**Yolak Aktifleşme Skorlarının Hesaplanması**

Gen seti zenginleştirme analizi (Gene Set Enrichment Analysis, GSEA) önceden tanımlanmış bir gen setinin ya da biyolojik yolağın iki biyolojik durum arasında istatistiksel olarak anlamlı farklılıklar gösterip göstermediğini belirlemeye yarayan bir yaklaşımdır [1]. Her ne kadar GSEA yöntemi transkriptomik alanıyla özdeşleşmiş görülse de zenginleştirme analizi, eldeki verilerin daha önceden oluşturulmuş biyolojik dizilerle karşılaştırılmasında ve yorumlanmasında sık kullanılan bir yöntemdir. Her bir örneklem için metabolomik veriler ve hedef gen listesi olarak metabolik yolaklar kullanılarak, yolakların örneklem üzerindeki zenginleştirme skorları (yolak aktifleşme skorları, ssGSEA) ile gerçekleştirilebilir. Hali hazırda literatürde metabolit seti zenginleştirme analizinde (Metabolite Set Enrichment Analysis, MSEA), GSEA yöntemine benzer şekilde önerilmiştir [4]. Bu nedenle proje kapsamında metabolit yolaklardan elde edilecek zenginleşme (aktifleşme) skorları GSEA yöntemindeki algoritma kullanılarak hesaplanacaktır. Burada sadece genler yerine metabolitler ve gen yolakları yerine metabolik yolaklar (örn. metabolom veri tabanı HMDB (<http://www.hmdb.ca/>) ve METLIN ([https://metlin.scripps.edu/)](https://metlin.scripps.edu/%29)) kullanılacaktır. Analizler MetaboAnalyst 5.0 ve açık kaynak kodlu BIOCONDUCTOR kütüphanesi MetaboAnalyst 3.0 yazılımlarında gerçekleştirilecektir. Aşağıda GSEA ve ssGSEA yöntemleri detaylı bir şekilde açıklanmıştır.

GSEA yönteminde bahsi geçen biyolojik durum ya da diğer adıyla fenotip kadın-erkek, hasta-sağlıklı, uzun-kısa gibi örneklemlerin mevcut sınıf etiketlerini belirtmektedir. Fonksiyonel zenginleştirme analizi olarak da bilinen bu yöntem 3 ana adımdan oluşmaktadır.

1. **Veri Seti İçerisindeki Genlerin Sıralanması**

İlk aşama elimizdeki veri seti $D$ içerisinde yer alan farklı genlerin sıralanmasıdır. Her genin ilgili sınıf etiketi (fenotip) ile ilişki derecesini gösteren değerin hesaplanması gereklidir. GSEA, gen korelasyon değerini hesaplamak için varsayılan olarak formül (1)’de verilen sinyal-gürültü oranı kullanmaktadır. $G=\{g\_{1},g\_{1},…,g\_{N},\}$ veri setinde bulunan $N$ adet eşsiz geni göstermek üzere, denklemde yer alan $A$ ve $B$ sınıf etiketlerini , $μ\_{A}(g\_{i})$ ve $σ\_{A}(g\_{i})$ ise ilk sınıf içerisinde yer alan $g\_{i}$ gen ifadesinin sırasıyla ortalama değerini ve varyans değerlerini temsil etmektedir.

$r\left(g\_{i}\right)= \frac{μ\_{A}(g\_{i})- μ\_{B}(g\_{i}) }{σ\_{A}(g\_{i})- σ\_{B}(g\_{i})}$ (1)

Hesaplanan değerler büyükten küçüğe sıralanarak genlerin sınıflar ile korelasyon gücünü gösteren $L = \{r\_{1},r\_{2},…,r\_{N}\}$ listesi oluşturulur. Listenin ilk elemanı $A$ sınıfı, son elemanı $B$ sınıfı ile yüksek bir ilişki değerine sahip olmuş olur.

1. **Zenginleştirme Skorunun Hesaplanması**

Şekil 1. Geçerli Toplam İstatistiği

Zenginleştirme skoru (Enrichment Score, ES), daha önceden oluşturulmuş gen listesi $S = \{g\_{1},g\_{2},…,g\_{M}\}$ ile verilerimizden sıralanmış gen listesi$ L$ içerisindeki değerlerin karşılaştırılması sonucunda sınıf etiketlerinin (fenotip) üstündeki aşırı temsil edilme derecesini yansıtır. Bir önceki aşamada $L$’de yer alan genler ilk sınıf ilişki değerinden ikinci sınıf ilişki değerine doğru sıralanmıştır. Dolayısıyla eğer $S$ içerisinde yer alan genler, $L’$in ilk genleriyle daha çok eşleşirse bu demektir ki $S$ listesi $A$ sınıfı ile daha çok temsil edilebilir. Bu temsil gücünün gösterilmesi için de zenginleştirme skoru kullanılmaktadır. $S$, yolak içerisinde yer alan genler olabileceği gibi bağımsız oluşturulmuş bir gen listesi de olabilir.

Zenginleştirme analizinde, $L$ içerisinde yer alan genler üzerinde sırayla dönülür. Liste içerisinde ilgili gene kadar olan genlerin zenginleştirme skoru eğer ilgili gen$ S$ içerisinde geçiyorsa formül (2)’deki gibi arttırılırken, eğer geçmiyorsa formül (3)’teki gibi azaltılamktadır. Dolayısıyla her gene kadar olan zenginleştirme skoru kümülatif toplanarak gelir. Bu yaklaşıma geçerli toplam istatistiği denilmektedir. Örneğin şekil 1’de A geni listede bulunduğu için skor artarken, B geni bulunmadığı için azaltılmıştır. Bu şekilde ulaşılan maksimum ya da minimum değer zenginleştirme skorunu belirtir. Artış miktarı ilgili gene ait korelasyon değerinin ağırlıklı değeri $\left|r\_{j}\right|^{α}$ ile orantılıdır. Yani her gen için artış aynı değildir.

$T\_{eşleşme}\left(S,i\right)=\sum\_{\begin{array}{c}gen\_{i} \in S\\j \leq i \end{array}}^{}\frac{\left|r\_{j}\right|^{α}}{N\_{R}}, N\_{R}= \sum\_{g\_{j} \in S}^{}\left|r\_{j}\right|^{α}$ (2)

$T\_{ıskalama}\left(S,i\right)=\sum\_{\begin{array}{c}gen\_{i} \notin S\\j \leq i \end{array}}^{}\frac{1}{(N-N\_{S})}$ (3)

$ES\left(S,i\right)=T\_{eşleşme}(S,i)-T\_{ıskalama}(S,i)$ (4)

Formüller içerisinde yer alan $i$ değeri $L$ dizisinde işleme tabi tutulan genin indisini;$ j$ ise 0’dan $i$’ye kadar olan indis değerlerini göstermektedir. Böylece $i$ indisinden önceki tüm indislerdeki genlerin zenginleştirme skorları kümülatif olarak toplanmış olur. Bunun sebebi zenginleştirme skorunun tekil gen için değil $L$ dizisindeki ilgili gene kadar olan tüm genleri içerecek şekilde hesaplanmasıdır. Çünkü zenginleştirme analizi karşılaştırılan gen sayısı arttıkça ulaşılan en yüksek mutlak skoru hesaplamayı amaçlar. Formülde yere alan diğer ifadelerden $N, L$ listesinin uzunluğunu; $N\_{S}, S$ listesinin uzunluğunu; $α$ ise zenginleştirme ağırlık katsayısını tanımlamaktadır. Varsayılan olarak zenginleştirme ağırlık katsayısı 1 kabul edilmektedir. Böylelikle korelasyon değeri birebir kullanılmış olunur. Eğer çok sayıda gen içeren kümeler kullanılıyor ve yalnızca ufak bir alt kümenin tutarlı olması bekleniyorsa zenginleştirme katsayısı yükseltilebilir.

Şekil 2. Gen Seti Zenginleştirme Analizi

$L$ listesinde her $i.$ gene kadar olan zenginleştirme skoru formül (4)’teki $ES\left(S,i\right)$ ile ifade edilir. Listedeki her genden sonra bu skor hesaplanır. $ES\left(S,i\right)$ içerisinden sıfırdan sapma değeri en yüksek olan ya da başka bir ifadeyle mutlak değeri en büyük olan değer nihai $ES\left(S\right)$ olarak belirlenir.

1. **Zenginleştirme Skorunun Anlamlılık Değeri**

Zenginleştirme analizi iki farklı sınıf etiketi farklılıklarının bir başka gen setiyle yorumlanmasına ve gen setinin hangi sınıf ile daha uyumlu olduğunu ortaya çıkartmaya yarayan yaklaşımdır. Ancak ortaya çıkan skorun rastgelelerken gelip gelmediği de önemlidir. Çünkü rastgele bir sınıf dağılımında da aynı zenginleştirme skoru elde edilebilir. Bu doğrulamanın yapılabilmesi için sıfır hipotezinden yararlanılmaktadır. İlk olarak permütasyon testi ile sıfır hipotezi için örneklemler oluşturulur. Permütasyon testi sıfır hipotezi altında test istatistiğine örneklem dağılımı oluşturmak için basit bir yol sağlar [2].

Şekil 3.P-Değeri Hesaplama

Permütasyon verilerinin oluşturulması için $D$ içerisinde yer alan örneklemlerin sınıf etiketleri rastgele değiştirilerek $D\_{sıfır}$ oluşturulur. $D\_{sıfır}$ içerisinde yer alan genlerin korelasyon skoru formül (1)’deki gibi tekrar hesaplanır. Daha sonra korelasyona göre sıralanan genlerin zenginleştirme skoru $ES\_{sıfır}$ 2.aşamada anlatıldığı gibi hesaplanır. Bu işlem 1000 defa tekrarlanır ve permütasyon testi için 1000 tane $ES\_{sıfır}$ elde edilir. Sonuç olarak rastgele oluşturulmuş bir veri setlerinden elde edilmiş 1000 tane zenginleştirme skoru hesaplanmış olur. Şekil 3’tekine benzer olarak skorların hangisinin kaç tane geçtiğini gösteren dağılım oluşturulur. Dağılım içerisinde gerçek zenginleştirme yani veri setinin herhangi bir rastgeleye uğratılmadan aldığı skorun kaç kere geçtiğine bakılır.

Örneğin gerçek zenginleştirme skoru $ES(S)$ 0.2 ve permütasyon testi tekrar sayısı 5000 olsun. Permütasyon testleri neticesinde oluşan 5000 $ES\_{sıfır}$’dan 25 tanesi 0.2 skor aldığını varsayalım. Bu durumda gerçek zenginleştirme skorunun dağılım içerisinde geçme sayısı da 25’tir. Bu sayının toplam tekrar sayısına bölerek, 25/5000, P-değeri elde edilir.

$NES= \frac{ES(S)}{\frac{\sum\_{}^{}ES\_{sıfır}}{K}}$ (5)

Gen seti boyutları, setler ile ifade veri seti arasındaki korelasyon farklılıkları zenginleştirme skorunu etkileyen önemli faktörlerdendir. Dolayısıyla farklı iki gen seti geldiğinde ortaya çıkacak zenginleştirme skorları birbirleri için çok fazla anlam ifade etmeyebilir. Bu sorun formül (5)’teki gibi gerçek zenginleştirme skoru $ES(S)$’in rastgele oluşturulmuş verilerden elde edilen skorların ortalamasına bölünerek elde edilir. K sayısı permütasyonun kaç kere tekrar edildiği ya da başka bir deyişle kaç tane $ES\_{sıfır}$ hesaplandığını göstermektedir.

**Örneklem Temelli Zenginleştirme Skoru**

Daha önceden de belirtildiği gibi zenginleştirme analizi veri seti içerisindeki genlerin hangi sınıf etiketinde daha çok zenginleştiğinin istatistiğini hesaplar. Ancak analizin örneklem bazında kullanılabilmesi için sınıf bilgisinden ayrıştırılmalıdır. Çünkü her bir örneklemin tek bir sınıfı olacağından, sınıfa göre bir hesaplama yapılması mümkün değildir. Bunun yerine örneklemin gen ifade değerleri üzerinden hesaplanan Tekil-Örnek Gen Zenginleştirme (Single-Sample GSEA, ssGSEA) analizinden faydalanılır. ssGSEA sınıf etiketinden bağımsız olarak her bir örneklem ve gen seti $S$ için ayrı ayrı zenginleştirme skoru hesaplar [3]. Her bir gen korelasyon değeri yerine gen ifadesinin mutlak değerine göre sıralanarak $L = \{gen\_{1},gen\_{2},…, gen\_{N}\}$ oluşturulur. Veri seti içerisindeki her bir tekil örneklem $m$, örneklemin içerdiği gen ifadesine göre sıralı genler $L$,$ L$ listesinin uzunluğu $N$, yolak ya da daha önceden belirlenmiş gen listesi $S$, $S$ listesinin uzunluğu $N\_{S}$ ile ifade edilmek üzere her bir örneklemin zenginleştirme skoru formül (6) ile hesaplanmaktadır.

$ES\left(S,m\right)= \sum\_{i=1}^{N}[P\_{S}^{w}\left(S,m,i\right)-P\_{NS}(S,m,i)]$ (6)

$P\_{S}^{w}\left(S,m,i\right)=\sum\_{gen\_{j }\in G, j\leq i}^{}\frac{\left|gen\_{j}\right|^{α}}{\sum\_{gen\_{j }\in S}^{}\left|gen\_{j}\right|^{α}} , P\_{NS}\left(S,m,i\right)=\sum\_{gen\_{j }\notin S, j\leq i}^{}\frac{1}{(N-N\_{S})}$ (7)

Sıralı liste $L$’de bulunan ve indisi $i$’e kadar olan her bir gen eğer $S$ seti içinde geçiyorsa bu genlerin ağırlıklı deneysel kümülatif dağılım fonksiyonu (Empirical Cumulative Distribution Functions, ECDF) formülde yer alan $P\_{S}^{w}\left(S,m,i\right)$ ile hesaplanırken, geçmeyenlerinki $P\_{NS}(S,m,i)$ ile hesaplanmaktadır. Dolayısıyla örneklemde yer alan $i.$ gen için o gene kadar olan kümülatif zenginleştirme skoru formüldeki $[P\_{S}^{w}\left(S,m,i\right)-P\_{NS}\left(S,m,i\right)]$ farkı ile belirlenir. Bu kısımda geçme toplamından geçmeme toplamı çıkartılmaktadır. Örneklemin nihai zenginleştirme skoru için ise $L$ içerisinde yer alan tüm genler için bu işlemin yapılması ve her gen sonucunda çıkan skorların toplanması gereklidir. Böylelikle $S$ gen setinin, $m$ örnekleminin tamamı üzerindeki zenginleştirme skoru $ES\left(S,m\right)$ hesaplanmış olur.

[1] Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proceedings of the National Academy of Sciences* 102.43 (2005): 15545-15550.

[2] Rice, Ken, Thomas Lumley, and U. W. Biostatistics. "Permutation tests." *UW Biostatistics: Seattle, WA, USA* (2008).

[3] Barbie, David A., et al. "Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1." *Nature* 462.7269 (2009): 108-112.

[4] Xia, Jianguo, and David S. Wishart. "MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data." *Nucleic acids research* 38.suppl\_2 (2010): W71-W77.