MİKRODİZİN ANALİZİNDE GÜRBÜZ GEN İFADESİ ELDE EDEBİLMEK MAKSADIYLA KULLANILAN ÖN İŞLEME ALGORİTMALARININ KARŞILAŞTIRILMASI VE TEKNİK TEKRARLARIN BAŞARIMLARININ ANALİZİ

INVESTIGATION AND COMPARISON OF THE PREPROCESSING ALGORITHMS FOR MICROARRAY ANALYSIS FOR ROBUST GENE EXPRESSION CALCULATION AND PERFORMANCE ANALYSIS OF TECHNICAL REPLICATES

H. Gökhan İlk [1,], Özlem İlk [2], Özlen Konu [3], Hilal Özdağ [4]*

ilk@eng.ankara.edu.tr , oilk@metu.edu.tr, konu@fen.bilkent.edu.tr, hilalozdag@gmail.com

[1] Ankara Üniversitesi, Mühendislik Fakültesi, Elektronik Mühendisliği Bölümü, Beşevler, Ankara
[2] Orta Doğu Teknik Üniversitesi, İstatistik Bölümü, Ankara
[3] Bilkent Üniversitesi, Moleküler Biyoloji ve Genetik Bölümü, Bilkent, Ankara
[4] Ankara Üniversitesi, Biyoteknoloji Enstitüsü, Beşevler, Ankara

Özetçe

cDNA ve oligo mikrodizin verilerinin, istatistiksel analizlerini gerçekleştirmeden önce arkaplan çıkarımı, normalizasyon, ve özetleme sırası ile açıklanabilecek ön-işlemlerden geçirilerek standardizasyonu gerekmektedir. Affymetrix verilerinin analizi için kullanılmakta olan belli başlı ön-işleme algoritmaları arasında, RMA, dChip, ve MAS5 gelmektedir. Önceki çalışmalar RMA metodunu en gerçekçi algoritmalardan biri olarak gösterirken, MAS5 algoritması daha fazla hata payı içeren bir algoritma olarak karakterize edilmiştir. Bu çalışmada, RMA, dChip ve MAS5 algoritmalarının performansı mikrodizin teknik tekrarları arasındaki değer farklılıklarının ROC karakterleri göz önüne alınarak karşılaştırılmıştır. Her üç algoritmanın da "latin square" deneylerinden [1] seçilen teknik tekrarların kalitesini benzer şekilde sıraladığı gözlenmiştir. Diğer yandan, RMA diğer metodlarla karşılaştırıldığında ROC eğrisi altında kalan alanı maksimize ettiğinden daha yüksek performans gösterdiğini söylemek mümkündür. Bu makalede önerilen metod, mikrodizin deneylerindeki teknik tekrarlarda yer alabilecek lokal ve global hataların tespitinde de kullanılabilir.

Abstract

Preprocessing of microarray data involves the necessary steps of background correction, normalization and summarization of the raw intensity data obtained from cDNA or oligo-arrays before statistical analysis. Several algorithms, namely *RMA, dChip,* and *MAS5* exist for the preprocessing of Affymetrix microarray data. Previous studies have identified *RMA* as one of most accurate algorithms while MAS5 was characterized with lower accuracy and sensitivity levels. In this study, performance of different preprocessing algorithms have been compared in terms of ROC characteristics of pairwise intensity differences of microarray replicates. Our findings indicated that all three algorithms predicted in similar order the quality of the technical replicates obtained from a selected set of latin square experiments [1]. On the other hand, *RMA* exhibited higher performance in terms of accuracy by maximizing the area under the receiver operating curve. The proposed method also is useful for detection of global and/or local artifacts associated within the technical replicas of a microarray experiment. Therefore this study is unique in the sense that it provides an extensive investigation and comparison of preprocessing algorithms and proposes a novel method for the detection and identification of fine technical replicate pair.

1. Giriş:

Genetik biliminden genombilime geçişte ortaya çıkan ve giderek güçlenen teknolojilerin başında DNA mikrodizin teknolojisi gelmektedir. Bütün bir genomun bağıl ifadelenme profilini mRNA düzeyinde çıkaran bu teknolojinin ilk uygulamaları cDNA parçalarının cam slaytların üzerine basılması ile gerçekleştirilmiştir [2]. Mikrodizinler cDNA parçalarının sentetik oligonükleotidler halinde cam yüzeylere direkt olarak fotolitografik sentezi ile de üretilebilmektedir [3].

1.1. Mikrodizin analizi:

Mikrodizin üretiminde Affymetrix™ şirketi (Affymetrix, Inc., Santa Clara, CA, USA) tarafından kullanılan fotolitografik sentez yöntemi ile şirketin tasarladığı, birçok organizmanın bütün genomunu içeren mikrodizinler genom araştırmalarında kullanılmaktadır. Bir genin 11 ila 20 adet 25 baz çifti uzunluğunda DNA probları tarafından temsil edildiği bu mikrodizinlerde özgün olmayan hibridizasyonu modellemek üzere herbir eşleşmiş probun (perfect match) bir uyumsuz probu (mismatch) da mikrodizin üzerine sentezlenmektedir. Uyumsuz problar 25 bazlık dizilerinde eşleşmiş problardan yalnızca 13. bazlarında farklılık gösterecek şekilde tasarlanırlar. Bu durumda örneğin bütün insan genomunu temsil eden 47,000 transkript içeren insan dizisinde (Affymetrix HG_U133 Plus2) yaklaşık toplam 1.5 milyon prob çeşidi bulunmaktadır. Deney sonucunda alınan 1.5 milyon veri noktası değişik ön işleme algoritmalarının (RMA, dCHIP, MAS5) uygulanması ile herbir transkript için bir değer verecek şekilde arkaplan çıkarımı (background correction), normalizasyon ve özetleme (summarisation) aşamalarına uğrar. Bioconductor yazılımları (*affybatch*) bu tür ön işlemleri yapmak için gerekli paket programları içermektedir. Örneğin, arkaplan ayarlaması için RMA konvolüsyon veya MAS 5.0 arkaplan yazılımları kullanılabilir (R, www.bioconductor.org).

1-4244-0239-5/06/$20.00 ©2006 IEEE

1.2. Ön işleme algoritmaları

Ön işleme algoritmaları incelendiğinde temel olarak üç algoritma dikkat çekmektedir. Bunlar RMA [4], dchip [5] ve MAS5 [6] olarak genelleştirilebilir. Ayrıca tüm metodlar arkaplan çıkarımı (background correction), normalizasyon (normalisation) ve özetleme (summarisation) sırasında aşamaları içermektedir. Bu aşamaların ne şekilde yapıldığı ve hangi veriyi temel aldığı algoritmalar arasındaki temel ancak önemli farklılıkları vermektedir. İlgili kaynaklarda [4,5,6] algoritmaların detayları ve çalışma prensiplerinde kullanılan kuramsal bilgiler detaylı olarak açıklanmıştır.

Örneğin, RMA (Robust Microarray Analysis) metodu sadece PM (perfect match) problarını kullanır ve bu probların normal dağılımlı bir hata (arka plan gürültüsü) ile üstel dağılımlı bir sinyal bileşeni olduğunu varsayar. Bunun yanında MAS 5.0 algoritması ise çipi 16 eşit dikdörtgen alana ayırarak her alandaki en düşük ışımalı probların (tüm probların %2'si) ortalama ışıma değerini mazgala özgün arkaplan değeri olarak kabul eder. Daha sonra her bir probdan, mazgalların merkezlerine olan uzaklığı ile ters orantılı olarak bir arkaplan sinyali çıkarılır, ve bu işlem hem PM (perfect match) hem de MM (mismatch) probları için gerçekleştirilir.

Normalizasyon aşaması farklı mikrodizin çiplerinden elde edilen arkaplan düzeltmeleri gerçekleştirilmiş olan verilerin birbirleri ile uyumlu ve karşılaştırılabilir olmaları için gereklidir. Burada dikkat edilmesi gereken husus, arka plan gürültüsünden temizlenmiş ham verilerin farklı çipler kullanmasından dolayı normalize edilmeleri gerekliliğidir.

Özetleme aşaması aynı prob setine ait prob değerlerinin anlamlı bir biçimde tek bir değer verecek şekilde "özetlenmesini" içerir. Tablo 1'de mümkün olabilecek tüm arkaplan çıkarımı, normalizasyon ve özetleme yöntemleri sunulmuştur. Tablo 1'den de açıkca görülebileceği üzere bir kısmı anlamsız 420 farklı kombinasyonda ön işleme algoritması önermek mümkündür.

Tablo 1. Ön-işleme algoritmalarına ait metodlar

AŞAMA	***METOD***
Arkaplan çıkarımı	"mas","none","rma","rma2"
Normalizasyon	"constant","contrasts","invariantset" "loess","qspline","quantiles", "quantiles.robust"
PM düzeltme (gerekli ise)	"mas", "pmonly", "subtractmm"
Özetleme	"avgdiff", "liwong", "mas", "medianpolish","playerout"

Farklı ön-işleme metodları farklı hassasiyetlerde verileri analiz kabiliyetine sahiptirler. Şimdiye kadar yapılan çalışmalar [7] RMA metodunun oldukça başarılı olduğunu göstermiştir. Diğer bir çalışma ise 30'dan fazla önişleme algoritmasını ROC karakteristiklerine dayanarak karşılaşılaştırmış ve proba-özel arka alan çıkarımını kullanan GCRMA metodunun diğer metodlara olan üstünlüğünü göstermiştir [9]. RMA, dChip, ve MAS5 metodları, 'latin-square' diye bilinen verisetinin bir altkümesinin kullanıldığı bir çalışmada, ortalaması alınmış üçlü veri grupları arasındaki farkları ölçmedeki başarısı açısından 'rank' ya da 'küçükten büyüğe sıralama' metodu kullanılarak karşılaştırılmıştır [4]. Sözü edilen çalışma, RMA metodunun, birbirinden farklı konsantrasyonlarda eklenmiş (spike-in) prob setlerinin hemen hepsini gerçekte fark göstermemesi gereken probsetlerinden diğer metodlarlarla karşılaştırıldığında daha başarılı bir şekilde ayırabildiğini göstermiştir.

1.3. Bu çalışma ile önerilen yaklaşım

Bütün bir genomun ifade profilinin diğer bir deyişle moleküler imzasının güvenilir ve sağlam bir şekilde çıkarılabilmesi için deneysel değişkenlerden kaynaklanabilecek hataların en aza indirgenmesi hedeflenmektedir. Mikrodizin deneyleri tasarlanırken deneysel değişkenlerin güvenilirliği, teknik tekrarlar yapılmak suretiyle sağlanmaya çalışılır.

Bu çalışmada mikrodizin analizlerinde güvenilir ve sağlam gen ifadesi elde etmek üzere kullanılan algoritmaların (RMA, dchip, MAS5) karşılaştırmaları yapılmış ve teknik tekrarların başarımlarını tayin eden ROC tabanlı bir yaklaşım uygulanmıştır. ROC analiz sonuçları geliştirdiğimiz optimizasyon tabanlı diğer bir yaklaşım ile de doğrulanmış olup bu karşılaştırmanın detayları bu makalenin kapsamı dışında tutulmuştur..

2. Metod

Affymetrix mikrodizin analizleri GeneLogic™ Latin Square verileri üzerinde gerçekleştirilmiştir [1]. Bu veritabanı içinde BIOB, BIOC, DAPX, ve CRE bakteriyel genlerinin farklı bölgelerine bağlanacak şekilde dizayn edilen 11 adet cRNA fragmanının herbirinin farklı konsantrasyonlarda bulunduğu tekrar edilmiş oligo dizi verileri Tablo 2'de sunulmuştur. Bu diziler HU95A (Human Genome) GeneChip'leri olup toplam altı adet chip içermektedir. Her bir dizide kullanılan ortak komplex cRNA akut myeloid lösemi hücre hattından edinilmiş olduğundan prob setlerin 11'i hariç diğer genler için çipler arası farklılık göstermemesi beklenmektedir. Bu nedenle, Tablo 2'de sunulan *11 adet spike-in genlerindeki konsantrasyon farkları "differentially expressed"* (anlamlı fark) çip üzerinde bulunan *diğer 12,615 gen ise "non differantially expressed"* (anlamsız fark) olarak tanımlanabilir.

Bu nedenle "non differentially expressed" genlerin doğru olarak tanımlanmaları gerçek pozitif (TP), "differentially expressed" genlerin hatalı olarak tanımlanmaları yanlış pozitif (FP) olarak belirlenmiştir. Doğal olarak TP ve FP tanımlamaları değiştirilebilir.

Tablo 2'nin ilk sütununda isimleri yer alan mikrodizinler diğer sütunlarda yer alan konsantrasyon bilgilerinden de kolaylıkla anlaşılabileceği gibi iki adet farklı grubun üçer adet teknik tekrarıdır. Bu mikrodizinler için RMA, dChip ve MAS5 ön işleme algoritmaları uygulanmış, elde edilen gen ifadesi değerleri her bir grubun her bir teknik tekrarı için karşılaştırılmıştır (www.bioconductor.org). Bu sayede toplam 9 adet fark değerine ulaşılmıştır. Bunlar 1-1, 1-2, 1-3, 2-1, 2-2, 2-3 ,3-1, 3-2 ve 3-3 teknik tekrar karşılaştırmalarıdır. Bu karşılaştırmalarda ilk indis grubu, ikinci indis ise teknik tekrarı ifade etmektedir. Gen ifadesi farklarından elde edilen toplam 9 set veri üzerinde "differentially expressed" ve "non differentially expressed" sayıları ROC (receiver operating curve) eğrileri [8] kullanılarak karşılaştırılmıştır. Bu karşılaştırma sonuçları Tablo 3'de sunulmuştur.

Tablo 2. GeneLogic firmasından temin edilen "Latin Square" tasarımı için spike-in konsantrasyonları. Parantez içinde her bir gruptaki teknik tekrarların indisleri belirtilmiştir.

GeneChip array	***BioB-5_atpM***	***BioBM_atpM***	***BioB-3_atpM***	***BioC-5_atpM***	***BioC-3_atpM***	***BioDn-3_atpM***	***DapX-5_atpM***	***DapXM_atpM***	***DapX-3_atpM***	***CreX-5_atpM***	***CreX-3_atpM***
92561hgu95a11 (1)	0.5	37.5	25	75	100	50	1.5	1	3	2	5
92561hgu95a21 (2)	0.5	37.5	25	75	100	50	1.5	1	3	2	5
92561hgu95a31 (3)	0.5	37.5	25	75	100	50	1.5	1	3	2	5
92557hgu95a11 (1)	100	1	0.5	2	25	1.5	5	3	35.7	12.5	50
92557hgu95a21 (2)	100	1	0.5	2	25	1.5	5	3	35.7	12.5	50
92557hgu95a31 (3)	100	1	0.5	2	25	1.5	5	3	35.7	12.5	50
Göreceli konsantrasyon	***200***	***37.5***	***50***	***37.5***	***4***	***33.3***	***3.3***	***3***	***12.5***	***6.25***	***10***

Tablo 3. ROC eğrileri altında kalan "yaklaşık" alan değerleri ve gruplar arası teknik tekrar karşılaştırmaları

Yöntem Tekrar	RMA (sıralama)	RMA (alan)	Dchip (sıralama)	Dchip(alan)	MAS 5 (sıralama)	MAS 5 (alan)
1-1	5	0.9835	5	0.9687	6	0.9596
1-2	8	0.9732	8	0.9682	8	0.9595
1-3	2	0.9937	2	0.9689	3	0.9596
2-1	6	0.9788	6	0.9685	2	0.9597
2-2	1	0.9969	1	0.9689	1	0.9598
2-3	7	0.9757	7	0.9684	7	0.9596
3-1	4	0.9859	3	0.9688	4	0.9596
3-2	9	0.9698	9	0.9679	9	0.9592
3-3	3	0.9859	4	0.9687	5	0.9596

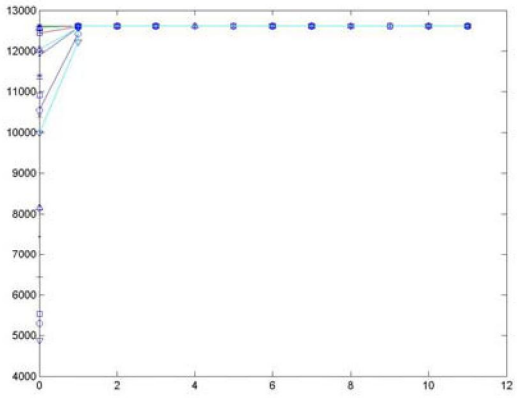
3. Tartışma

Tablo 3'de verilen ve ROC eğrileri altında kalan alanlar 1.0'a normalize edilerek hesaplanmıştır. Bu değerler aynı yöntem için anlamlı olup, farklı yöntemler için karşılaştırma yapılması anlamlı değildir. Bunun nedeni farklı yöntemlerde kullanılan TP (doğru pozitif) ve FP (yanlış pozitif) değerlerinin, gen ifadesi değerlerinin aralığı ile değişmesidir. Bu durum gen ifadesi değerlerinin histogramlarını çizdirerek rahatlıkla gözlemlenebilir. Tablo 3'den elde edilen en önemli gözlem her üç yönteminde en başarılı ve en başarısız teknik tekrar çiftini aynı şekilde bulmasıdır. Kullandığımız veri için bu teknik tekrar 2-2 çiftini önermektedir. 3-2 çiftinin en kötü karşılaştırma sonucunu vermesi ise ikinci grubun 2 numaralı teknik tekrarının en iyi örnek olmadığını göstermektedir. Bu sonuçtan yola çıkarak teknik tekrarın değil, teknik tekrar çiftlerinin göz önünde bulundurulması gerektiği görülmektedir. Ayrıca önerdiğimiz bu yöntem, kalite kontrol açısından da kullanışlı olup, ROC eğrileri altında kalan alanın belirlenen bir değerden düşük çıkması halinde bu teknik tekrarlardaki problemlere işaret edebilir.

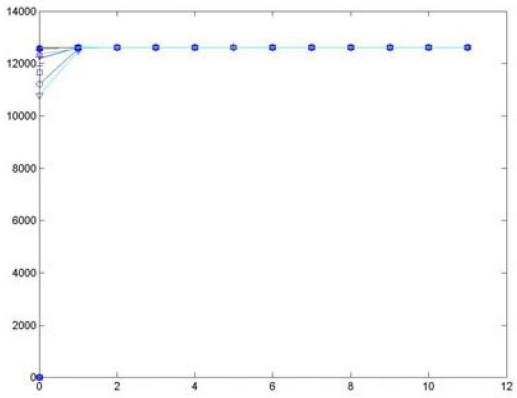
Şekil 1, 2 ve 3'de sırasıyla RMA, dChip ve MAS5 yöntemleri ile elde edilen tüm teknik tekrarlara ait ROC eğrileri verilmiştir. Bu şekillerden de açıkca görüldüğü üzere RMA ve dChip yöntemleri biri birine çok yakın sonuçlar vermekle birlikte MAS5 algoritması diğer iki algoritma kadar başarılı sonuçlar üretememektedir.

4. Sonuç

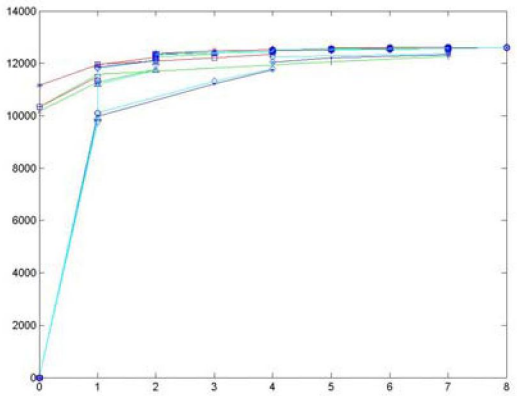
ROC eğrilerinin altında kalan alan kadar önemli bir diğer parametrede maksimum TP'e karşılık gelen noktada minimum FP veren sistemin optimum çalışma noktasıdır (operating point). Tablo 4'de en iyi teknik tekrar (2-2) için optimum çalışma noktasına tekabül eden doğru pozitif ve yanlış pozitif değerleri sunulmuştur.



Şekil 1. RMA yöntemi ile elde edilen tüm teknik tekrarlara (toplam dokuz adet) ait ROC eğrisi. X aksı FP ve Y aksı ise TP sayısını göstermektedir.



Şekil 2. dChip yöntemi ile elde edilen tüm teknik tekrarlara (toplam dokuz adet) ait ROC eğrisi. . X aksı FP ve Y aksı ise TP sayısını göstermektedir.



Şekil 3. MAS5 yöntemi ile elde edilen tüm teknik tekrarlara (toplam dokuz adet) ait ROC eğrisi. . X aksı FP ve Y aksı ise TP sayısını göstermektedir.

Bu değerlerden açıkca görülebileceği üzere RMA yöntemi tüm gerçek pozitifleri belirlerken hiç bir yanlış pozitif belirlememiştir. Bu veri setine bakarak RMA ön işleme yönteminin dChip ve MAS5 yöntemleri ile karşılaştırıldığında konsantrasyon değerlerine göre en doğru gen ifadesi değerlerini sunduğunu söylemek mümkündür. Genelleştirilmiş bir sonuç elde edebilmek maksadıyla yazarlar farklı "latin square" tasarımı ile elde edilmiş spike-in konsantrasyonları üzerinde karşılaştırma yapmaya devam etmektedirler.

Tablo 4. En iyi teknik tekrarın optimum işletim noktası değerleri.

	RMA	dChip	MAS5
TP	12,615	12,615	12,615
FP	0	4	6

Sonuç olarak bu makalede, teknik tekrarlardan en verimli olanların ROC eğrileri kullanılarak tespit edilmesine yönelik bir yöntem önerilmiş ve en iyi teknik tekrar çifti incelendiğinde konsantrasyon değerine karşılık gelen gen ifade değerlerinin RMA yöntemi ile daha doğru bir şekilde çıkarıldığı gösterilmiştir.

Kaynakça

[1] GeneLogic (2002) Datasets http://www.genelogic.com.

[2] **De Risi JL, Iyer VR, Brown PO,** (1997) "*Exploring the metabolic and genetic control of gene expression on a genomic scale*" Science, Vol. 278: 680-686.

[3] **Wodicka L, Dong H, Mittmann M, Ho MH, Lockhar DJ.** (1997) "*Genomewide expression monitoring in Saccharomyces cereverisiae*", Nat. Biotechnol,Vol. 15: 1359-1367

[4] **Irizarry,R.A., Hobbs,B., Colin, F., Beazer-Barclay,Y.D., Antonellis,K.,Scherf,U. and Speed, T.P**. (2003) "*Exploration, normalization and summaries of high density oligonucleotide array probe level data*" Biostatistics, Vol.4, 249– 264.

[5] **Li,C. and Wong,W.H.** (2001) "*Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error applications*" *Genome Biol.*, **2(8)**, 1–11.

[6] Affymetrix, Statistical algorithms reference guide, Technical report, (2001)

http://www.affymetrix.com/support/technical/manuals.affx

[7] **B. M. Bolstad, R. A. Irizarry, M. Astrand and T. P. Speed,** (2003) "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias" *Vol. 19 no.2, Pages 185–193*

[8] The magnificent ROC (Receiver Operating Characteristic curve) http://www.anaesthetist.com/mnm/stats/roc/#ssize

[9] Irizarry,R.A., Wu, Z., and Jafee H.A. (2006) "Comparison of Affymetrix GeneChip Expression Measures" Bioinformatics. Jan 12. [Epub ahead of print]