Kısa Süreli Mikrodizi Serilerinin Analizi ve Biyolojik Anlamlandırması
Short Time Series Microarray Data Analysis and Biological Annotation

Zerrin Sökmen, Volkan Atalay

Bilgisayar Mühendisliği Bölümü
Orta Doğu Teknik Üniversitesi
{zerrin.sokmen,volkan}@ceng.metu.edu.tr

Rengül Çetin Atalay

Moleküler Biyoloji ve Genetik Bölümü
Bilkent Üniversitesi
rengul@bilkent.edu.tr

Özetçe

Mikrodizi deney verilerini analiz ederek oluşturulan anlamlı gen listesinin, biyolojik işlevler açısından da anlamlandırması gerekmektedir. Bu çalışmanın amacı, ele alınan kısa süreli mikrodizi serisinde biyolojik açıdan ilintili genleri içeren kümeleri gözetimsiz yöntemlerle ortaya çıkartmak ve bu kümelerin otomatik olarak biyolojik anlamlandırılmasını yapmaktır. Çalışmanın ilk aşamasında, kısa süreli gen ifadesi içeren mikrodizi verisi benzer profile sahip olma özelliğine göre kümelenmektedir. İkinci adımda ele alınan bir kümedeki genlerle ilgili farklı kaynaklardan gelen biyolojik bilgiler bütünleştirilecek ve bütünleştirilmiş veriye göre yeni altküme(ler) oluşturulacaktır. Üçüncü adımda ise elde edilen altkümedeki genlere ait bilgiler kullanılarak biyolojik anlamlandırma yapılacaktır.

Abstract

Significant gene list is the result of microarray data analysis should be explained for the purpose of biological functions. The aim of this study is to extract the biologically related gene clusters over the short time series microarray gene data by applying unsupervised methods and automatically perform biological annotation of those clusters. In the first step of the study, short time series microarray expression data is clustered according to similar expression profiles. After that, several biological data sources are integrated to get information related with the genes in one of those clusters and new sub-clusters are created by using this unified information. As a last step, biological annotation of gene sub-clusters is performed by using information related with those sub-clusters.

1. Giriş

Mikrodizi deneyleri, onbinlerce genin anlık deney uygulanan hücre içersindeki gen ifadesine ait bilgiye ulaşılmasını mümkün kılmaktadır. Binlerce verinin analizi için parametrik veya parametrik olmayan istatistiksel testler ve işlemsel algoritmalar kullanılmaktadır. Analizler sonucu elde edilen uzun gen listelerinin (ortalama 2000-4000 gen içeren) biyolojik anlamlandırılması için özelleşmiş veri tabanlarından elde edilen bilgiler ile birleştirilmesi gerekmektedir. Kısa zaman serisi deneyleri, bir zaman süreci içersinde belli aralıklarla örnek alınarak gerçekleştirilir ve genelde 6-20 mikrodizi deneyini içerirler. 50 örneğin altındaki mikrodizi deneylerinin analiz edebilmek için geleneksel yöntemlerin kısa süreli veriye uyarlanması yada yeni yöntemlerin geliştirilmesi gerekmektedir. Geleneksel istatistiksel yöntemler araştırmacıya sadece deneye özgün anlamlı gen listesi vermektedir ve bu sonuçlar *genler arasındaki işlevsel ilişkiler* gözönünde bulundurulmadan elde edilir. Bu nedenlerle gen listesini biyolojik işlevler açısından anlamlandırmak için ikinci bir aşama olarak, otomatik yöntemleri (Onto-Express, FatiGO, Seq-Express, "Bioconductor" anlamlandırma paketleri vb.) uygulamak gerekmektedir.

Bu çalışmamızda, kısa süreli mikrodizi gen serilerinin analizi ve farklı veri kaynaklarından elde edilen gen bilgilerinin bütünleştirilerek biyolojik olarak anlamlandırılması için otomatik olarak çalışabilen çeşitli yöntemler geliştirmekteyiz. Çalışmada kullanılan mikrodizi verileri, hem genel erişime açık mikrodizi verileri hem de *Bilkent Üniversitesi Moleküler Biyoloji ve Genetik Bölümü Affymetrix Mikrodizi Laboratuarı*'ndan karaciğer kanseri için elde edilen özgün verilerinden oluşmaktadır. Dördüncü bölüm *Yöntem* kısmında detaylı olarak anlatılan aşamalı gen analizini, internet üzerinden sunulan biyolojik veri tabanlarından ve grubumuz tarafından geliştirilmiş öngörü araçlarından ya da kullanıma açık öngörü araçlarından elde edeceğimiz verileri birleştirerek yapmaktayız. Çalışmamızın ilk aşaması kısa süreli mikrodizi verisini benzer ifade profillerine göre gözetimsiz olarak kümelemektir. Sonraki aşamalarda ise benzer ifade profillerine sahip olan genlerin çeşitli veri kaynaklarından gelen bilgilerin de yardımıyla biyolojik olarak anlamlandırması yapılacaktır. Dolayısıyla çalışmamız sonuçlandığında, hem farklı veri kaynaklarından elde edilen bilgilerden yeni özniteliklerin çıkartılması ve bu değişik kaynaklı gen bilgilerinin mikrodizi gen analizi sırasında bütünleştirilmesi kısmında hem de kısa süreli mikrodizi gen serilerinin analizi konusunda özgün değerler ortaya çıkarmayı planlamaktayız.

2. Kısa Süreli Mikrodizi Analizi

Kısa süreli mikrodizi serilerinin analizi konusunda yapılan araştırmalar, zaman serisi analizi çalışmalarına göre oldukça kısıtlıdır. Bu alanda yapılan önemli bir çalışma Ernst ve çalışma arkadaşları tarafından gerçekleştirilmiştir [1]. Yöntemlerinin ilk aşaması, mikrodizi deneyi sırasında, herhangi bir gen tarafından sergilenebilecek tüm olası ifade profillerinin seçilmesidir. İkinci aşamada, her bir gen uygun profile atanır ve her bir profildeki genlerin zenginleştirme

978-1-4244-1999-9/08/$25.00 ©2008 IEEE

analizi yapılır, her bir profil için hesaplanan puana göre, anlamlı profiller tayin edilir, ve bu profiller analiz edilir. Sonuç olarak seçilen profiller Gen Ontoloji (GO) veri tabanı yardımıyla değerlendirilerek biyolojik fonksiyonlar belirlenmeye çalışılır. Başka bir çalışmada ise, gen ifadeleri arasındaki zaman-miktar bilgisinden de faydalanabilmek amacıyla polinomlara dayanan bir model geliştirilmiştir [2]. Gen ifadelerinin, deneyler sırasındaki dinamik ve birbirine bağlı yapısını dikkate alan Bayes tabanlı kümeleme yöntemini uygulamıştır. Yine başka bir çalışmada parçalı doğrusal fonksiyonlar olarak ifade edilen gen ifadeleri, "belirsiz kümeleme" yöntemi yardımıyla kümelenmiştir [3].

Zaman serisi mikrodizi verisini analiz etmek amacıyla, saklı Markov modelleri (SMM) de kullanılmıştır. Fakat bu çalışmaların çoğu kısa süreli zaman serisi üzerinde yoğunlaşmamıştır. Bir çalışmada karma SMM yöntemi kullanılarak mikrodizi gen ifadeleri kümelenmiştir [4]. Bu çalışmada SMM yöntemi, zaman serisi verisi içindeki zaman eksenindeki yatay bağlantıları daha iyi hesaba katabilmek amacıyla kullanılmıştır. Başka bir çalışmada ise, en başta n tane gen kümesi belirlenip, her bir gen kümesi için bir SMM eğitilip, tüm gen ifadeleri her bir SMM üzerinde sınanmıştır [5]. Mikrodizi gen ifadelerini kümelemek amacıyla uygulanan başka bir yaklaşım da sadece bir tane profil SMM kullanılmasıdır, fakat her bir zaman birimi için bir durum ve farklı gen ifade seviyeleri için farklı alt durumlar yaratılarak bu profil SMM oluşturulmuştur [6].

3. Biyolojik Bilgi Bütünleştirme

Mikrodizi gen ifadeleri üzerinde çeşitli eş-kümeleme çalışmaları da yapılmıştır. Bu çalışmalar çoğunlukla, tek bir tip mikrodizi gen ifadesi veri setini, biyolojik işlev bilgisini de kullanarak eş zamanlı olarak kümelemeyi amaçlamaktadır. Bu alandaki öncü çalışma Hanisch ve çalışma arkadaşları tarafından yapılmıştır [7]. Geliştirdikleri eş-kümeleme yöntemiyle, mikrodizi gen ifadeleri ile metabolik ağ (KEGG) yapısından gelen bilgiler, bir uzaklık fonksiyonu içinde birleştirilip, hiyerarşik kümeleme yönteminde kullanılmıştır. Başka bir çalışmada ise mikrodizi gen setindeki genler arasındaki benzerlik derecesi, GO bilgisini de kullanarak bulunmaktadır ve daha sonra hiyerarşik kümele yönteminde kullanılmaktadır [8]. "Memetic algoritma" kullanan bir çalışma ise, mikrodizi gen ifadeleri ile GO dizgesindeki uzaklık bilgilerini birleştirerek yüksek puanlı kümeleri belirlemeye çalışır [9]. Öz-düzenmeli haritalar kullanılan başka bir çalışmada ise, mikrodizi gen ifadeleri ile GO dizgesindeki uzaklık bilgileri birleştirilerek eş-kümeleme uygulanmıştır [10].

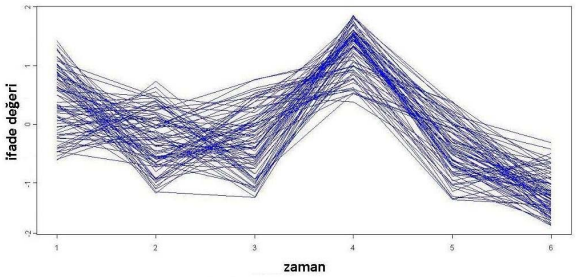
4. Yöntem

4.1. Kısa Süreli Mikrodizi Serilerinin Analizi

Mikrodizi deneyleri Bilkent Üniversitesi'ndeki laboratuarlarda tasarlanıp gerçekleştirildiği için, deney sonucunda oluşan gen ifadelerinin belli bir yada birkaç davranışı göstermeleri beklenmektedir. Bu şekildeki gen kümelerini tespit edebilmek amacıyla çalışmamızın ilk aşamasında, "k-orta değer kümeleme" yöntemi ile "saklı Markov modelleri" melezlenmektedir [11]. Bu melez yöntemde, ilk önce işlenmemiş mikrodizi verisi, "RMA" ön-işleme yöntemi kullanılarak normalleştirildi. Ardından bu veri üzerinde, her hücre aşamasında benzer davranışlar (tepkiler) gösteren gen gruplarını tespit edebilmek amacıyla k-orta değer kümeleme algoritması uygulandı. Kümeleme işlemi esnasında, iki gen arasındaki uzaklığı hesaplamak amacıyla "öklid uzaklık" ölçütü kullanıldı. Algoritmadaki toplam küme sayısı 100 olarak tayin edildi. Kümeleme işlemi bittikten sonra, bu 100 küme arasından gen ifadeleri açısından önemli gen kümelerini belirleyebilmek amacıyla, her kümenin kendi içindeki değişimi 1 nolu denklem ile hesaplandı ve aday olarak 13 tane küme seçildi.

$$\sigma_c^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^c - \mu^c)^2 \tag{1}$$

Buradaki x_i^c değeri, c nolu küme merkezinin i. sütun değerini; μ^c ise c küme merkezindeki tüm sütunların ortalama değerini gösterir. σ_c^2 ise c küme merkezi içindeki sapmanın ortalamasıdır ve σ_c^2 değeri belirli bir eşik değerinden yüksek olan kümeler aday olarak seçilmiştir. Belirlenen bu 13 aday küme içinden, biyolojik olarak anlamlı bir ifade örüntüsü sergilediği düşünülen bir tane küme seçildi ve *anlamlı küme* olarak adlandırıldı. Anlamlı küme içinde yer alan genler, SMM'ini eğitebilmek için gerekli olan ifade profilini oluşturdular (bakınız Şekil 1).



Şekil 1: Anlamlı küme içindeki genlerin ifade profili.

SMM'nin eğitim işleminden önce, ifade değeri kabul edilebilir bir aralıkta olmayan (belirli bir sapma değerinden fazla sapma gösteren) genler anlamlı kümeden çıkartıldı ve anlamlı kümede toplam 83 tane gen ifadesi kalmış oldu. Tasarladığımız SMM, ilgilendiğimiz veri kümesi için toplam 6 aşamadan (durum) oluşmaktadır. Orjinal mikrodizi verisinde, her bir genin ifade değeri reel sayılar ile gösterilirken, SMM içindeki her bir aşama için bu reel değerler 2 nolu denklem kullanılarak tam sayıyla ifade edilen sembollere (1, 2, ve 3) dönüştürülmüştür; bu semboller düşük, değişmeyen ve yüksek düzeydeki gen ifadelerine karşılık gelmektedir.

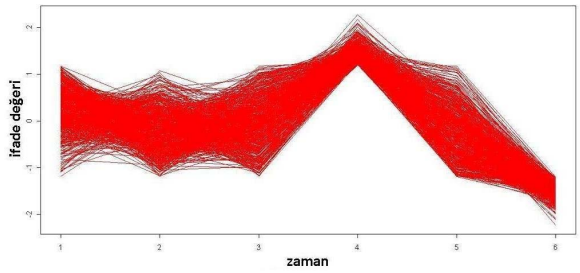
$$D_i = \begin{cases} 1, & S_i < 0 \\ 2, & 0 \le S_i < 1 \\ 3, & S_i \ge 1 \end{cases} \tag{2}$$

SMM'nin eğitimi sırasında anlamlı kümede yer alan 83 tane genin ifade değerleri ve Baum-Welch eğitme algoritması kullanıldı. Tüm mikrodizi veri kümesi içinde, daha önceden seçilen anlamlı gen kümesindeki gen ifadeleriyle benzer örüntüler gösteren genleri belirleyebilmek amacıyla, eğittiğimiz SMM'ni tüm veri kümesi üzerinde sınadık. Bu

sınama sırasında SMM'nden yüksek bir olasılık (denklem 3 ile hesaplanır) üreterek çıkan genler, yeni "anlamlı gen" listesine eklendi, bu şekilde toplam 620 tane gen tespit edildi.

$$P(Y) = \sum_{i=1}^{n} P(Y|S_i).P(S_i) \quad (3)$$

Bu 620 genin ifade profilleri, SMM'ni eğitmek için kullanılan 83 genden oluşan anlamlı gen kümesinin ifade profiline oldukça benzer bir davranış sergilemektedir (bakınız Şekil 2). Bu nedenle, geliştirmiş olduğumuz bu melez yöntem, ortak ifade örüntüleri sergileyen gen kümelerini tespit etme işini başarabilmektedir. Bu çalışmanın ileriki safhalarında, geliştirdiğimiz bu melez yöntem, halka açık veri tabanlarından elde edeceğimiz çeşitli özellikteki kısa süreli mikrodizi verilerinin analizini yapmak için kullanılacak, böylelikle daha da geliştirilip irdelenecektir.



Şekil 2: Tüm veri içindeki anlamlı genlerin (620 adet) profili.

4.2. Değişik Kaynaklardan Gelen Bilgilerin Bütünleştirilerek Anlamlı Genlerin Kümelenmesi

Mikrodizi gen ifadesi verilerinin güvenilirliğinin düşük olmasından dolayı tek başına bu verilerden biyolojik olaylar hakkında sebep-sonuç ilişkisine yönelik öngörü yapmak da güvenilir değildir. Bu nedenle genler hakkında başka biyoenformatik bilgilerinin (örneğin işlev bilgisi, ileti yolu analizi, dizge analizi vb.) kullanılması gerekmektedir. Bu çalışmanın ikinci aşamasında, kısa zamanlı mikrodizi verisi üzerinde yapılan kümeleme sonuçları, farklı veri kaynaklarından gelen öznitelikler ile bütünleştirilerek ikinci bir kümeleme işlemi gerçekleştirilecektir. Yani, değişik kaynaklardan gelen bilgiler nicemlendirilerek bir gen ile ilgili tüm bilgiler bir vektör halinde gösterilecek yada iki gen arasındaki uzaklık bilgisine dönüştürülecek ve böylece geleneksel kümeleme algoritmaları ile analiz edilebilecektir. Yazında şimdiye kadar yapılmış tüm çalışmalarda, sadece mikrodizi verisine ve tek bir tip çizge (GO yada KEGG) içindeki biyolojik işlev bilgilerine dayanarak kümeleme işlemi uygulanmıştır. Fakat bu çalışma kapsamında, birden fazla mikrodizi gen ifade verisi, yine birden fazla ek bilgi kaynağından gelen öznitelikler kullanılarak kümeleme işlemine tabi tutulacaktır. Bu amaçla, belirtilen şu veri kaynaklarını kullanmayı planlıyoruz: mikrodizi gen ifadesi, gen ontoloji (GO) bilgisi, KEGG ileti yolu bilgisi, yazından gelen metinsel bilgi, dizi benzerlik puanı, protein etkileşim ağları, protein aile ve alan bilgisi.

Çalışmamızda, hem genel erişime açık mikrodizi verileri (GEO veri tabanı) hem de kendi laboratuarlarımızda gerçekleştirdiğimiz mikrodizi deneylerinden elde edilen özgün veriler kullanılacaktır [12]. GEO veri tabanından elde edilen gen ifadelerini önem sıralarına göre derecelendirmek için kendi geliştirdiğimiz yöntem kullanılacaktır [13]. GO veri tabanından, genlere ait biyolojik işlev ve moleküler süreç bilgisi alınacaktır [14]. GO içinde yer almayan, yani işlevi bilinmeyen genler için kendi geliştirdiğimiz SPMap öngörü yöntemi kullanılacaktır [15, 16]. KEGG ileti yolundan ise, herhangi bir proteinin belirli bir ileti yolunda bulunup bulunmadığına bakılarak, bulunma bilgisi elde edilecektir [17]. Yazında, üzerinde çalışma yapılmış tüm genlere ait en geniş metinsel bilgi MEDLINE makale özetlerinde yer almaktadır ve bunları tarayabilmek, içlerinde geçen ilgili gen isimlerini ve terimleri çıkartabilmek amacıyla, TXTGate uygulaması kullanılacaktır [18]. Ayrıca yine bu makale özetlerinden protein etkileşim bilgilerine ulaşmak için de iHOP uygulamasını kullanmayı planlıyoruz [19]. Her bir gene, dizi olarak en çok benzeyen diğer genler, BLAST yöntemiyle tespit edilecektir [20]. BLAST, verilen diziye en çok benzeyen diğer dizileri e-puanına göre sıralar ve bu puan benzerlik ölçütlerimizden birisi olacaktır. Mikrodizi veri setimizde yer alan genlerden sentezlenen proteinlerin aynı etkileşim ağı içinde yer alıp almadığı ve birbirlerine olan benzerlikleri, UniHi (Unified Human Interactome) veri tabanı kullanılarak araştırılacaktır [21]. Aynı etkileşim ağında bulunan proteinlerin benzerlik puanını hesaplamak için "difüzyon çekirdek" yöntemi kullanılacaktır [22]. Veri setimizdeki genlerden sentezlenen proteinlerin, aile bilgileri InterPro veri tabanından elde edilecektir [23]. Benzer işleve sahip proteinlere yüksek bir benzerlik puanı verilecektir.

Farklı veri kaynaklarından gelecek olan bu öznitelikler, gen çiftleri esas alınarak uzaklık matrisi içinde birleştirilecektir. Farklı türdeki öznitelikleri birleştirmek için çeşitli yöntemler bulunmaktadır. Kullanılan her veri kaynağına eşit katsayı verilebileceği gibi, verilerin önemine göre farklı değerde katsayılar verilerek doğrusal bir kombinasyonu da alınabilir (denklem 4).

$$d(x,y) = \sum_{i=1}^{n} k_i d^i(x,y) \quad (4)$$

Buradaki $d(x,y)$, gen x ve y arasındaki genel uzaklık miktarı olup, farklı özniteliklerin toplanmasıyla elde edilir; $d^i(x,y)$ ise veri kaynağı i'den gelen x ve y arasındaki uzaklık bilgisini gösterir; k_i ise veri kaynağı i'ye, kümele işlemindeki biyolojik önemine göre verilecek katsayıdır.

Bütünleştirilen veriler üzerinde yeni bir kümeleme işlemini uygulamak amacıyla hiyerarşik, uzaklık yada çizge tabanlı herhangi bir kümeleme yöntemi uygulanabilir. Bu amaçla çalışmamızda, izgesel, çekirdek, hiyerarşik ve k-orta değer kümeleme yöntemlerinden bir yada birkaçı uygulanacaktır.

5. Sonuçlar

Gerçekleştirilen bu çalışmada, kısa zamanlı mikrodizi serilerinin analizi ve otomatik olarak biyolojik anlamlandırması için, bir araç geliştirilmesi hedeflenmektedir. Çalışmanın ilk safhasında geliştirilen melez kümeleme yöntemi, kısa zamanlı mikrodizi deneylerinin analizi konusunda ümit verici sonuçlar ortaya çıkarmıştır. Şimdiye kadar geliştirilmiş çoğu mikrodizi analiz aracında, ayırt edici özellikteki 2000-4000 uzunluğundaki gen listelerinin biyolojik açıdan sebep-sonuç

ilişkilendirilmesi için, ikincil araçların ya da veri tabanlarının kullanılmasını gerektirmektedir. Bu nedenle, bu çalışmamız sonunda geliştireceğimiz kısa zamanlı mikrodizi serilerine özgün analiz aracı, ayırt edici gen örüntülerini otomatik olarak biyolojik bilgilerle zenginleştirip anlamlandıracağı için, kısa süreli verilerin analizinin yanı sıra ayrı bir özgün değer taşımaktadır.

6. Kaynakça

[1] J. Ernst, G. J. Nau, and Z. Bar-Joseph, "Clustering short time series gene expression data," *Bioinformatics*, vol. 21, pp. i159–i168, 2005.

[2] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, "Cluster analysis of gene expression dynamics," *Proc Natl Acad Sci USA*, vol. 99, pp. 9121–9126, 2002.

[3] C. S. Moller-Levet, F. Klawonn, K.-H. Cho, H. Yin, and O. Wolkenhauer, "Clustering of unevenly sampled gene expression time series data," *Fuzzy Sets and Sys.*, vol. 152, pp. 49–66, May 2005.

[4] A. Schliep, A. Schonhuth, and C. Steinhoff, "Using hidden markov models to analyze gene expression time course data," *Bioinformatics*, vol. 19, no. Suppl. 1, pp. i255–i263, 2003.

[5] X. Ji, J. Li-Ling, and Z. Sun, "Mining gene expression data using a novel approach based on hidden markov models," *FEBS Letters*, vol. 542, no. 1, pp. 125–131, 2003.

[6] Y. Zeng and J. Garcia-Frias, "A novel hmm-based clustering algorithm for the analysis of gene expression time-course data," *Comput. Stat. and Data Anal.*, vol. 50, no. 9, pp. 2472–2494, 2006.

[7] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer, "Co-clustering of biological networks and gene expression data," *Bioinformatics*, vol. 18, pp. S145–S154, 2002.

[8] J. Cheng, M. Cline, J. Martin, D. Finkelstein, T. Awad, D. Kulp, and M. A. Siani-Rose, "A knowledge-based clustering algorithm driven by gene ontology," *J. Biopharm. Stat.*, vol. 14, pp. 687–700, Aug 2004.

[9] N. Speer, C. Spieth, and A. Zell, "A memetic clustering algorithm for the functional partition of genes based on the gene ontology," pp. 252–259, Proc. of IEEE Sym. on Comp. Intel. in Bioinf. and Comp. Bio., 2004.

[10] M. Brameier and C. Wiuf, "Co-clustering and visualization of gene expression data and gene ontology terms for saccharomyces cerevisiae using self-organizing maps," *J. Biomedical Infor.*, vol. 40, no. 2, pp. 160–173, 2007.

[11] Z. Sokmen, M. Ozturk, V. Atalay, and R. Cetin-Atalay, "A hybrid method for the identification of expression patterns from microarray data," 15th Inter. Conf. on Intel. Sys. for Mol. Bio.(ISMB) and 6th Euro. Conf. on Comp. Bio.(ECCB), July 21-25, 2007.

[12] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W.-C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar, "Ncbi geo: mining tens of millions of expression profiles-database and tools update," *Nuc. Acid. Res.*, vol. 35, pp. D760–D765, 2007.

[13] L. Carkacioglu, T. Can, O. Konu, V. Atalay, and R. Cetin-Atalay, "Expression pattern analysis of housekeeping genes across large number of microarray experiments," 6th Euro. Conf. on Comp. Bio.(ECCB), Sept. 10-13, 2006.

[14] G. O. Consortium, "The gene ontology (go) database and informatics resource," *Nuc. Acid. Res.*, vol. 32, pp. D258–D261, 2004.

[15] O. Sarac, O. Gursoy-Yuzugullu, R. Cetin-Atalay, and V. Atalay, "Protein function annotation by subsequence based feature map," AFP and SIG meeting in ISMB-ECCB, July 2007.

[16] O. Sarac, O. Gursoy-Yuzugullu, R. Cetin-Atalay, and V. Atalay, "Subsequence based feature map for protein function classification," *Journal of Comput. Biology and Chem.*, to appear.

[17] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, "From genomics to chemical genomics: new developments in kegg," *Nucleic Acids Res*, vol. 34, pp. D354–357, 2006.

[18] P. Glenisson, B. Coessens, S. V. Vooren, J. Mathys, Y. Moreau, and B. D. Moor, "Txtgate: Profiling gene groups with text-based information," *Genome Biology*, vol. 5(6), pp. 1–12, 2004.

[19] R. Hoffmann and A. Valencia, "A gene network for navigating the literature," *Nature Genetics*, vol. 36, p. 664, 2004. http://www.ihop-net.org/.

[20] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, 1990.

[21] G. Chaurasia, Y. Iqbal, C. Hanig, H. Herzel, E. E. Wanker, and M. E. Futschik, "Unihi: an entry gate to the human protein interactome," *Nucleic Acids Res.*, vol. 35, pp. D590–D594, 2007.

[22] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete input," pp. 315–322, In Proceedings of International Conference on Machine Learning, Morgan Kaufmann Press, 2002.

[23] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley, E. Courcelle, U. Das, L. Daugherty, M. Dibley, R. Finn, W. Fleischmann, J. Gough, D. Haft, N. Hulo, S. Hunter, D. Kahn, A. Kanapin, A. Kejariwal, A. Labarga, P. S. Langendijk-Genevaux, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, A. N. Nikolskaya, S. Orchard, C. Orengo, R. Petryszak, J. D. Selengut, C. J. A. Sigrist, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats, "New developments in the interpro database," *Nucleic Acids Res*, vol. 35, pp. D224–D228, 2007.