilir bir aşı olma özelliği sunan bitkilerin üretilmesi mümkün olabilecektir.

Kaynaklar

Anderson TW. 1974. An Introduction to Multivariate Statistical Analysis. John Wiley&Sons, Inc., New York, USA,374s.

Barash Y, Dehan E, Krupsky M. 2004. Comperative Analysis of Algorithms for Signal Quantitaion from Oligonucleotide Microarrays. Bioinformatics. 20(6):839-846.

Brown PO, Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. Nature genetics. 21:33-37.

Brown SM, Grundy NW, Lin D, Cristianini N. 2000. Knowledge-Based Analysis of Microarray Gene Expression Data by Using SVM. Pnas.97(1): 262-267.

Causton HC, Quackenbush J, Brazma A. 2003. A Beginner's Guide: Microarray Gene Expression Data Analysis. Blackwell Publishing, Malden,USA,160s.

Cooley WW, Lohnes PR. 1971. Multivariate Data Analysis. John Wiley&Sons, Inc., New York, USA,364s.

Collins FS, Morgan M, Patrinos A. 2003.The Human Genome Project: Lessons from Large- scale Biology. Science. 300:286-290.

Cox JM. 2001. Applications of nylon membrane arrays to gene expression analysis. Journal of Immunol meth. 250:3-13

Draghici S. 2003. Data Analysis Tolls for DNA Microarrays. Chapman&Hall/Crc., NewYork, Usa, 477s.

Eisen MB, PT Spellman, PO Brown, D Botstein. 1998. Cluster Analysis and Display of Genome-wide Expression Patterns. Proc. Natl. Acad. Sci. 95(25), 14863-14868.

Gardeux V, Natowicz R, Wanderley MFB, Chelouah R. 2013. Optimization for feature selection in DNA microarrays. Heuristics: Theory and Applications.

Golub, TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Bloomfield CD. 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286(5439), 531-537.

Heinloth AN, Irwin RC, Boorman AG. 2004. Gene Expression Profiling of Rat Livers Reveals Indicators of Potential Adverse Effects. Toxıcologıcal Sciences. 80:193–202.

Gonzalo R, Sanchez A. 2018. Chapter Three - Introduction to Microarrays Technology and Data Analysis. Comprehensive Analytical Chemistry. 82: 37-69.

Herrero J, Diaz R, Dopazo J. 2003. Gene Expression Data Preprocessing. Bioinformatics. 19: 655-656.

Johnson RA, Wichern N. 1982. Applied Multivariate Statistical Analysis. Prentice-Hall. London,333s

Jolliffe IT. 2002. Principal Component Analysis. Springer,New York,487s.

Jordan B. 2001. DNA Microarrays: Gene Expression Aplications. Springer, New York, USA, 140s.

Kamberova G, Shah S. 2002. DNA Array Image Analysis. DNA Pres LLC, USA, 206s.

Kerr MK, Churchill GA. 2001. Statistical Design and the Analysis of Gene Expression Microarray Data. Genetical Research. 77(2):123-128.

Landgrebe J, Welzl G, Metz T, Van Gaalen MM, Ropers H, Wurst W, Holsboer F. 2002. Molecular Characterisation of Antidepressant Effects in the Mouse Brain Using Gene Expression Profiling. Journal of Psychiatric Research, 36(3), 119-129.

Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. 1999. High density synthetic oligonucleotide arrays. Nat Genet. 21:20-4.

Leung FY, Cavalieri D. 2003. Fundamentals of cDNA Microarray Data Analysis. Trends in Genetics. 19: 649-659.

Lin SM, Johnson KF. 2002. Methods of Microarray Data Analysis. Kluwer Academic Publishers, Boston,182s.

Minitab 13 Statistical Software. 2004. [Computer software]. State College, PA: Minitab, Inc.

Newton MN, CM Kendziorski, CS Richmond, FR Blattner, KW Tsui. 2001. Improving Statistical Inference About Gene Expression Changes from Microarray Data. Journal of Computational Biology, 8: 37-52.

Ocampo RV, Sanchez GA, Luna M, Vega A. 2016. Improving pattern classification of microarray data by using PCA and logistic regression. Intelligent Data Analysis. 53-67

Özdamar K. 2002. Paket Programlar ile İstatistiksel Veri Analizi. Kaan Kitabevi. 513s.

Peterson LE. 2001. Factor analysis of Cluster-spesific gene expression levels from cDNA microarrays. Computer methods and programs in Biomedicine. 69:179-188.

Peterson LE. 2003. Partitioning Large-sample Microarray-Based Gene Expression Profiles Using Principal component Analysis. Computer methods and programs in Biomedicine. 70:107-119.

Rao MS, Van VTR, Ciurlionis R, Buck WR, Mittelstadt SW, Liguori MJ. 2019. Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. Frontiers in Genetics. (9): 636

Raychaudhuri S, Stuart JM, Altman RB. 2000. Principal Component Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. Pacific Symposium on Biocomputing, 452-463.

Schena M. 2001. DNA Microarrays: A Practical Approach. Oxford University Pres, New York, USA, 206s.

Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H. 2000. Normalization Strategies for cDNA Microarrays. Nucleic Acid Research. 28:4,1-5.

Shao G, Li D, Zhang J, Yang J, Shangguan Y. 2019. Automatic microarray image segmentation with clustering-based algorithms. PLoS ONE 14(1): e0210075.

Shoemaker DD, Schadt EE, Armour CD. 2001. Experimental Annotation of The Human Genome Using Microarray Technology. Nature. 409:922-927.

S-PLUS. 2000. TIBCO Software Inc.

Taguchi YH. 2017. Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and DrugMatrix datasets. Sci Rep.7(1):13733.

Tatlıdil H. 1996. Çok Değişkenli İstatistiksel Analiz, Akademi matbaası,424s.

Wagner A, Kumar R, Conley Y, Kochanek P, Berga S. 2015. Multiple Aromatization Mechanisms Influence Mortality and CNS Secondary Injury Profiles after Severe TBI. J Neurotrauma. 2011;28:871–888.