BÜTÜNLEŞTİRİCİ MODÜL AĞLARIYLA GEN DÜZENLEME ANALİZİ

GENE REGULATION ANALYSIS WITH INTEGRATIVE MODULE NETWORKS

GİRAY SERCAN ÖZCAN

Başkent Üniversitesi
Lisansüstü Eğitim Öğretim ve Sınav Yönetmeliğinin
BİLGİSAYAR Mühendisliği Anabilim Dalı İçin Öngördüğü
YÜKSEK LİSANS TEZİ
olarak hazırlanmıştır.

2013

Bütünleştirici Modül Ağlarıyla Gen Düzenleme Analizi başlıklı bu çalışma, jürimiz tarafından, 27/01/2014 tarihinde, **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI'nda YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Başkan (Danışman) : Doç. Dr.Hasan OĞUL

Üye : Yrd. Doç. Dr. Mustafa SERT

Üye : Yrd. Doç. Dr. Yunus Kasım TERZİ

ONAY

..../..../.....

Prof. Dr. Emin AKATA

Fen Bilimleri Enstitüsü Müdürü

TEŞEKKÜR

Yazar, bu çalışmanın gerçekleşmesinde katkılarından dolayı, aşağıda adı geçen kişi ve kuruluşlara içtenlikle teşekkür eder.

Sayın Doç. Dr. Hasan OĞUL'a (tez danışmanı), çalışmanın sonuca ulaştırılmasında ve karşılaşılan güçlüklerin aşılmasında her zaman yardımcı ve yol gösterici olduğu için…

Çok değerli aileme her zaman yanımda oldukları için...

Bu tez çalışması TÜBİTAK tarafından 110E160 nolu proje ve Başkent Üniversitesi tarafından BA12/FM-10 nolu proje kapsamında desteklenmiştir.

ÖZ

BÜTÜNLEŞTİRİCİ MODÜL AĞLARIYLA GEN DÜZENLEME ANALİZİ

Giray Sercan ÖZCAN

Başkent Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Gen düzenlemesi karmaşık bir biyolojik olgudur. Bu sürecin güvenilir bir analizi, çok sayıda veri kaynağının kullanımını gerektirir. Bu tezde, Bayes modül ağları kullanılarak transkripsiyon sırası ve transkripsiyon sonrası gen düzenlemesinin aynı anda modellenmesi için bir yaklaşım sunulmaktadır. Model mRNA, mikroRNA ve transkripsiyon faktörlerinin birlikte düzenlenen elemanlarına ek olarak düşük seviyeli düzenlenme devrelerinin üretimi için mRNA ve mikroRNA ifade ve dizilim bilgisinin eşleştirilmiş örneklerini kullanır. Gerçek kanser veri seti üzerinde yapılan deneylerde, biyolojik olarak anlamlı birçok küme ve anlaşılabilir motifler elde edilmiştir. Sonuçlar, bazı test edilebilir biyolojik hipotezler üretilmesini de sağlamıştır.

ANAHTAR SÖZCÜKLER: gen düzenlenmesi, mikroRNA, transkripsiyon faktörü, bayes ağlar, veri birleştirme, modül ağlar, düzenleme ağları.
Danışman: Doç.Dr. Hasan OĞUL, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

ABSTRACT

GENE REGULATION ANALYSIS WITH INTEGRATIVE MODULE NETWORKS

Giray Sercan ÖZCAN

Baskent University Institude of Science and Engineering
Department of Computer Engineering

Gene regulation is a complex biological phenomenon. A reliable analysis of this process requires the integration of several data sources in a rigorous pipeline. Here, we propose an approach for simultaneous modeling of transcriptional and post-transcriptional gene regulation over a Bayesian module network. The framework uses paired samples of mRNA and microRNA expressions and their sequence data to produce low-level regulatory circuits in addition to the co-regulated entities of mRNAs, microRNAs and transcription factors. The experiments performed on a real cancer dataset reveal that several biologically meaningful clusters and motifs can be inferred. The results lead to the generation of some testable biological hypotheses.

KEYWORDS: Gene regulation, microRNA, transcription factor, bayes network, data integration, module network, regulation network.
Advisor: Assoc.Prof.Dr. Hasan OĞUL, Baskent University, Computer Engineering Department.

İÇİNDEKİLER LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR LİSTESİ

DNA	Deoksiribonükleik asit
RNA	Ribonükleik asit
mRNA	mesajcı RNA
tRNA	Taşıyıcı RNA
rRNA	Ribozomal RNA
cDNA	Bütünleyici DNA
miRNA	Mikro RNA
GO	Gen Ontolojisi
BAP	Bilimsel Araştırma Projesi
TÜBİTAK	Türkiye Bilimsel ve Teknik Araştırma Kurumu

1. GİRİŞ

Genler, hücrenin DNA (Deoksiribonükleik asit)'sında bulunan, canlı bireylerin kalıtsal karakterlerini taşıyıp ortaya çıkışını sağlayan ve bu kalıtsal karakterleri nesilden nesile aktaran kalıtım faktörleridir. Yaratımın şifrelerini taşıyan küçük biyolojik yapı taşlarıdır. Gen ifade verilerinin analizi, bu şifreleri çözebilmek açısından büyük önem taşır. Türler hakkında detaylı bilgi edinme, türler arası benzerlikleri ortaya çıkarıp hangi türün hangi türe evrildiğine yanıt arama, çeşitli hastalıkların sebebini araştırıp biyologlara test edilebilir hipotezler sunma amacıyla sıkça kullanılan bir yöntem haline gelmiştir. Bu amaçlar doğrultusunda çok sayıda gen ifade verisi analizi yöntemi geliştirilmiştir.

Bu tezin motivasyonu benzer görevleri yapan genleri ve bu genlerin düzenleyicilerini bulmaktır. Bu amaçla biyologlar için gerekli yazılımsal ve matematiksel işlemlerden soyutlanmış araç geliştirilmiştir. Genlerin etkileşimlerini bulan araçlar olmasına rağmen uygulanan yöntem bakımından bu çalışma bir farklılık yaratmaktadır. Geliştirilen araç belli bir veri seti üzerinde test edilmiş ve çeşitli sonuçlar alınmıştır. Elde edilen analiz sonuçları biyolojik anlamda tahmin yeterliliği açısından değerlendirilmiştir. Çeşitli veri setlerinden alınan farklı türdeki verilerle yapılan bu çalışma, bu tür verileri bir araya getiren ilk çalışma olması nedeniyle kanser araştırmaları ve ilaç tasarımı gibi biyotıp çalışmalarına katkı sağlayacaktır.

Tez en genel tanımıyla, biyolojik bazı süreçlerin anlaşılabilmesi amacıyla çeşitli veri setlerinden farklı türdeki verilerin alınıp, bu verilerden gen etkileşimlerini ve benzer görevleri yapan genleri listelemeyi hedefleyen bir biyobilişim çalışmasıdır.

Bu tez yedi bölümden oluşmaktadır: Birinci bölümde tezin motivasyonundan, katkılarından bahsedilmiş ve temel bilgiler verilmiştir. İkinci bölümde önceki çalışmalar hakkında bilgi verilirken, üçüncü bölümde yöntemlerden bahsedilmiştir. Dördüncü bölümde geliştirilen araç detaylı olarak anlatılırken, beşinci bölümde kullanılan veri kümelerine değinilmiş, altıncı bölümde sonuçlar verilmiştir, yedinci bölümdeyse tartışma ve gelecek çalışmalardan bahsedilmiştir.

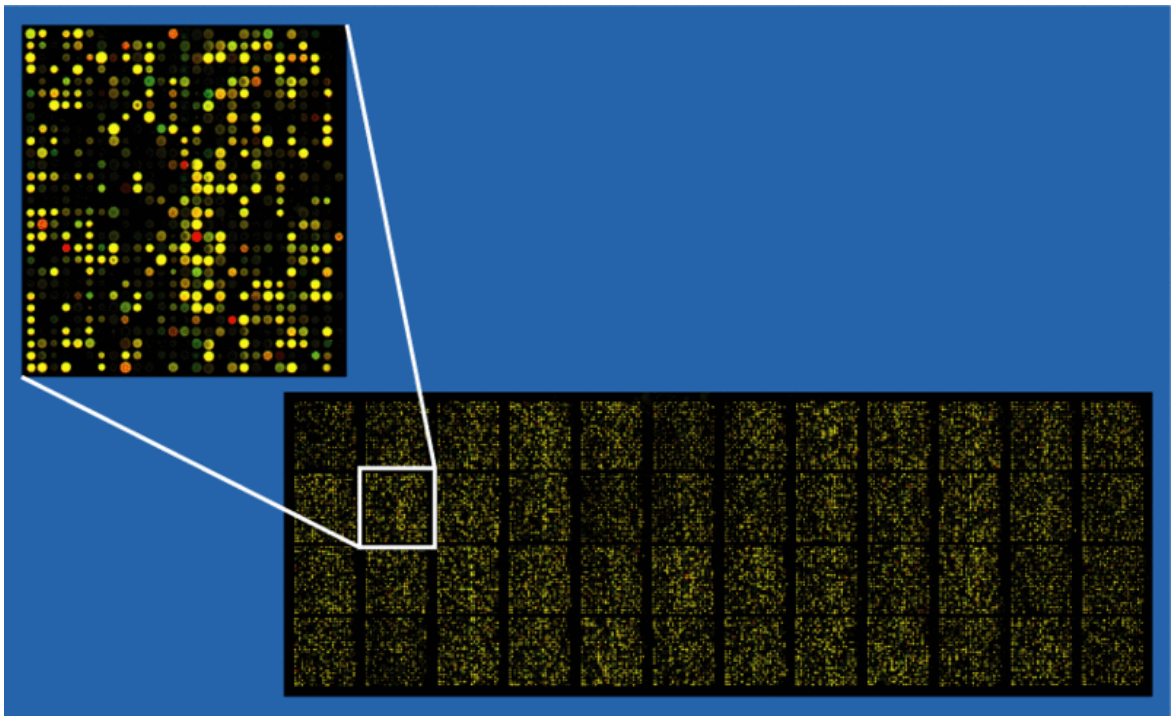
1.1. Gen İfadesi (Gene Expression)

İnsan vücudundaki hücrelerin hepsi aynı genetik materyali içerse de her hücrede aynı genler etkin değildir. Hangi genin aktif olup hangisinin olmadığı bilgisi biyologlara, normalde bu hücrelerin nasıl işlediği ve bazı genler doğru çalışmadığı zaman hücrenin bundan nasıl etkileneceği bilgisini vermektedir. Bu aktiflik bilgisine hücrenin gen ifadesi denmektedir. Geçmişte biyologlar birkaç genin gen ifade verisini aynı anda ölçebilirken, DNA (Deoksiribonükleik asit) mikroçip teknolojisinin gelişimiyle binlerce genin gen ifade verisi eşzamanlı olarak ölçülebilmektedir.

Mikroçip teknolojisi ve bu verilerin analizi araştırmacılara kalp rahatsızlıkları da dahil olmak üzere birçok farklı hastalık, ruhsal rahatsızlıklar ve bulaşıcı hastalıklar hakkında bilgi vermektedir ve hastalıklar oluşmadan önleyici önlemler alma imkanını vermektedir.

Gen ifadesi tüm canlılarda kullanılır: ökaryotlar (çok hücreli yapılarda dahil), prokaryotlar (bakteri ve arkea) ve virüsler. Yaşam için gerekli olan makro moleküler yapıların üretimi için gen ifadelerini kullanırlar. Transkripsiyon, RNA yapıştırma, çeviri ve proteinin translasyon sonrası değişikliği olmak üzere gen ifadesi süreci birkaç aşamada ayarlanabilir.

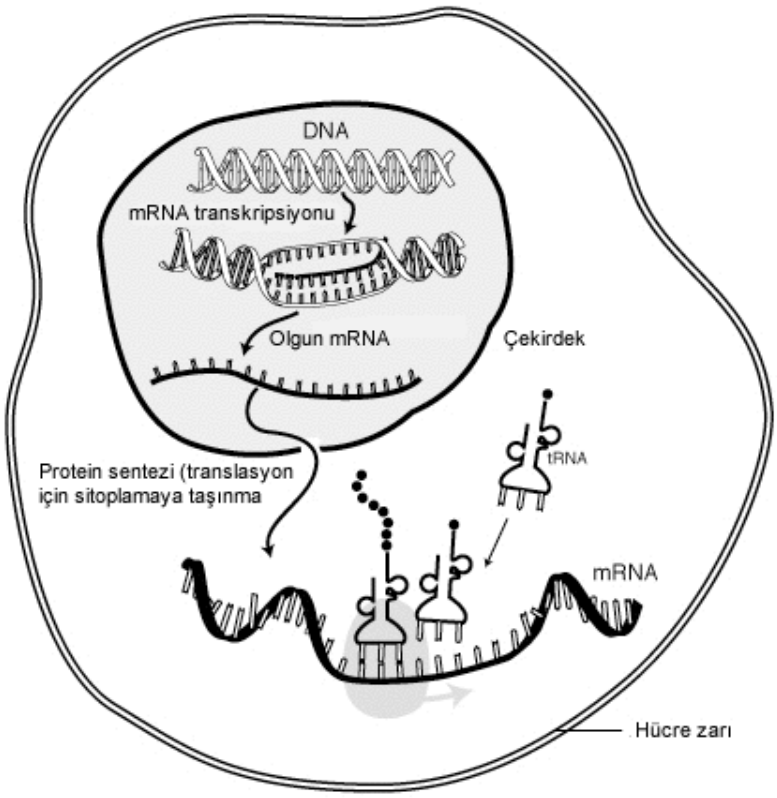
DNA mikroçipleri, bir mikroskop lamı üzerinde gen dizilimlerinin yüzlercesinin ya da binlercesinin miktarını ayarlayabilen robotik makineler tarafından üretilir. Gen aktive edildiğinde hücresel makine o genin belirli kısımlarını kopyalamaya başlar. Elde edilen ürün, proteinleri üretmek için hücrenin şablonları olan mesajcı RNA (mRNA)'dır. Hücre tarafından üretilen mRNA, tamamlayıcı olması nedeniyle kopyalandığı DNA sarmalının orijinal kısmına bağlanacaktır. Şekil 1.1'de örnek olarak bir mikroçip deneyi görülmektedir.



Şekil 1.1 Örnek mikroçip deneyi görünümü [1]

1.2. Mesajcı RNA (mRNA)

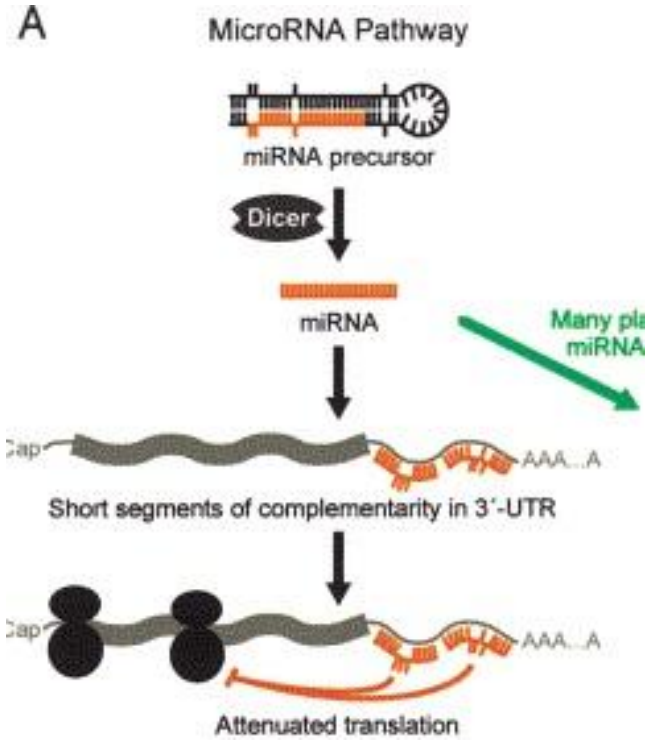
Mesajcı RNA, genetik bilgiyi DNA'dan proteinlerin yapılarında bulunan amino asitlerin birbirine bağlandığı ribozoma taşıyan RNA molekülünün bir türüdür. Transkripsiyon, hücrede DNA'dan RNA oluşması sürecidir. Polimeraz enzimiyle transkripsiyon sırasında DNA'dan sentezlenmiş olan birincil transkript mRNA (pre-mRNA) hücrede işlenir ve olgun mRNA oluşur. Olgun mRNA ribozomda amino asit polimerlerine çevrilir. DNA'da olduğu gibi mRNA da nükleotidlerin dizilimi şeklinde oluşur. Nükleik asitlerin amino asit dizilerine karşılık gelen bölgelerindeki her üç baz, proteindeki bir amino asite karşılık gelir. Bu üçlülere kodon denir. Her kodon farklı bir amino asit içindir. Durdurma kodonu protein sentezini bitirir. Bu kodonların amino asitlere çevrimi süreci iki RNA türü daha gerektirir: taşıyıcı RNA (tRNA) kodonu tanımlar ve ilgili amino asidi sağlar, ribozomal RNA (rRNA) ise ribozomun protein üretimi sırasında katalizör görevini görür. Şekil 1.2'de bir mRNA molekülünün çalışması özetlenmiştir.



Şekil 1.2 Örnek mRNA çalışma mekanizması [2]

1.3. MikroRNA (miRNA)

MikroRNA, protein üretilmeyen, bitkilerde ve hayvanlarda bulunan, transkripsiyon sırasında ve transkripsiyon sonrasında gen ifadesinin düzenlenmesinde görev yapan küçük RNA molekülleridir (22 nm). DNA'lar tarafından kodlanan miRNA'lar, mRNA moleküllerinin dizilimleriyle eşleşerek o mRNA'nın gen ifadesinde yükselmelere ya da alçalmalara sebep olabilir. İnsan genomu 1000'den fazla miRNA kodlayabilir. Bunlar memeli genlerinin %60'ını ve çoğu insan hücre tiplerinde fazlalıkta olan mRNA'ları hedefleyebilir. miRNA'lar ökaryotik organizmalarda fazlalıkla görülür ve gen düzenlemesinin hayati ve evrimsel antik bileşeni olduğu düşünülmektedir. miRNA temel bileşenleri bitkilerde ve hayvanlarda değişmeden korunmuştur fakat iki türdeki miRNA repertuvarı fonksiyonlar farklı olacak biçimde bağımsız evrimleşmişlerdir. Bu yüzden çalışmada tamamen insan veri setlerinden alınan verilerde çalışılmıştır.



Şekil 1.3 miRNA fonksiyonu için model

Şekil 1.3'te bir miRNA'nın çalışma modeli görülebilir. miRNA, mRNA'nın belli bir kısmına tutunarak o mRNA'nın gen ifadesinin değişmesine sebep olmaktadır.

1.4. Trankripsiyon Faktörü (TF)

Transkripsiyon faktörü (TF), transkripsiyon sırasında belli DNA dizilimlerine bağlanan proteindir. Bu yüzden DNA'dan mRNA'ya genetik bilginin akışı sırasında görev yapar ve protein üretiminde rol oynar. TF'ler bu fonksiyonu tek başına yapabildikleri gibi diğer proteinlerle birlikte de yapabilirler. Bunu belli genlere RNA polimeraz enzimi (DNA'dan RNA'ya genetik bilginin geçişi sırasında düzenleyici) erişimini yükselterek veya engelleyerek gerçekleştirirler. Kısacası, TF'ler mRNA'ların gen ifadelerini yükseltebilir veya düşürebilir.

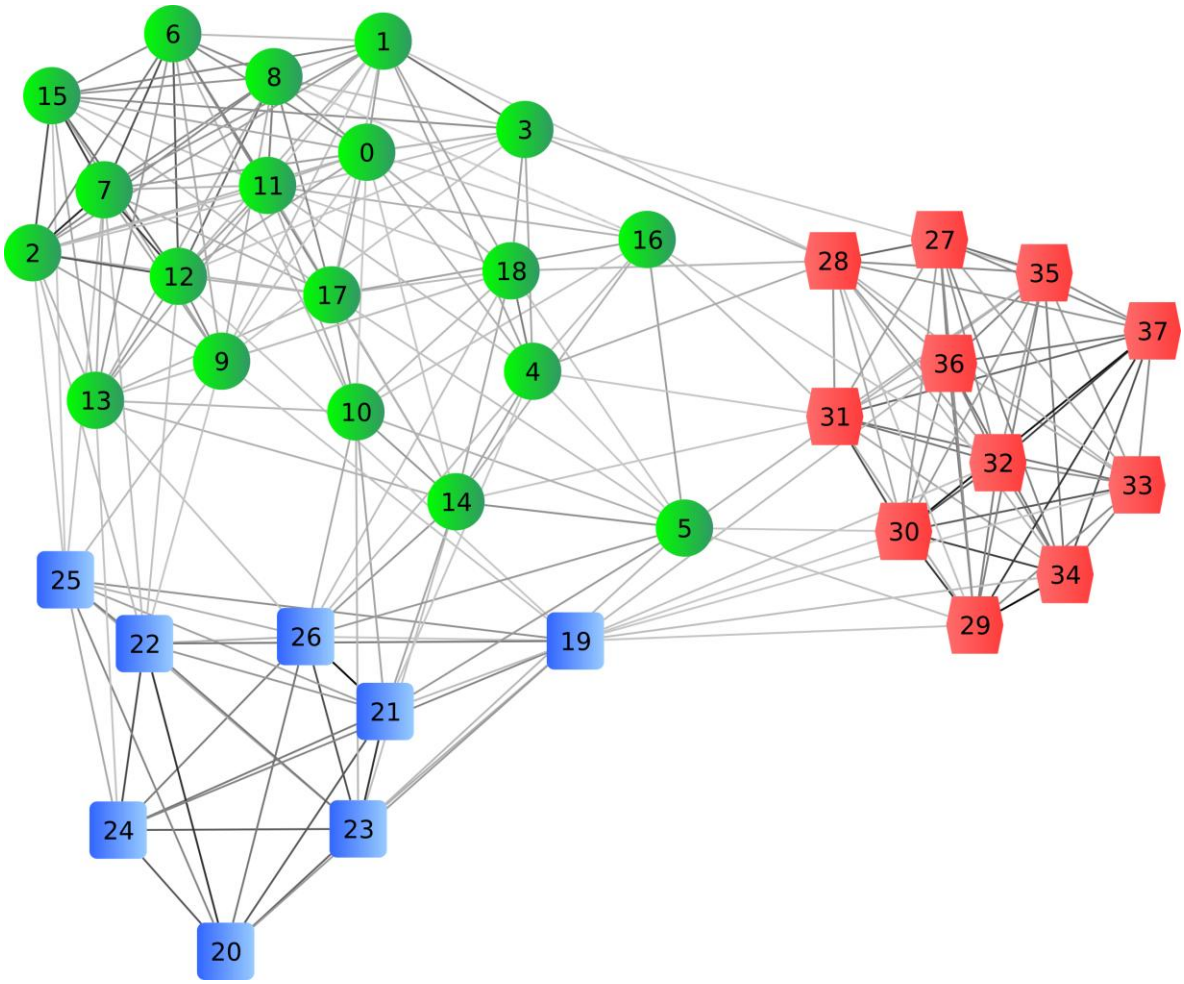
TF'lerin tanımlayıcı özelliği bir ya da daha fazla DNA bağlanma alanı içermeleridir. Bu alanlar düzenlenecek genlere komşu DNA'nın belli dizilerine bağlanırlar. TF'ler gen ifadesinin düzenlenmesinde temel bileşendir ve sonuç olarak yaşayan her canlıda bulunmaktadır. Canlıda bulunan TF sayısı canlının kompleks olup

olmadığına göre değişmektedir, büyük genom, gen başına daha fazla sayıda TF içermektedir. İnsan genomunda DNA bağlanma alanı içeren yaklaşık 2600 protein vardır ve bunların çoğu TF olarak kabul edilmektedir.

1.5. Kümeleme (Clustering) Analizi

Kümeleme analizi, aynı gruptaki nesnelerin başka gruptakilere göre daha fazla benzerlik göstermesini sağlayacak şekilde nesne kümelerini gruplamaktır. Keşfedimsel veri madendiciliğinin ana görevidir. Aynı zamanda makine öğrenme, örüntü tanıma, resim analizi, bilgi çıkarma ve biyobilişim olmak üzere birçok istatistiksel veri analizi alanlarında kullanılır.

Kümeleme analizi belirli bir algoritma değildir. Daha çok çözülecek genel bir görevdir. Bu, kümenin nelerden oluşacağına ve etkin biçimde kümelerin nasıl bulunacağına bakılarak seçilmiş çeşitli algoritmalarla gerçekleştirilebilir. Popüler kümeleme yöntemleri küme elemanları arasında grafiksel olarak küçük mesafe olan, veri uzayının, aralıkların veya belirli istatistiksel dağılımların yoğun alanlarını içeren kümeleme algoritmalarını uygulayan yöntemlerdir. Kümeleme, bu nedenle çok amaçlı optimizasyon problemi olarak formüle edilebilir. Sonuçların kullanım amacı ve tekil veri setine göre uygun kümeleme algoritması ve parametre ayarlaması seçilir. Kümele analizi, otomatik bir görev değildir. Deneme ve başarısızlıkla bilgi keşfetme, etkileşimli çok amaçlı optimizasyonu içeren yinelemeli bir süreçtir. Şekil 1.4'te kümeleme yapılması sonucu elde edilen üç farklı grup görülmektedir. Her bir grup farklı renklerle gösterilmiştir.



Şekil 1.4 Kümeleme analizinde üç farklı grup

Kümeleme teriminin yanı sıra benzer anlamlara gelen terimlerde vardır. Bunlar otomatik sınıflandırma, sayısal taksonomi ve tipolojik analizlerdir. Genel farklılık sonuçların kullanımında ortaya çıkmaktadır. Veri madenciliğinde sonuç gruplar önemliyken, otomatik sınıflandırmada grupların ayırt edici özelliği önemlidir. Bu, veri madenciliği ve makine öğrenme alanlarından gelen araştırmacılar arasında yanlış anlaşılmalara sebebiyet vermektedir. Aynı terimleri ve aynı algoritmaları kullanmalarına rağmen farklı amaçları vardır.

Biyobilişimdeyse kümeleme gen ifade verilerinin analizinde en sık kullanılan yöntemlerden birisidir. Kümelemenin temel mantığı benzer ifade örüntülerine sahip genleri (benzer transkripsiyon faktör (TF) ve miRNA tarafından aktive edilen, benzer biyolojik işlemlerde yer alan) kümelemek ve ilişkilendirmektir.

Kümeleme algoritmaları çok fazladır. Her bir sorun için fazla sayıda kümeleme algoritması vardır. Bunun nedeni kümelemenin neye göre yapıldığının tam olarak tanımlanamamasıdır. Sınıflandırmada olduğu gibi öğretme veri kümesi yoktur. Kümeleme kendi kendine öğrenmeye çalışır. Ortak bir payda vardır: veri nesneleri grupları. Farklı araştırmacılar farklı kümeleme modelleri geliştirmiştir ve bu modellerin her birisi içinde farklı algoritmalar vardır. Farklı algoritmalarla bulunan kümeler özelliklerine göre değişiklik gösterir. Algoritmalar arasındaki farkları anlamak için kümeleme modellerini anlamak işin temelidir.

Nesnel olarak “doğru” kümeleme algoritması yoktur. Kullanıcı, gözüne en iyi gelen, amacına en uygun algoritmayı uygulamalıdır. Bir kümeleme modelinin seçilmesi için matematiksel neden yoksa belirli bir sorun için en uygun kümeleme algoritması deneysel olarak seçilmelidir. Bir model için hazırlanmış algoritmanın farklı model içeren bir veri setindeki kümeleri büyük olasılıkla bulması çok zordur.

1.6. Gen İfadesinin Düzenlenmesi (Regulation of Gene Expression)

Belirli gen ürünlerinin üretimini artırmak veya düşürmek için hücrenin birçok mekanizması vardır ve gen düzenlemesi olarak adlandırılır. Gelişmiş gen düzenlemesi mekanizmaları biyolojide bolca görülmektedir. Örnek olarak çevresel uyarılara yanıt vermek, yeni yemek kaynaklarına adapte olmak verilebilir. Transkripsiyon işleminin başlangıcından RNA işlemine ve çevirim sonrası proteinin üzerindeki değişikliklere kadar gen düzenlemesinin hemen hemen her aşaması ayarlanabilir.

Bir hücrenin çok yönlülüğünü ve uyumunu arttırdığından gen düzenlemesi tek hücreli canlılar ve virüsler için önemlidir. Çünkü protein ihtiyacı olduğunda hücrenin bunu üretebilmesi gereklidir. Gen düzenlemesi yoluyla bunu yapar. Ayrıca çok hücreli organizmalarda, gen düzenlemesi hücresel farklılığı sağlar. Bu da farklı gen düzenlemesi profillerini içeren farklı hücre tipleri oluşmasına sebep olur. Hücrelerin tamamı aynı DNA dizilimine sahip olsa da gen düzenlemesi, farklı hücrelerin kendi fonksiyonlarına uygun nano ölçümsel yapıların gelişmesini sağlar.

Gen düzenlemesi hücreye, hücrenin yapısı ve fonksiyonları üzerinde kontrol sağlar ve hücresel farklılık, morfogenez, organizmanın çok yönlülüğü ve adaptasyonu için bir temeldir. Gen düzenlemesi aynı zamanda evrimsel değişikliğin nedeni olabilir çünkü gen ifadesinin zamanlaması, konumu ve miktarı kontrol edilirse hücrede veya çok hücreli yapılarda genlerin fonksiyonlarını çok büyük şekilde etkileyebilir.

Genotip, organizmanın genetik yapısına verilen addır. Fenotip ise, genetik ve çevresel etkenlerin yarattığı özelliklerin canlının görünüşünde veya iç yapısında oluşturduğu değişikliktir. Genetikte, gen ifadesi hangi genotipin hangi fenotipe ne kadar yükseliş derecesi verdiğini bulmanın temelidir. DNA'da depolanan genetik kod, gen ifadesi ile yorumlanır ve ifadedeki özellikler organizmanın fenotipinde yükselmelere ya da alçalmalara sebep olabilir. Bu fenotipler genelde organizmanın şeklini kontrol eden proteinlerin senteziyle ya da organizmayı karakterize eden belirli metabolik yolları katalize eden enzimler olarak ifade edilebilir.

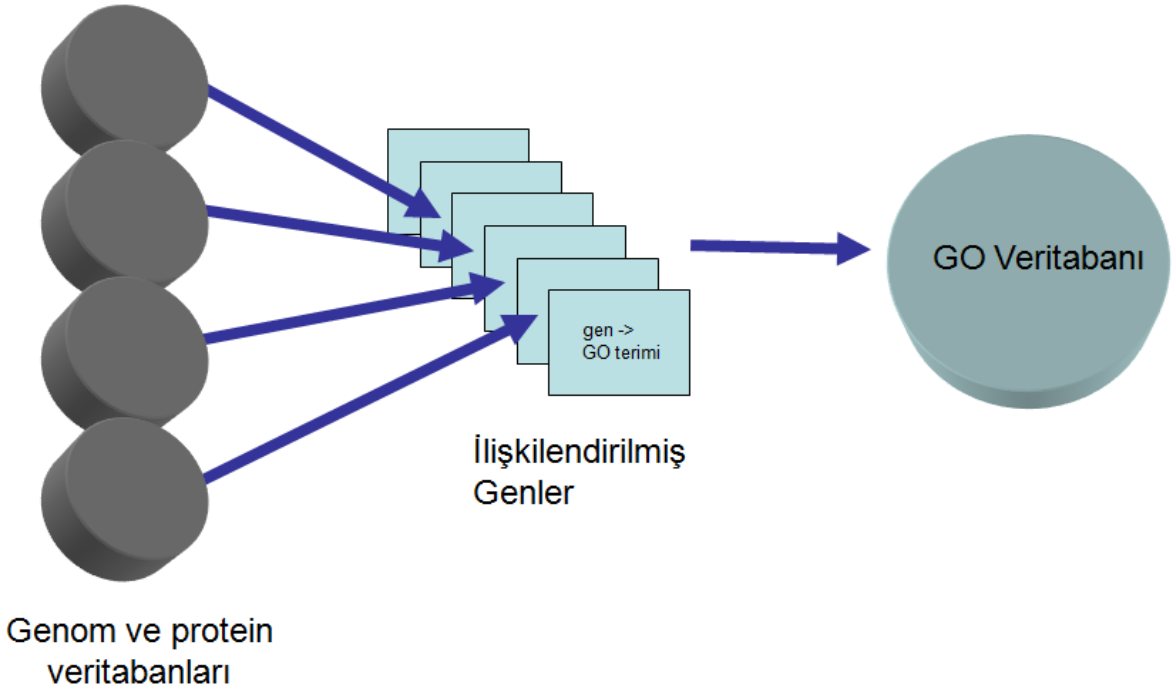
Gen, kalıtımsal bilgilerin kodlandığı DNA dizisidir. Genomik DNA'nın, anti-paralel ve ters bütünleyici şerit olmak üzere iki temel özelliği bulunur. Bu kısımların her birisi 3'UTR ve 5'UTR uçlarına sahiptir. MiRNA veya TF bu kısımlarla etkileşime girerek gen ifadesini artırabilir veya azaltabilir. Gen ifadesindeki bu değişiklik protein üretiminin artmasına ya da azalmasına sebep olur. Bu da canlıdaki değişikliklerin sebebidir.

Transkripsiyon sırasında ve transkripsiyon sonrasında olmak üzere gen ifadesi düzenlemesi iki aşamada olabilir. Bu çalışmada gen ifadesi verileri kullanılmış, dizilim verisi kullanılarak daha kesin sonuçlar alınmış, kümelenen gruplar şekilsel olarak gösterilmiş, etkileşim olasılıkları listelenmiş ve GO analizi bilgileri çıkartılmıştır. Aynı zamanda biyologlar için bir araç geliştirilmiştir. Bu araç transkripsiyon sırasında ve transkripsiyon sonrasında gen analizini eşzamanlı olarak yapabilmekte ve grupları, etkileşimleri ve GO analizi bilgilerini verebilmektedir.

1.7. GO (Gene Ontology) Analizi

Gen Ontolojisi veya GO, tüm türler arasındaki gen ve gen ürünleri niteliklerinin temsilini birleştirmek için yapılan bir biyobilişim girişimidir. Projenin amaçları arasında gen ve gen ürünleri niteliklerinin kelimelerinin bakımı ve geliştirilmesi, gen ve gen ürünlerini barındırma ve proje tarafından sağlanan verilere her yönüyle kolay erişim için gerekli araçların sağlanması sayılabilir. Kısacası, şu ana kadar araştırılan neredeyse tüm türlerin özelliklerini içeren bir veritabanıdır.

Biyologlar araştırmanın her alanındaki mevcut bilgileri ararken çok fazla zaman ve çaba harcamaktadırlar. Bu terminolojideki geniş varyasyonlarla daha da fazla zorlaşmaktadır. Hem insanlar hem de bilgisayarlar etkin bir arama yapamamaktadırlar. Örneğin antibiyotiklerin yeni hedefleri aranırken insandaki genlerden tamamen farklı dizilim ve yapılara sahip olan bakteriyel protein sentezinde rol oynayan tüm gen ürünleri bulunabilmektedir. Bir veritabanı bu moleküllerin 'çeviri' aşamasında rol aldıklarını söylerken başka bir veritabanı 'protein sentezi' ifadesini kullanabilir. İşlevsel olarak eşdeğer terimlerin bulunması insan için zordur, bilgisayar içinse daha da zordur.



Şekil 1.5 GO Veritabanı Oluşturulması Şematik Görünümü

Aramanın daha kolay yapılabilmesi için GO projesi başlatılmıştır. Proje, farklı veritabanlarındaki gen ürünlerinin açıklamalarını ve terimlerini tek bir çatı altında toplamayı hedeflemiştir. Başlangıçta birkaç veritabanından ibaret olan proje bugüne kadar bitki, hayvan ve mikrobiyal genomları içeren dünyanın büyük veri bankalarından çok sayıda veritabanını bünyesine katmıştır.

GO projesi gen ürünlerinin ilişkili biyolojik süreçleri, hücresel bileşenleri ve türden bağımsız moleküler fonksiyonlarını tanımlayan üç tane yapısal sözlük (ontolojiler) geliştirmiştir. Bu projenin üç tane amacı vardır: birincisi ontolojilerin geliştirilmesi ve devamlılığı, ikincisi ontolojiler, genler ve gen ürünleri arasında ilişkilendirme anlamına gelen gen ürünlerinin açıklamalarının yapılması, üçüncüsüyse ontolojilerin kullanımı, bakımı ve devamlılığının sağlanması için gerekli araçların teminidir.

Veritabanlarını birleştiren GO terimlerinin kullanımı tek tip sorguların yazılmasını kolaylaştırmaktadır. Kontrollü sözlükler farklı seviye sorguları yazılacak şekilde yapılandırılmıştır. Örneğin sinyal iletiminde görev alan farenin tüm gen ürünlerini görüntülemeyi veya tüm reseptör tirozin kinaz genleri üzerine yakınlaştırmayı kolaylıkla yapabilmektedir. Bu yapı aynı zamanda varlığın sahip olduğu özellik hakkındaki bilginin derinliğine göre farklı seviyelerde gen veya gen ürünlerine özellik atamaya da izin vermektedir.

Gen açıklaması, GO terimlerinin gen ürünlerine atanması işlemidir. Şekil 1.6'da görüldüğü üzere türe bağlı oluşturulan açıklama dosyalarında terim isimleri, GO ID'leri, tanımları, ontoloji türleri vb. bilgiler yer almaktadır.

id: GO:0006094	unique GO ID
name: gluconeogenesis	term name
namespace: process	ontology
def: The formation of glucose from noncarbohydrate precursors, such as pyruvate, amino acids and glycerol. [http://cancerweb.ncl.ac.uk/omd/index.html]	definition
exact_synonym: glucose biosynthesis	synonym
xref_analog: MetaCyc:GLUCONEO-PWY	database ref
is_a: GO:0006006 is_a: GO:0006092	parentage

Şekil 1.6 GO Çalışması Dosya Örneği [6]

2. ÖNCEKİ ÇALIŞMALAR

Son 50 yıldır moleküler biyolojideki araştırmalar genellikle indirgemeci düşünceler üzerine yoğunlaşmıştır. Birer birer genler veya proteinler araştırılarak organizmaların karmaşık çalışma yapıları çözülmeye çalışılmıştır. Son yıllarda moleküler biyologlar hücreyi daha farklı küresel bir bakış açısıyla görmeye başlamışlardır. Tam anlamıyla dizilimlenebilir genomların ve yüksek ölçekli fonksiyonel genomlama teknolojilerinin gelişimiyle, binlerce genin eşzamanlı gen ifade seviyeleri veya protein-DNA ilişkileri gibi moleküler özelliklerin ölçülmesi mümkün hale gelmiştir. Sonuç olarak teker teker çalışılması yerine genlerin, proteinlerin ve aralarındaki ilişkilerin biyolojik sistemler, olasılıksal dağılımlar anlamında çalışılması daha akla yatkın olmuştur. Biyobilişim bilimi böyle doğmuştur.

2.1. MikroRNA (miRNA)

miRNAlar yaklaşık 22 nt uzunluğunda, kodlanmayan fakat transkripsiyon sonrası bir seviyede gen ifadesini düzenleyen küçük RNAlardır. 1993'te Lee ve çalışma arkadaşları tarafından keşfedilmiştir [28], ancak mikroRNA terimi ilk 2001'de kullanılmaya başlamıştır [29]. Hedef olarak seçtikleri mRNAlara genellikle 3’UTR bölgesinde bağlanarak gen ifadesini etkilerler [30]. Dolayısıyla, bir miRNA’nın asıl aktifleşmesini sağlayan kendisinin ve bağlanacağı mRNA’nın dizilimidir. Bu bağlanma sonucunda miRNA, bağlandığı genin ifadesinde değişikliğe neden olur. Bu değişiklik genellikle negatif yönde (gen ifadesini baskılayıcı) olmak üzere iki yönlü de olabilmektedir [31]. miRNA ökaryotik hücrelerin normal işlevinde yer aldığı gibi, miRNA’nın bozuk çalışması da hastalığa neden olur. Birkaç miRNA ile bazı kanser tipleri arasında ilişkiler bulunmuştur. Lenfomalarda bulunan miRNA’ları özellikle çok miktarda üretmek üzere tasarlanmış farelerde 50 gün içinde kanser gelişmiş ve bunlar iki hafta ardından ölmüşlerdir [32]. Bir diğer çalışmada, hücre çoğalmasını düzenleyen E2F1 proteininin iki tip miRNA tarafından inhibe edildiği gösterilmiştir. miRNA, mRNA’ya bağlanarak gen aktivitesine etki eden proteinlerin çevrimini engellemektedir [33]. miRNA kodlayan 217 genin etkinliği ölçülerek farklı

kanser tiplerini ayırt edebilen gen ifade örüntüleri bulunmuştur. Bu çalışmayla, miRNA profillerinin kanser sınıflandırılmasında faydalı olduğu gösterilmiştir [27].

2.2. Dizilimle miRNA Hedef Tahmini

miRNAnın keşfiyle birlikte işlemsel biyolojide pek çok yeni problem ortaya çıkmıştır. miRNA genlerinin tanınması [34] ve belirli miRNA'ların hedef genlerinin tespiti [35] en çok üzerinde çalışılan konulardır. Sadece dizilim (sequence) verilerine dayalı olarak geliştirilen hedef tahmini algoritmaları, miRNA ile mRNA dizilimleri arasındaki eşleniklik veya dizilim yoluyla çıkarılan başka özellikleri kullanarak, bir potansiyel bağlanma skoru tanımlamakta ve buna bağlı olarak basit karar yapıları veya makine öğrenme yöntemleriyle sonuca ulaşmaktadır. Bu amaçla geliştirilmiş çok sayıda yöntem ve araç bulunmaktadır [36].

miRNA'nın 5' ucunda tohum (seed) diye adlandırılan bir bölge vardır. Bu bölge 1. ve 8. nükleotidler arasındadır. Hedef mRNA üzerinde bağlanma yeri (binding site) ile miRNAdaki tohum bölgesi arasındaki tamamlayıcı bir eşleşmenin hedefin seçilmesinde etkili olduğu düşünülmektedir [37]. Mevcut hedef tahmini algoritmalarının bazıları bu esas üzerine kurulmuştur [38,39]. Örneğin, TargetScan algoritması sadece miRNA tohum bölgesiyle mRNA 3'UTR dizlimi arasında bir eşleşme olup olmadığını kontrol ederek tahmin yapmaya çalışır [38]. Yakın zamanda yapılan çalışmalarda sadece tohum bölgesinde değil, olgun miRNA diziliminin diğer bölgesiyle mRNA dizilimi arasındaki tamamlayıcı eşleşmelerin de hedef tahmininde etkili olduğu gösterilmiştir [40,41]. RNA22 diye adlandırılan yöntemde miRNA diziliminin çeşitli bölgelerine göre hedef motifler tespit edilmiş ve potansiyel hedef mRNA dizilimlerinde bu motiflerin dağılımları analiz edilmiştir [42]. Dizilim eşleşmesinin yanı sıra, türler arasında korunan (conserved) dizilimlerin de hedef tahminin de etkili olduğu gösterilmiş ve bazı tahmin araçlarında bu özellik tohum eşleşmesine ek olarak kullanılmıştır [43,44]. miRNA hedef seçiminde etkili olduğu düşünülen başka bir özellik de termodinamik kararlılıktır. Potansiyel miRNA-mRNA dubleksi analiz edilerek serbest enerji hesaplanabilir. Dubleks serbest enerjisi ne kadar düşükse oluşan yapının kararlılığının ve dolayısıyla bağlanma olasılığının o kadar yüksek olduğu düşünülerek, enerji bilgisi tahmin

algoritmasında kullanılabilir [45,46]. Bazı çalışmalarda, mRNA dizilimi üzerinde bir bölgenin miRNA tarafından erişilebilirliği, ikincil yapı bilgisi kullanılarak değerlendirilmiştir [47,48]. Tahmin algoritmasında, tohum eşleşmesi, termodinamik kararlılık, korunum ve yapısal erişilebilirlik özeliklerini bütünleştiren çalışmalar da bulunmaktadır [49]. Aynı özellikleri birleştirerek tahmin yapmaya çalışan PicTar algoritması, aynı zamanda, farklı hücrelerde ifade edilen miRNAların birlikte düzenlediği genlerle ilgili bilgiyi Saklı Markov Modeli kullanarak entegre edebilmiştir [50]. miRNA hedef seçimini etkilediği düşünülen dizilime bağlı, yapısal veya termodinamik bu özellikler bazı çalışmalarda makine öğrenme yöntemleri kullanılarak eğitilmiş ve buradan edilen modellerle tahminler yapılmıştır [51,52,53]. Destek Vektör Makinaları bu amaçla, farklı özellikleri birleştiren çalışmalarda kullanılmıştır [54]. Çok sayıda çalışmaya rağmen miRNA hedeflerinin bulunmasında istenen tahmin yeterliliğine ulaşılamamış ve mevcut araçların birbirlerinden çok farklı sonuçlar ürettikleri gözlemlenmiştir [55].

2.3. Dizilim ve Gen İfadeleriyle miRNA Hedef Tahmini

miRNAların hedef olarak seçtikleri mRNA traskriptin gen ifadesi üzerinde etkili olduğu pek çok çalışmada gösterilmiştir [30, 56, 31, 57]. Bu bulgular, erişilebilir olması durumunda gen ifade verilerinin miRNA hedef tahmininde kullanılabileceği fikrini ortaya çıkarmıştır [58]. Gen ifadesinin farklı türlerde korunması hedef tahminine yardımcı olmuştur [60, 61, 62]. Bazı çalışmalarda dizilim bilgisi ile gen ifade bilgisini birleştiren miRNA hedef tahmini yöntemleri denenmiştir. Bunlardan birinde başka bir dizilim tabanlı hedef tahmin aracının ikili çıktısıyla, gen ifade profilleri olasılıksal bir model üzerinde birleştirilmiştir [59]. Başka bir çalışmada, yine dizilim tabanlı bir hedef tahmini aracıyla elde edilen miRNA-mRNA ikililerinin ifade verilerindeki ortak değişim rapor edilmiş, bu bulguyla, iki bilginin birleşimiyle miRNA fonksiyonel analizinin yapılabileceği tartışılmıştır [63]. Benzer bir çalışmada yine tahmin edilen miRNA-mRNA ikililerinden oluşan veri tabanları üzerinde, seçilen bir miRNA için olası hedef mRNAların gen zenginleştirme istatistiklerini çıkaran bir Excel araç kutusu tanıtılmıştır [64]. Web üzerinden hizmet veren başka bir araç, tahmin edilen ikililer ve gen ifade verilerini alıp GO analizlerini görsel olarak sumaktadır [65]. Daha yakın zamanda yayınlanan bir çalışmada, sadece

ifade profillerinin girdi olarak kullanıldığı destek vektör makinalarıyla bir miRNA hedef tahmini üretilmiş ve bu tahminle dizilim eşleşme skoru arasında bir uzlaşma aranmıştır [66]. Görülmektedir ki, miRNA hedef tahmini için iki tür veriyi birleştirmeye çalışan yaklaşımlardan hiçbiri bunu ortak bir model üzerinde yapmayıp, bilinen araçlarla herhangi bir tür veriyle elde edilen tahmini diğer veri türüyle ilişkilendirmekte veya iki ayrı tahmin arasında uzlaşma veya üstünlük bulmaya çalışmaktadır.

2.4. Gen İfadeleriyle miRNA Modül Analizi

miRNAlar üzerinde kuramsal, deneysel veya işlemsel çeşitli çalışmalar sürerken, ihmal edilen veya henüz yeterli bir çözüm bulunamamış olan bir konu, birlikte hareket eden miRNAların ve miRNAlar tarafından aynı anda düzenlenen gen kümelerinin bulunması olmuştur [34,67]. Yakın zamanda yapılan çalışmalarda bu konu sadece gen ifade verileri kullanılarak ele alınmıştır [68-71]. Ayrıca bu çalışmalar sadece modülleri tahmin etmekte, ağ ilişkisini belirleyememektedir.

2.5. Gen Düzenleme (Gene Regulation) Analizi

Moleküler biyolojiye yapılan yaklaşımlarda, sistemlerin amacı hücrenin fonksiyonlarını gerçekleştiren gen düzenlemelerini tersine mühendislikle bulmaktır. Özellikle deneysel verilerin çokluğundan dolayı transkripsiyonel düzenleme ağları çok büyük ilgi toplamıştır. Çeşitli çalışmalar, transkripsiyonel düzenleme ağlarına ışık tutmak için ifade verisi, kromatin immunopresipitasyon (ChIP) verisi, aktifleştirici motif verisi veya önceki fonksiyonel bilgileri (GO sınıfları [7] veya bilinen düzenleyici ağ yapıları) kullanmıştır [8-22]. Bu yöntemlerin çok büyük bir kısmı belirli ifade örüntülerindeki kontrol mantığını açığa çıkarmaya çalışmıştır. Bu analiz türleri ayrıntılı hesaplama programları gerektirmektedir. Düzenleyici ağların çıkarılması için özellikle olasılıksal grafik modelleri doğal matematiksel programlar olarak kabul edilmektedir [13]. Olasılıksal grafik modelleri Bayes ağlarını en iyi temsil eden modeldir. Model, her bir değişken (genler) için gözlemleri, sınırlı sayıdaki parent değişkenlerinin (düzenleyiciler) bir fonksiyonu olarak tanımlayan koşullu olasılık dağılımları cinsinden sistemi ifade etmektedir. Bu da gözlemlerin

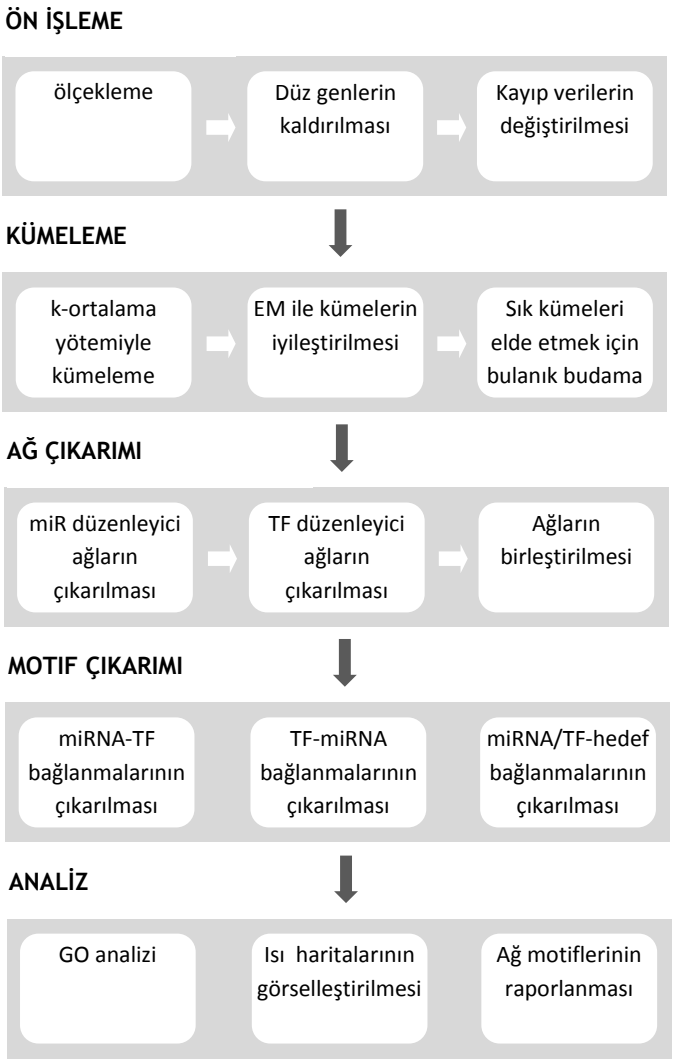
altında yatan düzenleyici ağları yeniden kurmak anlamına gelmektedir. Friedman ve diğerleri gen ifade verisinden düzenleyici ağları öğrenmek için Bayes ağları yöntemini ilk olarak kullanmışlardır [8,9]. Bu önceki çalışmalarda Bayes ağlardan elde edilen genler kendi parent genleri (düzenleyiciler) ve koşullu olasılıksal dağılımları gibi kendi düzenleme programlarıyla ilişkilendirilir. Bu yaklaşımın kısıtı, fazla sayıda yapısal özelliğin ve dağılım parametrelerinin sınırlı sayıda ifade profillerinden öğrenilmesi gerekliliğidir. Diğer bir deyişle gerçek ağın geri bulunması gerekliliği eksik kalmıştır. Bu sorun biyolojik ağların doğal modülerliklerinden yararlanan bir yolla çözülmek istenmiştir [23]. Bu yöntemde gen gruplarının aynı düzenleyiciler tarafından düzenlenebilmesi gerçeği dikkate alınmıştır. İlk olarak bu fikri Segal ve diğerleri, düzenleyici ağlar için modül ağlarını matematiksel model olarak alan bir yöntem uygulamışlardır [11,24]. Modül ağları olasılıksal grafik modelleridir. Modül, aynı koşullu dağılımları paylaşan kısacası benzer görevleri yapan gen grupları demektir. Modül ağlarda tahmin edilecek parametre sayısı tam bir Bayes ağdan çok daha az olacağından modül ağları öğrenmek için gen ifade veri kümesi yeterli büyüklükte olmalıdır [11,16,17,24].

Biyolojik olarak uyumlu düzenleyici ilişkilerini bulmakta modül ağ öğrenme algoritmalarının başarısı kanıtlanmasına [11,16,17,24] rağmen bu algoritmaların gerçek hassasiyetini ve farklı modül öğrenme stratejilerinin performansı nasıl etkilediğini ölçen sınırlı sayıda çalışma [17] bulunmaktadır. Son sorunu cevaplandırmak modül ağların geliştirilmesinde anahtar rol oynamaktadır.

Bu tez çalışmasında hem mRNA ve miRNA verilerinin birlikte kullanımı, hem de dizilim ve gen ifade verilerinin birlikte kullanımı gerçekleştirilmiştir. Bu veri bütünleştirmesi tezin yenilikçi yönünü oluşturmaktadır. Sunulan yaklaşımda diğer çalışmalardan farklı bir kümeleme algoritması kullanılmış, düzenleyici atamaları buna göre yapılmış ve dizilim bilgisinin kullanımıyla daha kesin motifler elde edilmiştir.

3. YÖNTEMLER

Geliştirilen sistem, aynı deneysel şartların eşlenmiş örneklerini içeren mRNA seti ve miRNA seti olmak üzere iki mikroçip deney sonucunu girdi olarak almaktadır. Şekil 3.1'de görülen çalışma aşamaları izlemektedir. Ön işleme aşaması birkaç veri hazırlama işlemi içerir. Kümeleme aşaması genlerin ifade değerlerine göre gen kümelerini oluşturur. Ağ çıkarım aşaması elde edilen kümelerin hangi nodlarla ilişkili olduğunu Bayes Ağlara dayanarak çıkarır. Motif çıkarım aşaması, çıkartılan ilişkiyi doğrulamak için iki üye arasında potansiyel bağlanmaları değerlendirmek için kullanılır. Analiz aşaması kümelerin biyolojik doğruluk olasılığını belirlemek için herbir kümeye Gen Ontoloji (GO) zenginleştirmesi uygular ve yerel düzenleme ilişkilerini açıklayacak ağ motiflerini listeler.



Şekil 3.1 Geliştirilen aracın çalışma aşamaları (EM: Expectation Maximization, TF: Transcription Factor)

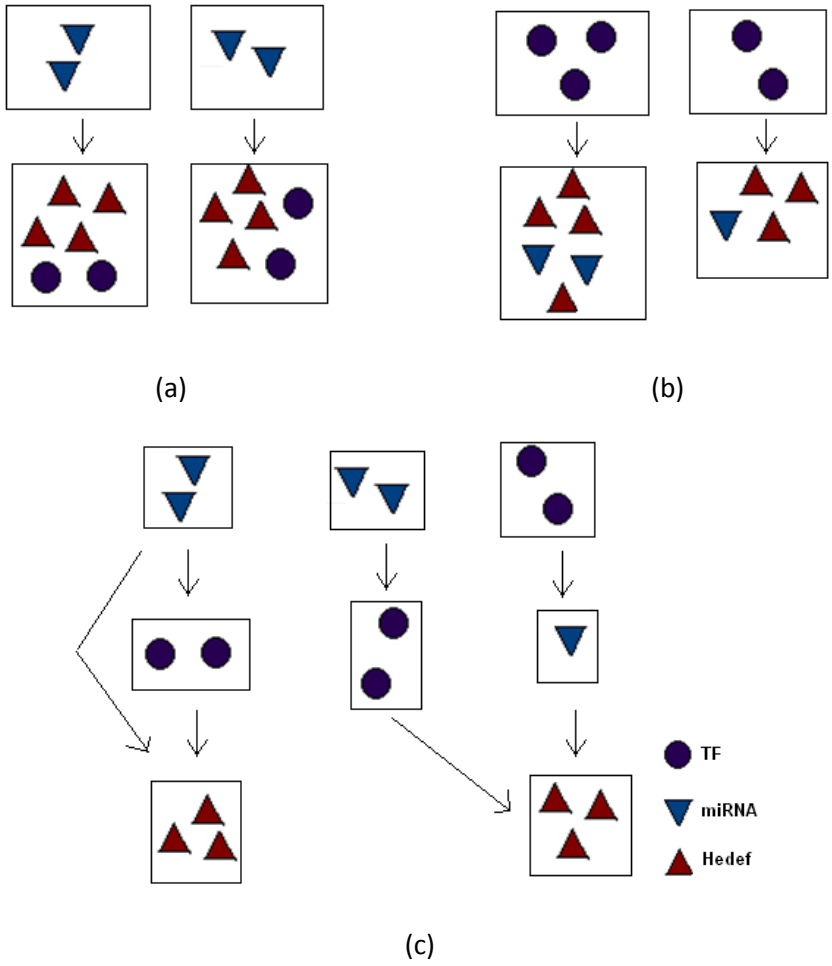
3.1. Ön İşleme Aşaması

Bu aşamada bir satırdaki verilerin ortalaması sıfır ve standart sapması bir olacak şekilde ölçeklendirilmiştir. İfade değerleri değişken olan profilleri tutmak için standart sapması belli bir değerin altında olan genler kaldırılmıştır. Bu değer veri setine bakılarak seçilebilir ama bu çalışmada Joshi ve diğ. (2008) tarafından önerilen şekilde 0.5 kullanılmıştır [4]. Veri matrisinde kayıp olan değerler sıfır ile değiştirilmiştir. Satırın ortalaması sıfır yapıldığından kayıp değerlerin sonuçları en az etkilemesi sağlanmıştır.

3.2. Kümeleme Aşaması

Kümeleme aşaması k-ortalamalar algoritmasına göre kümelemeyle başlar. Bayes modelle en uygun kümeler elde edilerek devam eder [4]. Optimal kümeler beklenti-makzimizasyon algoritmasıyla iyileştirilir. Bunu sıkı kümeleri elde etmek için bulanık eleme işlemi izler. Son olarak belli bir sayıdan düşük genleri içeren kümeler elenir. Bu çalışmada bu sayı dört olarak kullanılmıştır.

Modül ağının yapılandırılması ilk modüllerin oluşturulmasıyla başlar ve tekrarlayan iki açgözlü (greedy) adımla gerçekleştirilir. Başlangıç modülleri k-ortalamalar kümeleme algoritması ile seçilmiştir. Açgözlü adımlarda yukarıda bahsedilen Bayes skorunun iyileştirilmesi ve belirli sayıda tekrarlama sonucu bu skorun eniyilenmesi hedeflenmektedir. Birinci adım modüllerin oluşturulduğu ve gen veya miRNAların modüllere tekrar atandığı kısımdır. Bu kısımda Gibbs Örneklemesi yaklaşımı kullanılmıştır. İkinci adım yapısal atamaların, yani modüller arası kenarların belirlendiği bölümdür. Herhangi bir kenar değişikliği Bayes skorunu iyileştiriyorsa bu seçimle devam edilir. Düzenleyici ilişkiye aykırı olan (örneğin bir hedef genin başka bir TFnin düzenleyicisi haline gelmesine neden olan) kenarlar reddedilir.



Şekil 3.2 Bütünleştirici modül ağı oluşturulması. (a) miRNA-düzenleyici modül ağı, (b) TF-düzenleyici modül ağı (c) İki modül ağındaki hedef kümelerin kesişimi, TF->miRNA ve miRNA-TF ikililerinin birleşimiyle oluşturulan yeni modül ağı.

3.2.1. K-ortalamalar algoritması

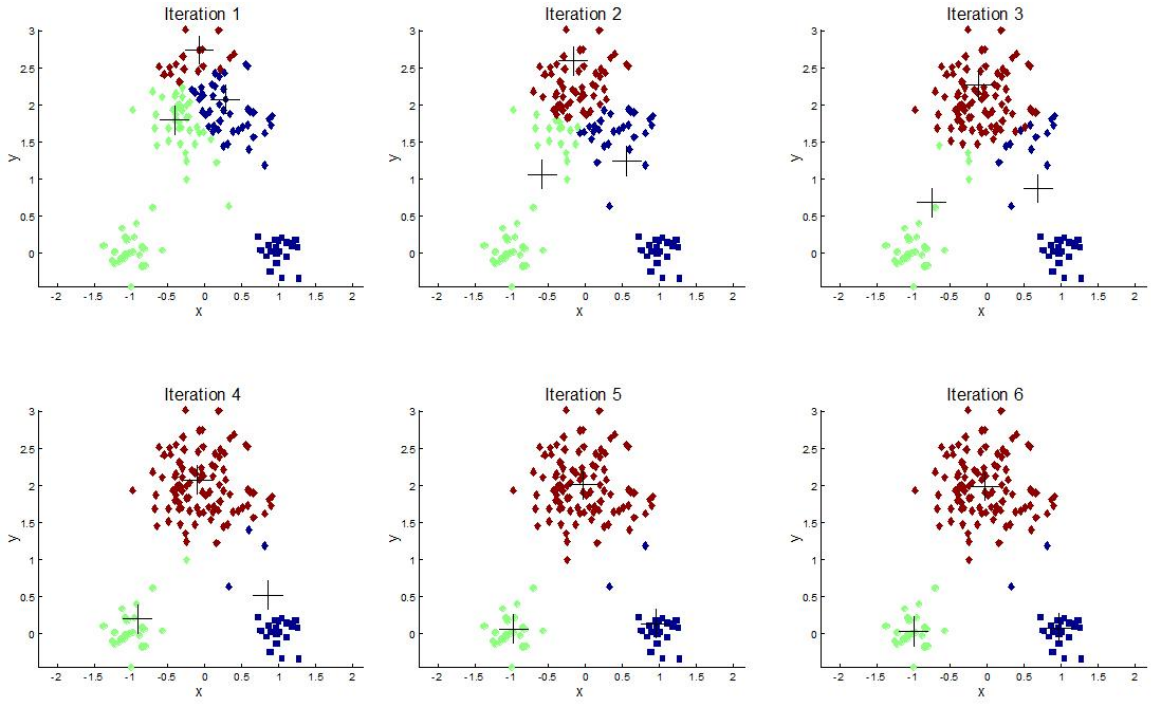
K-ortalamalar algoritmasının amacı *n* tane gözlemi *k* tane kümeye atamaktır. Verileri, her yinelemede kümeler için ortalama değerler hesaplayarak kümelemeye çalışır. K-ortalamalar algoritması eldeki verileri *k* adet kümeye ve kümelerin ortalamalarına göre ayırır. Kısacası *n* adet nesneyi küme içi benzerlik maksimum, kümeler arası benzerlik minimum olacak şekilde *k* tane kümeye böler.

K-ortalamalar algoritmasına göre öncelikle kümelerin başlangıç merkez noktalarını ve ortalamalarını temsil etmek üzere *k* adet nesne seçilir. Diğer nesneler kümelerin ortalama değerlerine olan Öklid uzaklıklarının [74] minimum değerlerine göre kümelere ayrılır. Bir *x* nesnesinin tüm kümelerin merkezlerine olan Öklid uzaklığı hesaplanır ve *x* nesnesi uzaklığın minimum olduğu kümeye alınır. Sonrasında tüm kümelerin ortalama değerleri tekrar hesaplanır ve yeni küme merkezleri bulunur. Herhangi bir değişim olmayıncaya kadar bu iterasyon tekrarlanır.

n Öklid uzayındaki boyut sayısı olmak üzere $P = (p_1, p_2, \ldots, p_n)$ ve $Q = (q_1, q_2, \ldots, q_n)$ noktaları arasındaki Öklid uzaklığı şu şekilde hesaplanır:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}. \qquad (3.1)$$

Başlangıç küme merkezlerinin seçimi için çeşitli teknikler olmasına rağmen bu çalışmada veri setinin merkezine en yakın noktalar başlangıç noktaları olarak seçilmiştir.



Şekil 3.3 K-means algoritmasının iterasyonları [25]

Şekil 3.2'de k-ortalamalar algoritmasının iterasyonları görülmektedir. İterasyon 1'de başlangıç noktaları atandığı için kümeler tam olarak birbirlerinden ayrılamamıştır. Her bir aşamada kümelerin merkez noktaları yani ortalama değerleri '+' (artı) işaretiyle gösterilmiştir. İterasyon 1'de görüldüğü gibi kümelerin merkez noktaları birbirine çok yakındır. Bu istenmeyen bir durumdur çünkü kümeler tam olarak birbirinden ayrılamamıştır. Her bir iterasyonda veri noktaları yeni kümelere atanmış ve merkez noktaları gitgide birbirinden uzaklaşmıştır. Son iterasyondaysa küme merkezleri birbirinden uzak durumda olduklarından amaca ulaşılmıştır. Örnekteki k-ortalamalar yöntemi sonucunda üç farklı küme elde edilmiştir.

K-ortalamalar algoritmasına göre kümeleme yapılırken öncelikle karışık halde verilmiş olan veri seti sıralanır. Veri setinin ortalaması alınır ve başlangıç noktaları belirlenir. Her noktanın belirlenmiş olan kümelerin merkez noktasına göre Öklid uzaklığı alınır. Veriler en yakın olduğu merkez noktasının kümesine dahil olur. Her küme için küme elemanlarının ortalaması alınır. Bu ortalamalar kümelerin yeni merkez noktasıdır. Bir önceki adımda hesaplanan merkez noktası, sonraki adımda hesaplanan merkez noktasıyla aynı çıkana kadar bu işlemler tekrarlanır.

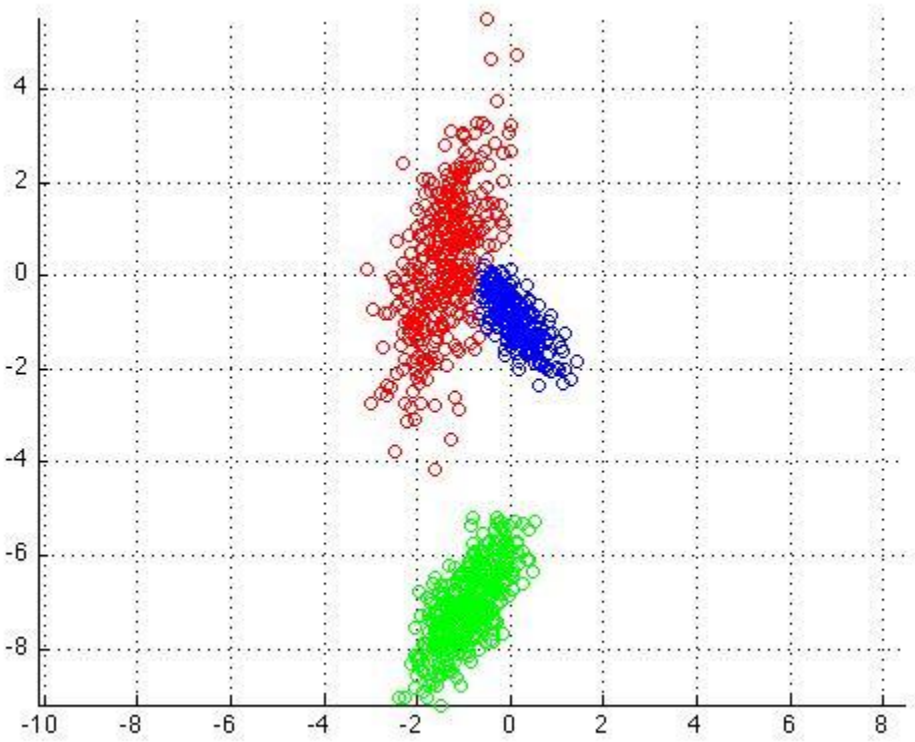
3.2.2. Beklenti eniyileme (Expectation maximization (EM)) algoritması

İstatistikte kullanılan beklenti eniyileme algoritması, istatistiksel modellerde, model gözlemlenmemiş, çıkartılmış değişkenlere dayanıyorsa kullanılır. Bu çalışmada bu değişkenler kümelerin merkez noktaları olmaktadırlar. EM algoritması [75] istatistiksel modellerde kullanılan parametrelerin maksimum komşuluğunu hesaplamak için kullanılan yinelemeli bir yöntemdir. Beklenti (E) ve eniyileme (M) adımı olmak üzere EM algoritması iki aşamadan oluşur. Beklenti (E) adımında parametrelerin o anki tahminleri kullanılarak hesaplanan log-komşuluğunun olasılığı için bir fonksiyon oluşturulur. Eniyileme (M) adımındaysa E adımında bulunan log-komşuluğu fonksiyonunu maksimize eden parametreler hesaplanır.

Örnek üzerinden anlatılmak istenirse normal bir yazı tura atma deneyi düşünülebilir. Bu deneyde iki adet bozuk para olduğu varsayılmaktadır. Beşer kez şu adımlar tekrarlanır: rastgele iki bozuk paradan birisi seçilir ve onar kez havaya atılır. İlk aşamada bozuk paraların hangisinin seçildiği ve hangi yüzlerinin geldiği bilinmektedir. Amaç bozuk paraların yazı gelme olasılıklarının bulunması ya da yakınsanmasıdır. Bu sorun belli matematiksel işlemler yardımıyla çözülebilir. İkinci aşamada başlangıçta bu bozuk paraların hangisinin seçildiği bilgisini bilmediğimiz varsayılmaktadır. Önceki paragrafta anlatıldığı gibi gözlemlenmemiş bir değişken vardır. Her bir bozuk para için olasılıkları normal yollarla hesaplamak artık mümkün değildir çünkü her bir atış için kullanılan bozuk paranın hangisi olduğu bilinmemektedir.

Sorunun çözümü için bir yinelemeli plan şu şekilde uygulanabilir: Her bir bozuk paraya olasılık başlangıç değerleri verilir ve gözlenen yüzleri o anki parametreler kullanılarak hangi bozuk paranın üretebileceği olasılıkları bulunur, sonrasında bu olasılıklar doğru kabul edilerek normal komşuluk hesaplama prosedürleri uygulanır. Bu adımlar bir sonraki aşamada aynı sonuçla karşılaşıncaya kadar uygulanır.

EM algoritması her bir aşamada hangi bozuk paranın atıldığı bilgisi yerine eksik verinin belirlenmesi için o anki parametreleri kullanarak belli bir olasılık hesaplar. Bu olasılıklar tüm olası verinin tamamlanması durumları için eğitim veri kümesi oluşturmakta kullanılır. Sonrasında bu olasılıklar bir sonraki adımda daha iyi eğitim veri kümesi oluşturmakta kullanılır ve en sonunda yakınsaklaştırılmış en iyi sonuç bulunur. Matematiksel açıklamalar sonraki sayfada verilmiştir.



Şekil 3.4 EM algoritmasının çıkarttığı kümeler

$\mathbf{X}$ gözlenen veriler, $\mathbf{Z}$ gözlenmeyen kayıp değişkenler ve $\boldsymbol{\theta}$ bilinmeyen parametrelerin vektörü olsun. Bu istatistiksel modelin komşuluk fonksiyonuysa $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ şeklinde verilsin.

E adımında $\mathbf{Z}$ nin koşullu dağılımıyla, $\boldsymbol{\theta}^{(t)}$ nin o anki tahminleriyle ve $\mathbf{X}$ verileriyle log-komşuluk fonksiyonunun beklenen değeri hesaplanır:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}\left[\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})\right] \tag{3.2}$$

M adımında bu büyüklüğü maksimize eden değer bulunur:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \tag{3.3}$$

3.2.3. Modüllerin Çıkarımı

Kümeleme Joshi ve diğ. (2008) tarafından tanımlanmış bir model esas alınmıştır [4]. Aynı modül içerisindeki genlerin aynı davranışı göstereceği ve ayrı koşul veya zamanlarda ölçülen gen ifadelerinin bağımsız olması varsayımıyla, herhangi bir modül/yapı atamasına karşılık gen ifadesi bir Gaussian dağılımla [76] tanımlanır.

N, genlerin sayısı, *M*, deney sayısı, $X=\{x_{ij},\ i=1..N,\ j=1..M\}$, herhangi bir gen veya miRNA için ifade profili, $E=\{E(i),\ i=1..N\}$, küme üyeliği belirteci ve *K*, küme sayısı olmak üzere, aynı kümede bulunan genlerin ortak bir normal dağılıma sahip olduğu varsayılır:

$$x_{ij} \sim N(\beta_{kj}, \sigma_{kj}^2) \tag{3.4}$$

Burada *i*, gen numarası, *k* küme numarası ve *j* ifade numarasını belirtir. Bu dağılım üzerinden olabilirlik (likelihood) fonksiyonu şöyle tanımlanır:

$$P(X \mid E, \beta, \sigma^2) \propto \prod_{k=1}^{K} \prod_{E(i)=k} \prod_{j=1}^{M} \left((\sigma_{kj}{}^2)^{-\frac{1}{2}} e^{\frac{-1}{2\sigma_{kj}{}^2}(x_{ij}-\beta_{kj})^2} \right) \tag{3.5}$$

Marjinal olabilirliği hesaplamak için *β* ve *σ* parametreleri üzerinden integral alınırsa şu sonuç elde edilir:

$$P(X \mid E) = \prod_{k=1}^{|E|} \prod_{j=1}^{M} \iint \prod_{E(i)=k} P(x_{ij} \mid \beta_{kj}\sigma_{kj}{}^2)\, p(\beta_{kj} \mid \beta_0 \sigma_{kj}{}^2)\, p(\sigma_{kj}{}^2)\, d\beta_{kj}\, d\sigma_{kj}{}^2 \tag{3.6}$$

β ve *σ* parametreleri için prior dağılımlar ise şu şekilde varsayılmıştır:

$$P(\beta_{kj} \mid \sigma_{kj}{}^2) \approx N(\beta_0, \sigma^2) \tag{3.7}$$

$$P(\sigma_{kj}{}^2) \approx InvGamma(a, b) \tag{3.8}$$

Burada β_0, a ve b parametrelerinin bilindiği varsayılmaktadır. β_0, değeri ifade seviyelerinin ortalama değeri, a değeri 1 ve b değeri standard sapmanın iki katı olarak alınmıştır.

E için prior dağılım Bayes modellerinde sıkça kullanılan Dirichlet dağılımı ile şu şekilde tanımlanmıştır ($n_{-i,k}$, k.kümenin eleman sayısı ve α ise apriori parametredir):

$$P(E(i)=j \mid E(-i)) = \begin{cases} \dfrac{n_{-i,k}}{N-1+\alpha} & j=0 \\ \dfrac{\alpha}{N-1+\alpha} & j>0 \end{cases} \tag{3.9}$$

Gibbs örneklemesi yinelemeli olarak genlerin kümelere atanması işlemini günceller. Herhangi bir yöntemle ilk kümeler oluşturulur. Her gen bulunduğu kümeden çıkarılır, sadece kendisinden oluşacak küme dahil tüm kümelere ait olma olasılığı mevcut küme dağılımına koşullu olarak hesaplanır, en yüksek olasılığın elde edildiği kümeye atama yapılarak *E(i)* değeri güncellenir. Bu adım tüm genler için sırayla tekrarlanır. Bir yakınsama durumuna veya maksimum yineleme sayısına kadar adımlara devam edilir. Bu yinelemeler boyunca marjinal olabilirlik izlenir ve en yüksek olabililirliğe karşılık gelen sonuç raporlanır. Yerel eniyilerden kaçınmak için algoritma birkaç kez çalıştırılarak en iyi çözüm alınır.

3.2.4. Bulanık kümeleme (Fuzzy Clustering)

EM algoritmasından alınan sonuçlar kullanılarak bir bulanık matris oluşturulur. Bu matris sadece alınan sonuçlar için hangi genin hangi kümede olduğunu göstermektedir. Bulanık matristen genlerin kümelerde olma olasılıklarını tutan ikili (pairwise) matris oluşturulur. İkili matristen bulanık kümelerin belirlenmesi için [73] makalesindeki yöntem izlenmiştir. Bu yöntem olasılık matrisinin eigen değerini ve karşılık gelen eigen vektörünü yinelemeli olarak hesaplar, bulanık kümeleri bu eigen vektörlere göre oluşturur ve son kümeye atanan genlerin ağırlıklarını (olasılıklarını) matristen çıkararak olasılık matrisini günceller. Yalnızca yüksek olasılık değerlerine sahip bir bulanık kümeyi tutarak, standart kümelemeye oranla

daha yüksek işlevsel tutarlılık gösteren sıkı kümeler elde edilmektedir. Aynı zamanda birçok bulanık kümeye ait küçük ama önemli olasılık değerlerine sahip genlerin tutulması, kısmi beraber ifade edilen genler arasındaki veya çok fonksiyonlu genleri tanımlamayı mümkün kılmıştır.

3.3. Ağ Çıkarım Aşaması

Ağ çıkarım aşaması iki bağımsız modül ağı çıkarmayı ve hedef kümelerin kesişimine göre bunları birleştirmeyi hedefler. Birinci ağda mRNA ve miRNA'lar potansiyel hedefler olarak alınırken girdi listesindeki TF'ler düzenleyici olarak alınmıştır. İkinci ağda tüm genler üzerinde miRNA düzenlemelerini bulmak için miRNA'lar ile TF'lerin yerleri değiştirilmiştir. Son modül ağı, genlerin ve miRNA'ların transkripsiyon sırasında ve transkripsiyon sonrasında düzenlenmesi bilgisini veren tek bir küme içerir. TF ve miRNA arasındaki ikili ilişkiler de tanımlanmıştır.

3.4. Motif Çıkarım Aşaması

İkili ilişkilerin çıkarımı düzenleme motiflerini sağlar. Bu da ortak düzenlemelerin yanı sıra tekil eleman terimlerini kullanarak son ağı görmekte yardımcıdır. Güvenilir bir analiz için bu aşamada dizilim bilgisi kullanılmıştır. TF-hedef ikililerinde TF bağlanma alanlarının ve hedef olma potansiyeli taşıyan genlerin dizilim bilgisi varsayılan bağı değerlendirmek için kullanılmıştır. miRNA-hedef ikililerinde olgun miRNA dizilimleri ve hedeflerin 3'UTR dizilimleri alınmıştır. CircuitsDB'de belirtildiği şekilde eşleştirme yapılmıştır [5].

3.5. Analiz Aşaması

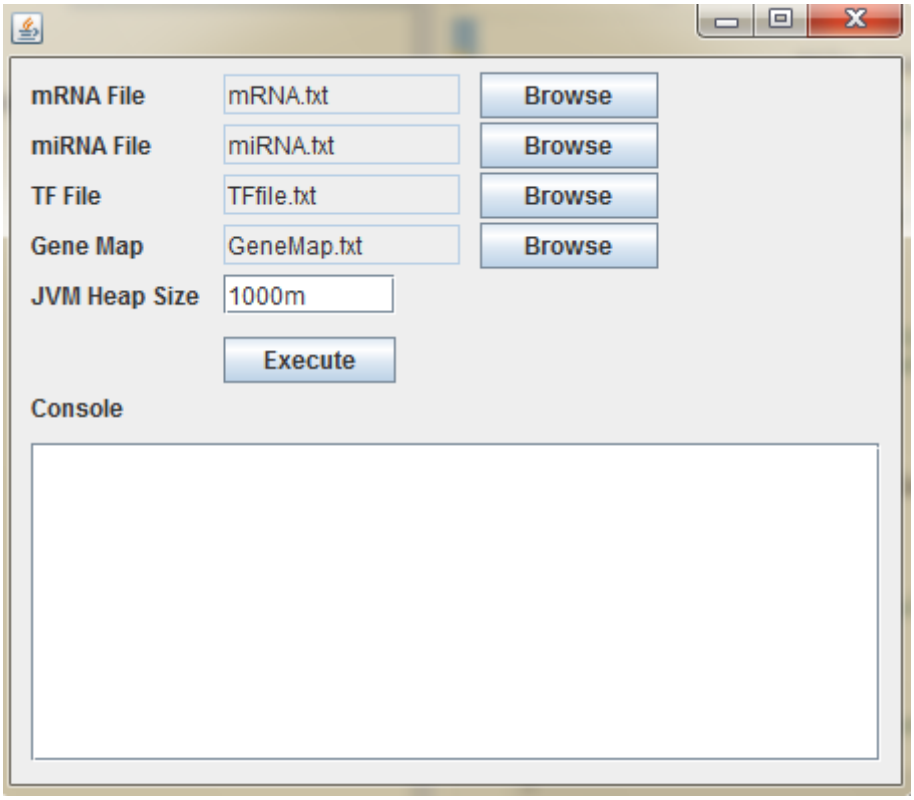
Kümeleri doğrulamak için Go terimleri zenginleştirme analizi yapılmıştır. Çalışma aşamaları eşleşen gen listesiyle GO terimlerinin, modül ısı haritaları ve düzenleyici motiflerin raporlanmasıyla bitmektedir.

4. GELİŞTİRİLEN ARAÇ

Geliştirilen araç biyologların kullanımı için gerekli sayısal ve tekniksel ayrıntılardan ayrıştırılmıştır. Kullanıcılar veri kümelerini araca yükleyebilir, ısı haritasını görüntüleyebilir ve analiz sonuçlarını elde edebilir. Aracın amacı birden fazla programlama dilinde yazılmış farklı programları çalıştırmaktır. Araç SegExpReg adıyla anılacaktır.

4.1. İşlevler ve Kullanıcı Arayüzü

SeqExpReg aracı çalıştırıldığında Şekil-4.1'de görünen arayüz kullanıcıyı karşılamaktadır. Burada biyoloğun tek yapması gereken veri setini yüklemesi ve çalıştır tuşuna basmasıdır. Bu tür programların çalışması büyük zaman aldığından dolayı programın altına kullanıcının programın gidişatını izlemesi için konsol ekranı yerleştirilmiştir. Büyük veri setlerinde programın çalışması günlerce sürebileceğinden ötürü herhangi bir aksaklıktan dolayı sürecin başına dönmeyi engellemek için program sürekli olarak yedekleme yapmaktadır.



Şekil 4.1 Geliştirilen Araç Kullanıcı Arayüzü

Kullanıcı veri kümesini yükledikten sonra, başka hiçbir işlem yapmasına gerek kalmadan program analiz sonuçlarını (Kümelerin ısı haritaları, Go analiz dosyaları ve Ağ motifleri) belli bir klasöre kaydetmektedir. Yöntemler kısmında belirtilen tüm işlemler kullanıcıdan bağımsız bir şekilde program tarafından yapılmaktadır.

4.2. Teknik Altyapı

Yazılımın geliştirilmesinde Java, Matlab ve Perl dilleri kullanılmıştır. Algoritmanın uygulanması, arayüzlerin geliştirilmesinde ve GO analizinde Java, hesaplama kısmının gerçekleştirilmesinde Matlab, dizilim verilerinin gen ifadesi verileriyle birleştirilmesi ve düzenleyicilerin belirlenmesi işlemi ise Perl diliyle yapılmıştır. Java'dan başka bir Java programı, Matlab ve Perl betikleri çalıştırılmıştır.

Kod geliştirme ortamı olarak Eclipse seçilmiştir. Yöntemin geliştirildiği ve farklı parametrelerle test edilmesi, 2.10 GHz işlemci hızına sahip Intel(R) Core(TM)2 işlemcili, 4.00 GB RAM'e sahip bilgisayarda yapılmıştır.

Geliştirilen yazılım aracı işletim sisteminden bağımsız Java 1.7, Perl 5.14 veya üstü sürümlerinin yüklü olduğu her bilgisayarda çalışabilmektedir.

5. VERİ KÜMELERİ

Geliştirilen yöntemin performansı farklı kanser türlerine ve meme kanserine sahip veri kümelerinde denenmiştir. Bu veri setleri NCBI GEO [72] veri merkezlerinden alınmıştır. Bu veri setlerindeki kayıp değerler sıfıra eşitlenmiştir. Çalışmada kullanılan veri setleri hem miRNA hem de mRNA için yapılmış olmalıdır. mRNA ifade profilleri çıkartılırken kullanılan örnekler miRNA ifade profilleri çıkartılırken de kullanılmış olması gerekmektedir. Ayrıca bu veri kümelerinin aynı sütun (örnek) sayısına sahip olması gerekmektedir. Veritabanlarında bu koşulları içeren çok az sayıda veri kümesi bulunmaktadır. Farklı kanser türleri ve meme kanseri örneklerini içeren iki ana veri topluluğu kullanılmıştır. Bu veri topluluklarının her biri için mRNA ve miRNA veri setleri olacağından toplam olarak dört veri seti kullanılmıştır.

Birinci ve ikinci veri setlerinde farklı kanser tipleri içeren, eşlenmiş mRNA ve miRNA profillerinin normal ve hastalıklı örnekleri kullanılmıştır. Kanser tipleri arasında kolon, pankreas, böbrek, idrar kesesi, prostat, yumurtalık, uterus, akciğer, mezo, mela ve meme kanseri sayılabilir. mRNA ifade profilleri [26] makalesinden alınmıştır. Veri kümesi her doku için 11 sınıf tümörden ve bazı normal örneklerden alınan 16,063 genin ifade profillerini içeren 89 örnekten oluşmaktadır. [27] makalesinde [26] makalesinin örnekleri kullanılarak 217 memeli miRNA'sının çubuk tabanlı sitometrik miRNA ifade profilleme yöntemiyle sistematik ifade analizi yapılmıştır. Çalışmada bu 217 miRNA'nın, mRNA ifade profilleri veri kümesinde olduğu gibi 89 örneğinin bir alt kümesi kullanılmıştır. miRNA verisi GSE2564 erişim numarasına sahiptir.

Üçüncü ve dördüncü veri setlerinde meme kanseri tiplerini içeren örnekler kullanılmıştır. mRNA veri seti GSE19783, miRNA veri seti GSE19536 erişim numaralarına sahiptir. Bu veri setlerinde 101 meme kanseri örneğinden miRNA profilleme uygulanmıştır. Deneyler, farklı dizi ve zaman noktalarında yinelenen hibridizasyonlar (99 Örnek) kullanılarak gerçekleştirilmiştir. İki örneklem yalnız bir kere profillenmiştir. Probların kopyalarının miRNA sinyal yoğunlukları için platform üstünde ortalamaları alınmıştır. Log2 dönüştürülmüş ve yüzde 75

normalleştirilmiştir. 114 meme kanseri örneği için mRNA profilleme uygulanmıştır. Bu çalışmada miRNA profillemede kullanılan örnek sayısı, mRNA profillemede kullanılan örnek sayısından daha az olduğu için mRNA ve miRNA için aynı 101 örnek kulanılmıştır. Kullanılan mRNA veri setinde 101 tane örnek, 40996 mRNA, miRNA veri setinde ise 101 örnek, 902 tane miRNA vardır. Farklı mikroçip uçları aynı mRNA'ya karşılık gelebildiğinden bazı genler tekrar edebilmektedir. miRNA veri setinde hiç ifade verisi olmayan miRNA'lara rastlanmıştır. Bu miRNA'lar silinmiş ve son olarak miRNA veri setinde gen sayısı 460'a düşürülmüştür.

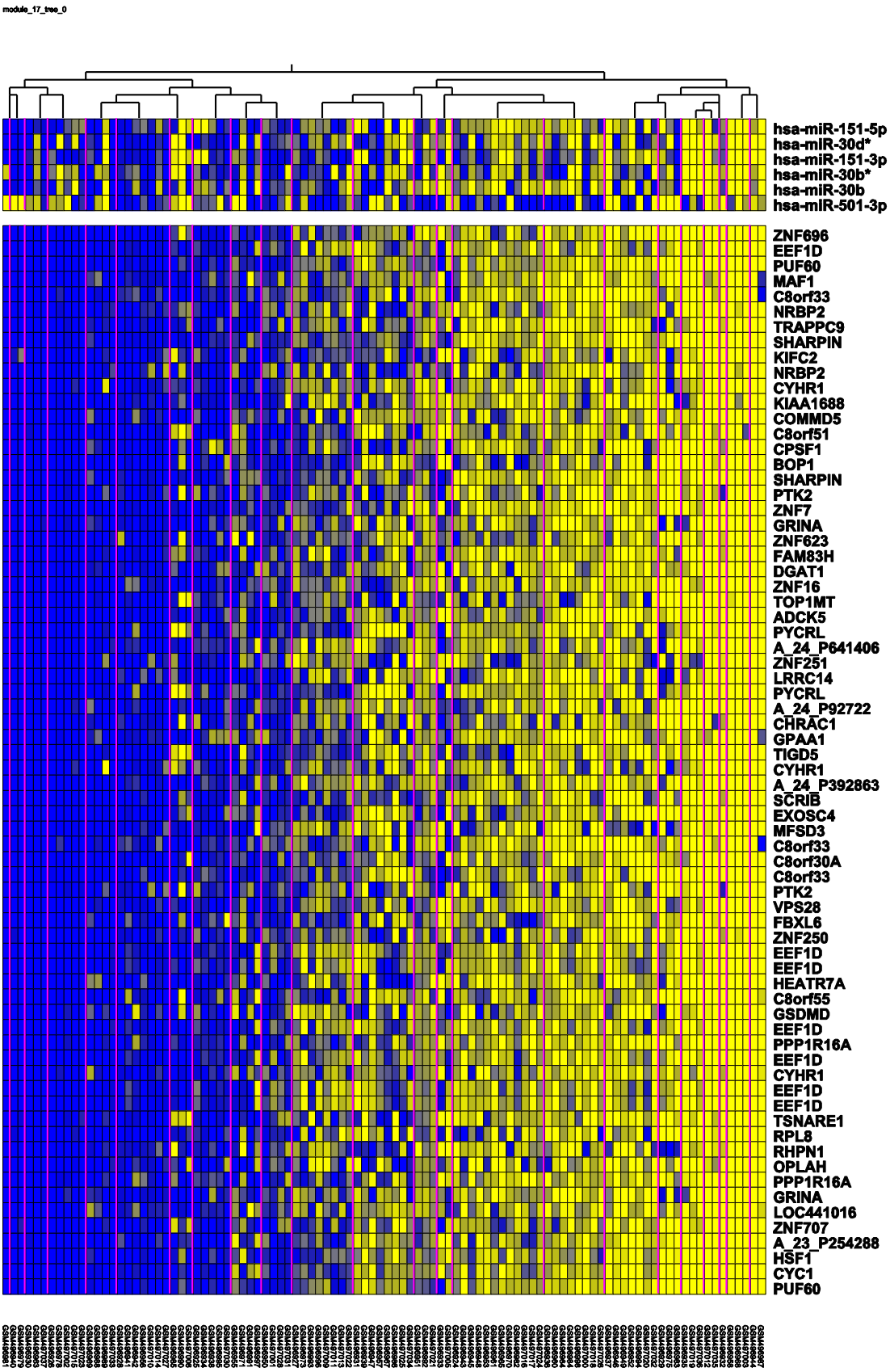
6. SONUÇLAR

6.1. Çıkarılan Modüller

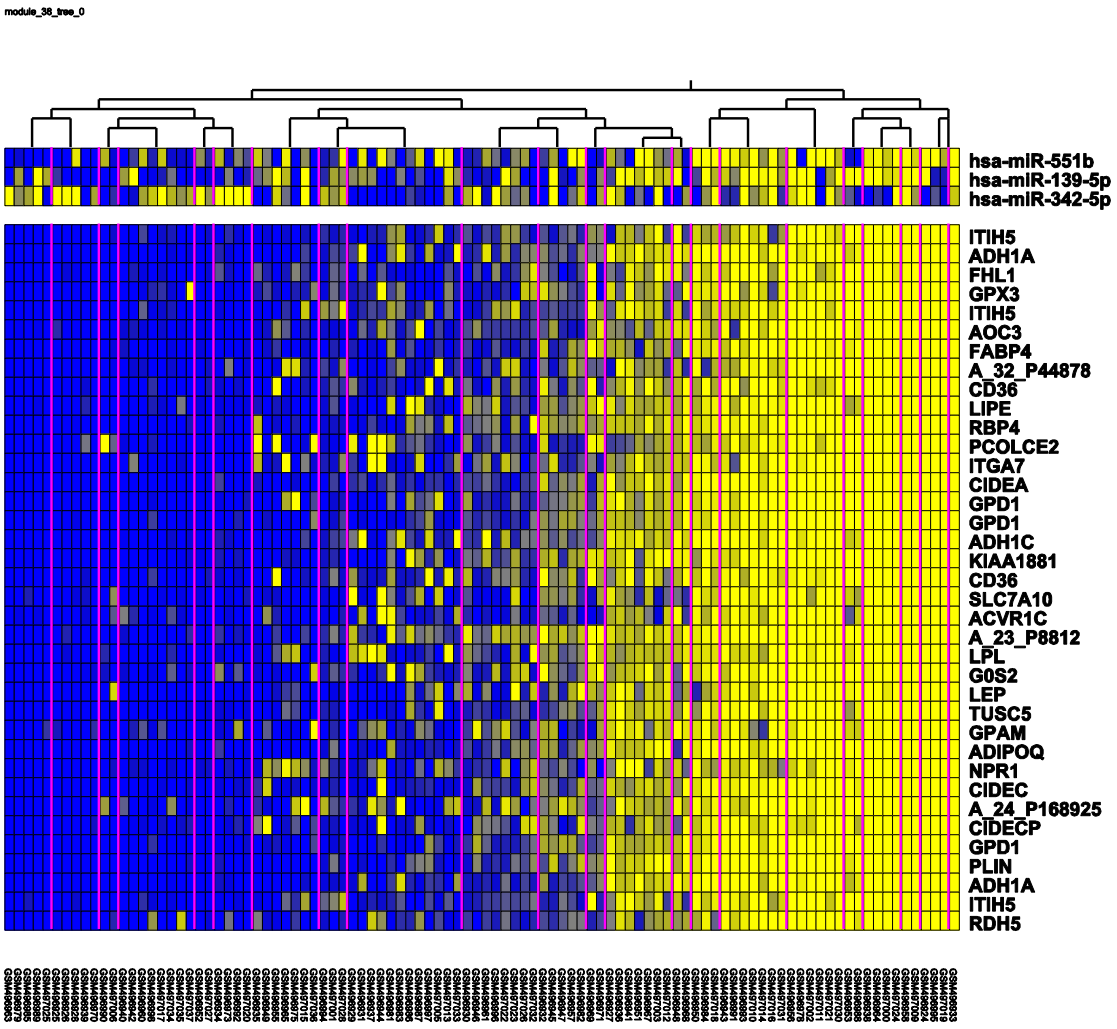
6.1.1. Düzenleyici TF alındığında

Çoklu kanser veri kümesinden 86 modül çıkarılmıştır. Bu modüllerin en az 26'sında en az bir GO terimi zenginleştirilmiştir. Modüllerden bazıları belirli kanser tiplerine ilşkin ilginç mRNA ve miRNA grupları ortaya çıkarmıştır. Modüller ısı haritalarına ve GO terimi zenginleştirme analizine göre değerlendirilmiştir.

Meme kanseri veri kümesindense 361 modül elde edilmiştir. Bu modüllerden 358 tanesinde en az bir düzenleyici miRNA modül ataması yapılmıştır. Bu durumda hedef modüller miRNAları da içerdiğinden GO analizi yapılmamıştır.

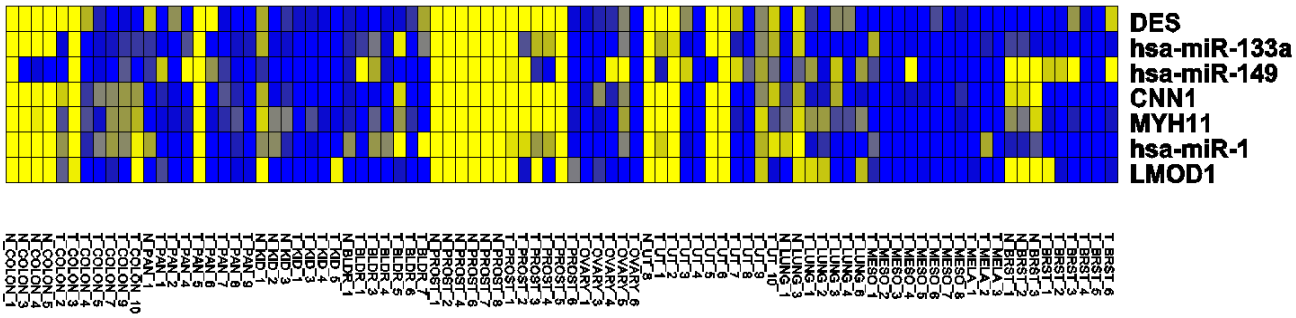


Şekil 6.1 Meme kanseri modül ağında 17 nolu modül için ısı haritası



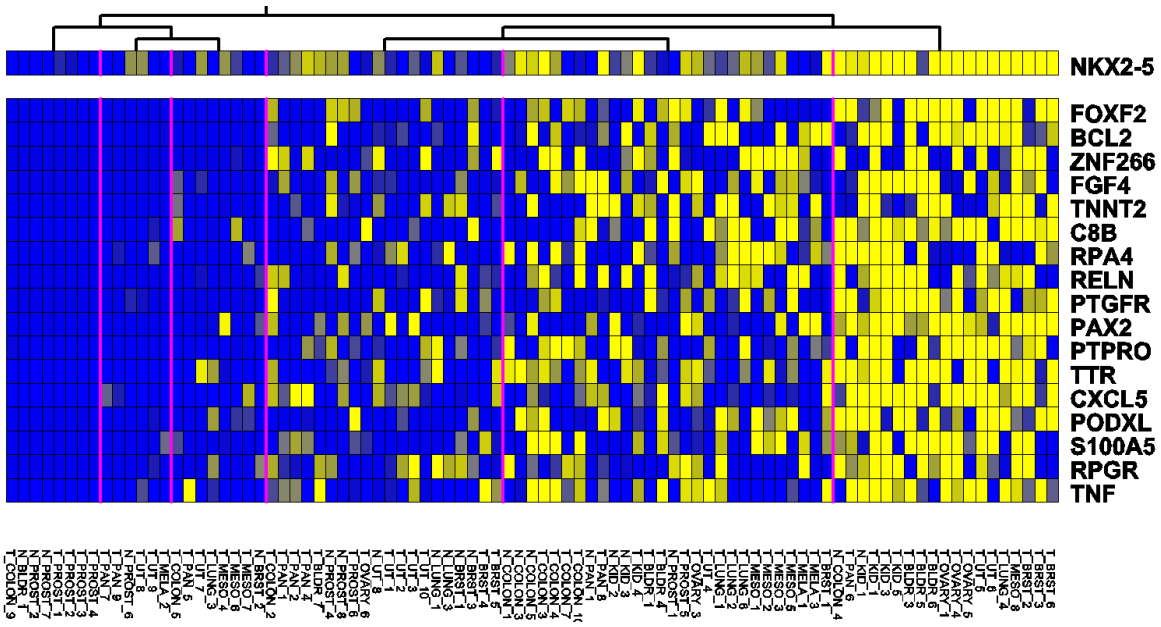
Şekil 6.2 Meme kanseri modül ağında 38 nolu modül için ısı haritası

Şekil 6.1 ve 6.2'de meme kanseri ısı haritalarının örnekleri görülmektedir. Dikkat edilmesi gereken nokta birden fazla düzenleyici ataması olduğudur.



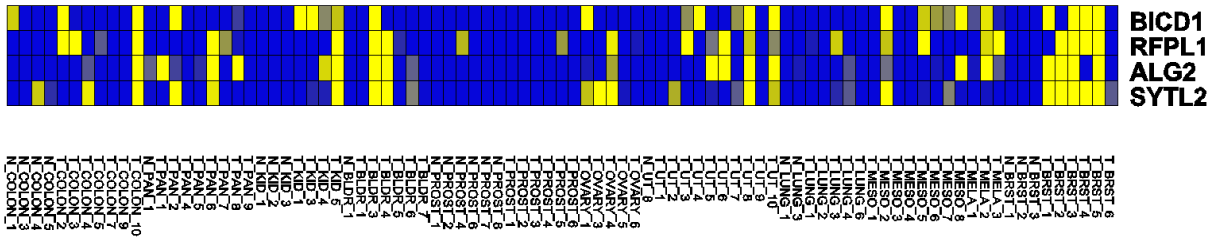
Şekil 6.3 Çoklu kanser modül ağında 61 nolu modül için ısı haritası

Şekil 6.3'de gösterilen modülde aynı ifade profillerini paylaşan gen grupları ve miRNA'lar görülmektedir. GO terimleri istatistiklerine göre bu genler organel organizasyonu, biyogenesis, hücresel bileşen organizasyonu ve hücre iskeleti organizasyonu ile ilişkilendirilir. Bu gruptaki genler ve miRNA'larda meme ve kolon kanseri örneklerinde ilginç bir şekilde azalma yönlü düzenleme gözlenmiştir. Normal ve hasta prostat örnekleri arasında farklı gen ifadelerine rastlanmamıştır. Gruptaki miRNA'lar HMDD veritabanında meme neoplasmasıyla ilişkilendirilmiştir.



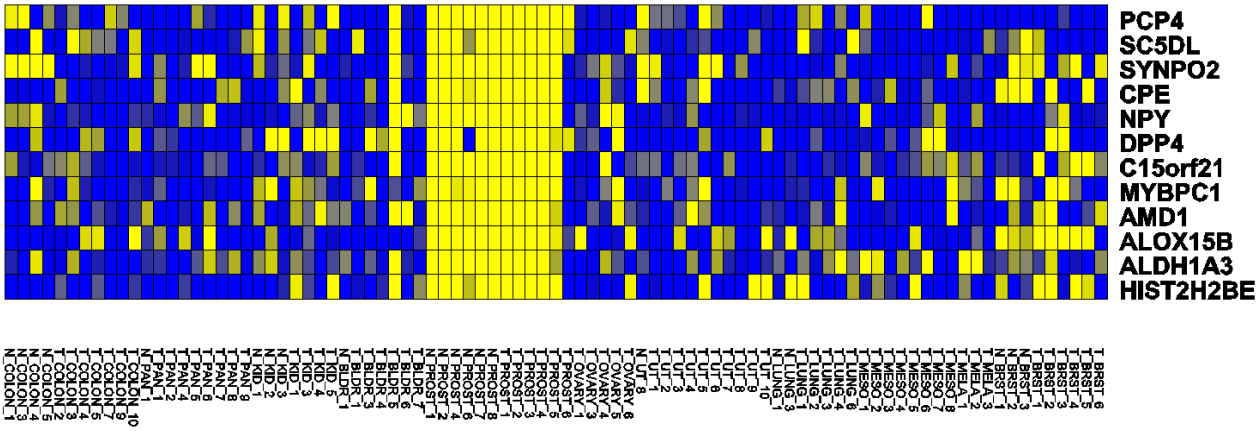
Şekil 6.4 Çoklu kanser modül ağında 44 nolu modül için ısı haritası

Şekil 6.4'de görülen başka bir modülde transkripsiyon faktörü NKX2-5 tarafından belli bir gen kümesi gen ifadesini yükseltici yönde etki göstermiştir. Normal dokularda bu gen ifadesini yükseltici etki görülmemektedir. Birkaç GO terimi dışında en ilgi çekicileri anatomik yapı morfojinezleri ve humoral bağışıklık tepkisi biyolojik işlemleri olmaktadır. NKX2-5'in dokuya özel gen ifadesinin düzenlenmesi ve doku farklılaşmasında görev yaptığı bilinmektedir. Şekil 6.4'deki modülde tümör dokularındaki mRNA'ların gen ifadesini yükseltici rol oynadığı görülmektedir.



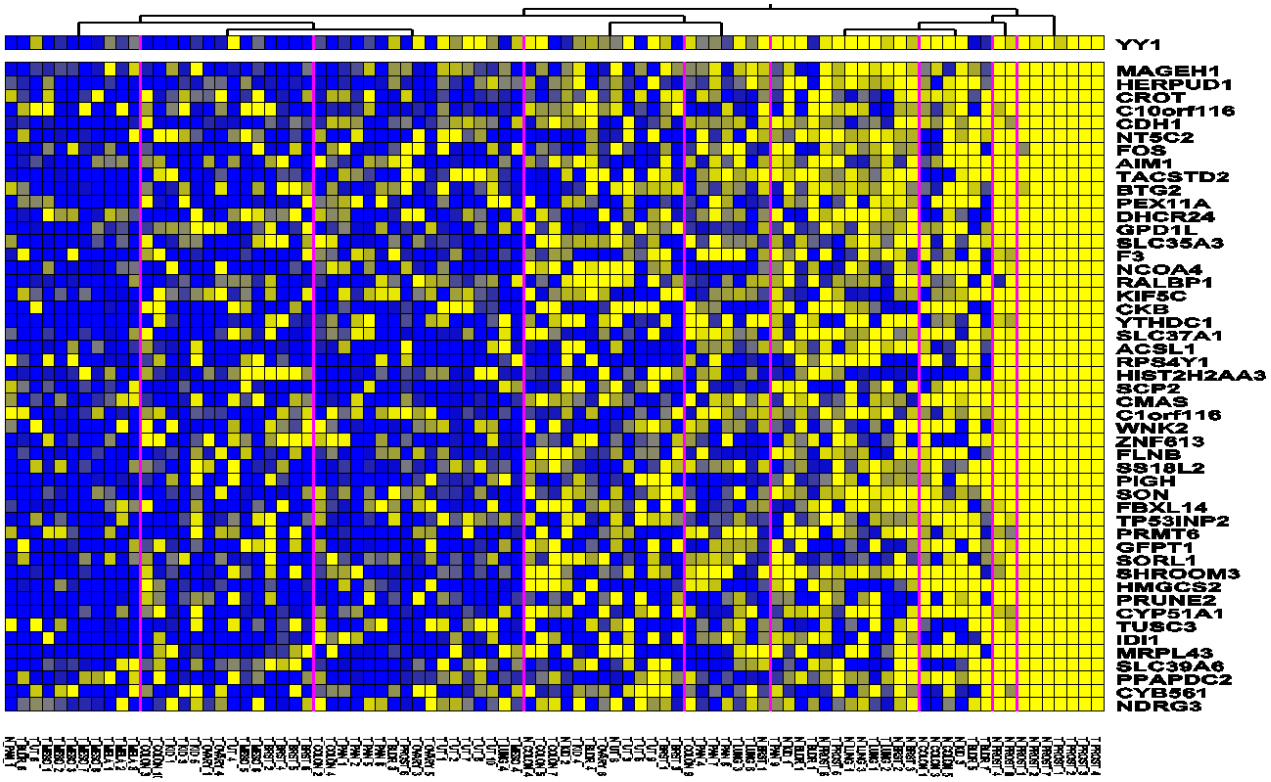
Şekil 6.5 Çoklu kanser modül ağında 54 nolu modül için ısı haritası

Şekil 6.5'den çıkan tek GO terimi makromolekül yerelleştirilmesidir. Bu kümedeki genler, muhtemelen karşılık gelen hastalar tarafından alınan makromoleküler ilaçların etkisiyle mutasyon geçirmiş olabilir. Başka bir dokuya özgü modül şekil 6.6'da gösterilmiştir.



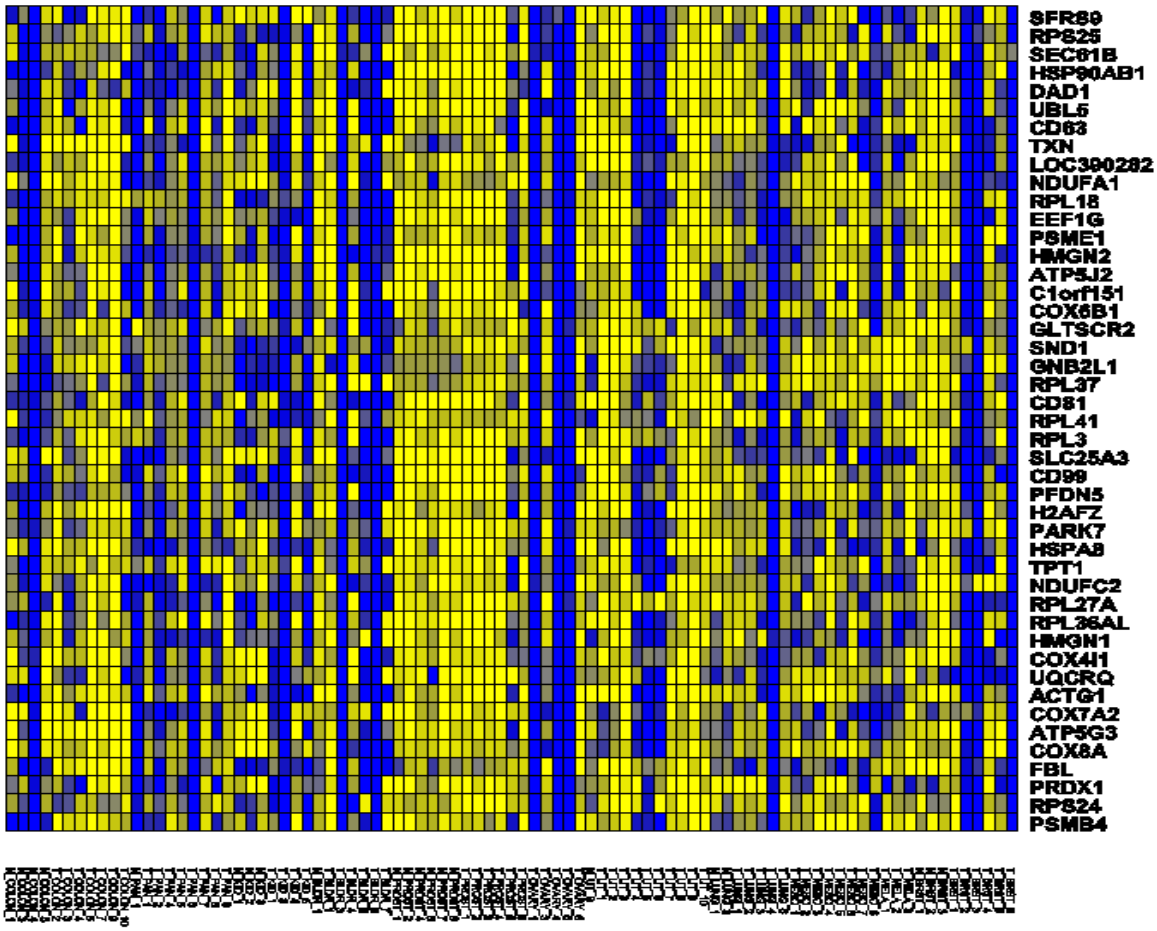
Şekil 6.6 Çoklu kanser modül ağında 42 nolu modül için ısı haritası

Şekil 6.6'da belirtilen gruptaki genlerin prostat dokularında farklı gen ifade değerleriyle ifade edildiği net bir şekilde görülebilmektedir. Modüldeki GO terimleriyse nöropeptid sinyal yolu ve hormonal metabolik süreçlerdir. Prostat dokularında gen ifade verilerini yükselten yönde etki gösteren başka bir örnek şekil 6.7'de verilmiştir.



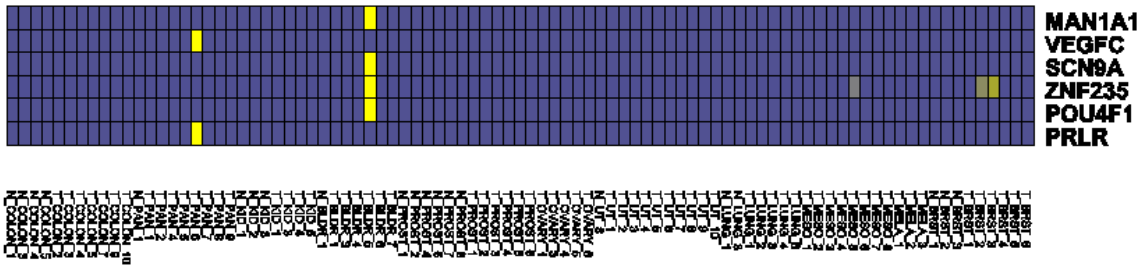
Şekil 6.7 Çoklu kanser modül ağında 15 nolu modül için ısı haritası

Bu düzenlemenin transkripsiyon faktörü YY1 ile yapıldığı görülebilmektedir. Başka örneklerde de tutarlı davranışlar gözlenmiştir.



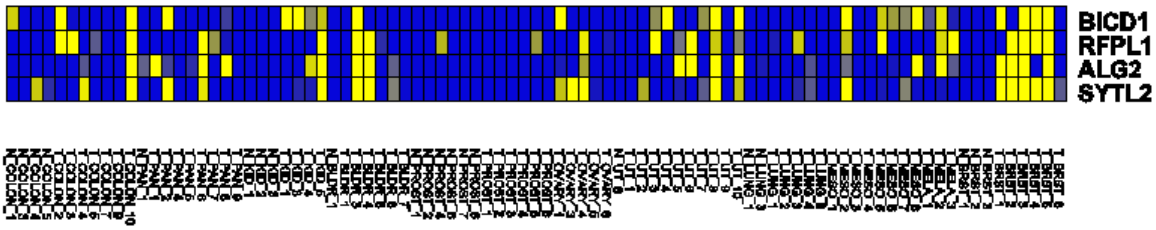
Şekil 6.8 Çoklu kanser modül ağında 10 nolu modül için ısı haritası

Şekil 6.8'da görüldüğü gibi prostat kanseri örneklerinde artış yönlü düzenleme gözlenmiştir. İlişkilendirilen GO terimleri arasında translasyon, translasyon sürdürmesi, öncü metabolitler ve enerji üretimi gösterilebilir. Diğer örneklerden alınan genler bu küme içerisinde benzer davranışlar göstermektedir.



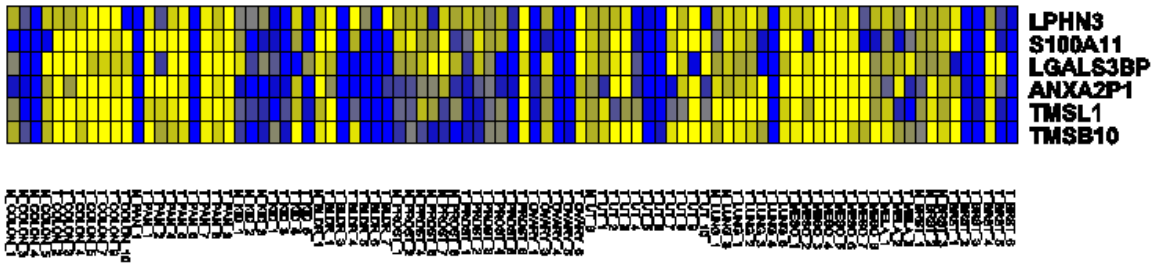
Şekil 6.9 Çoklu kanser modül ağında 52 nolu modül için ısı haritası

Şekil 6.9'da görülebildiği gibi kümede artış yönlü düzenleme gösteren bir tek örnek vardır. Bu bir tek örnek incelenmeli ve artış yönlü düzenleme göstermesi sebebinin ne olduğu araştırılmalıdır. GO terimleri analizinden ilgi çekici bir sonuç çıkmamıştır.



Şekil 6.10 Çoklu kanser modül ağında 54 nolu modül için ısı haritası

Normal ve hastalıklı örneklerde gen düzenlemesinin başka bir örneği şekil 6.10'da görülmektedir. Normal kişilerden alınan meme kanseri örneklerinde azalma yönlü düzenleme gözlenirken, hastalıklı örneklerde artış yönlü düzenleme gözlenmiştir. GO terimleri analizinden çıkan ilginç sonuçsa makromolekül sınırlandırmasıdır.



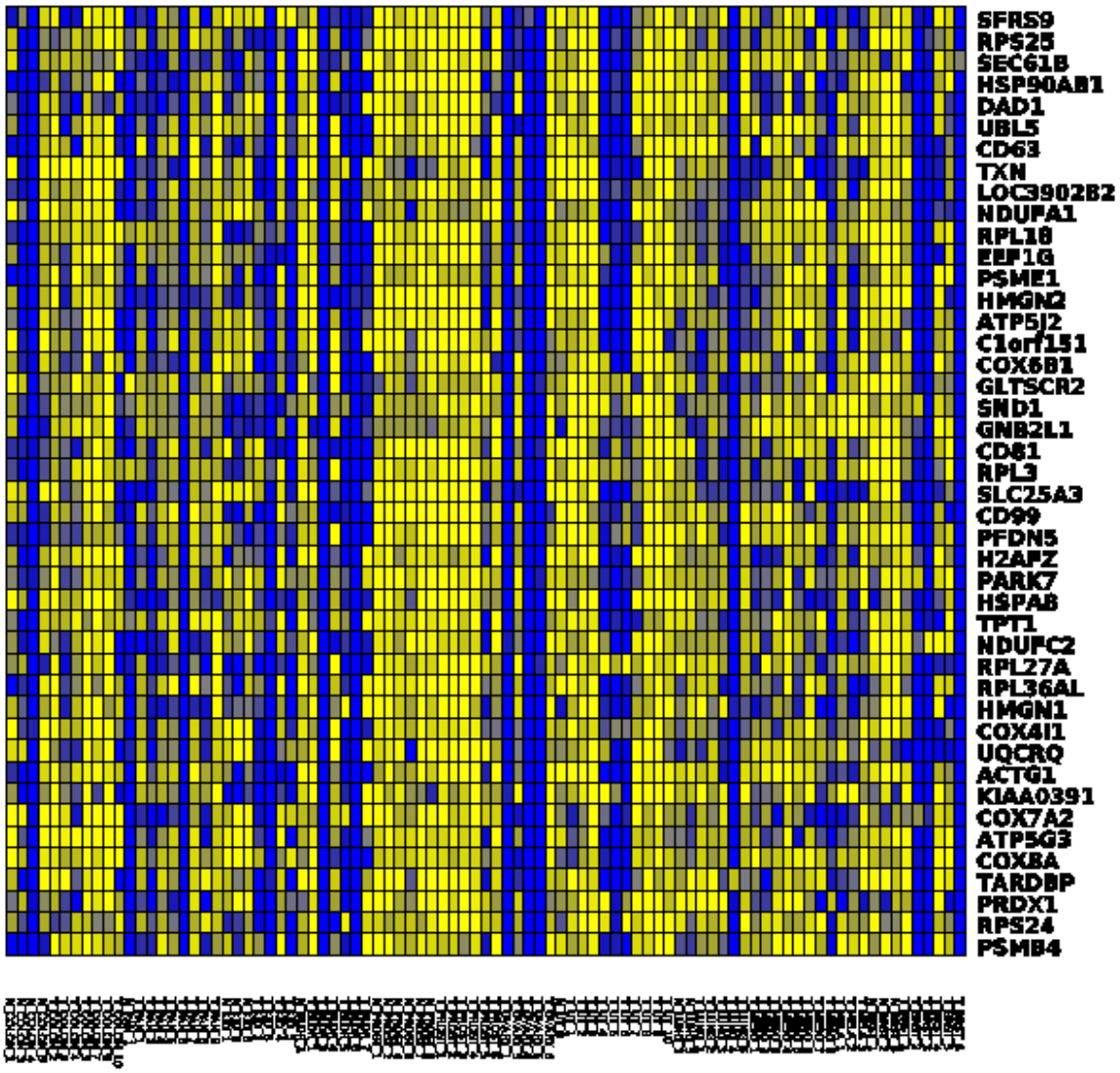
Şekil 6.11 Çoklu kanser modül ağında 83 nolu modül için ısı haritası

Şekil 6.11'de normal kolon kanseri örneklerinde azaltıcı yönlü düzenleme, hasta kolon kanseri örneklerindeyse artma yönlü düzenleme görülmüştür. GO terimleri analizinde birden fazla sonuç çıkmasına rağmen ilgi çekici olan 'hücresel metabolik süreçlerin negatif düzenlenmesi' dikkat çekmektedir. Bir tek kümedeki ANXA2P1 geni düzenleyici olarak bilinmemektedir. Bu yüzden araştırılıp, incelenmelidir.

6.1.2. Düzenleyici miRNA alındığında

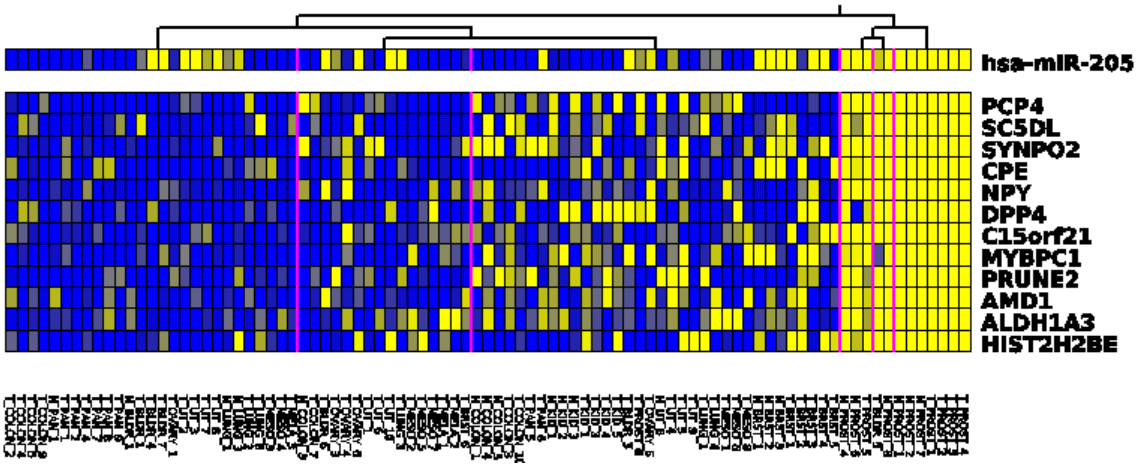
Çoklu kanser veri kümesinden 86 modül çıkarılmıştır. Bu modüllerin en az 26'sında en az bir GO terimi zenginleştirilmiştir. Modüllerden bazıları belirli kanser tiplerine ilşkin ilginç mRNA ve miRNA grupları ortaya çıkarmıştır. Modüller ısı haritalarına ve GO terimi zenginleştirme analizine göre değerlendirilmiştir.

Meme kanseri veri kümesindense 259 modül elde edilmiştir. Bu modüllerden 252 tanesinde en az bir düzenleyici miRNA modül ataması yapılmıştır. Modüllerin 144 tanesinde en az bir GO teriminin zenginleştiği, bunların %90.3'ünde zenginleşen GO terimi sayısının birden fazla olduğu görülmüştür.



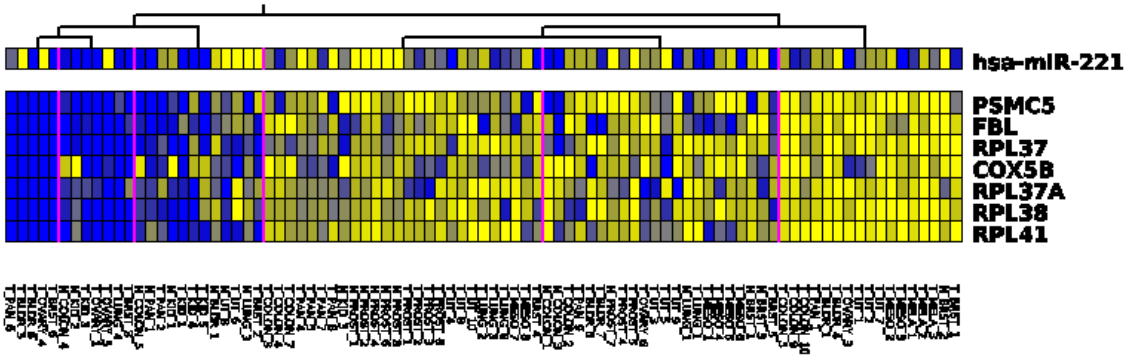
Şekil 6.12 Çoklu kanser modül ağında 9 nolu modül için ısı haritası

Şekil 6.12'da görüldüğü üzere hastalıklı ve normal prostat örneklerinde artma yönlü düzenleme görülmüştür. GO terimleri analizinden birkaç gende translasyonal sürdürme çıkmıştır. GO terimleri analizi, bu genlerin aynı kümede olmasını desteklemektedir. Aynı kümedeki genler benzer görevleri yapabileceğinden şekildeki genler araştırılmalıdır.



Şekil 6.13 Çoklu kanser modül ağında 35 nolu modül için ısı haritası

Şekil 6.13'de bir miRNA tarafından düzenlenen genler görülmektedir. GO terimleri analizinden ilginç bir sonuç çıkmamıştır. Ancak hsa-miR-205 miRNA'sındaki değişimlerin kümedeki genleri yaklaşık olarak orantılı bir şekilde etkilediği görülebilmektedir. miRNA'daki azalma hasta ve normal prostat kanserleri üzerinde artma yönlü düzenlemeye sebep olmuştur denilebilir.



Şekil 6.14 Çoklu kanser modül ağında 46 nolu modül için ısı haritası

Şekil 6.14'deki kümedeki genlerin çoğunun gen ifadesinin düzenlenmesinde görev aldığı bilinmektedir. Bu genlerin bir miRNA tarafından düzenlendiği bilgisiyse önemli olabilir. GO analizinden çıkan sonuçlar hücresel ve makromolekül biyosentetik süreç olmaktadır. miRNA'nın bu genleri düzenleyip düzenlemediği biyologlar tarafından araştırılabilir.

6.2. Çıkarılan Motifler

Dizilim bilgisi kullanılarak birleştirilmiş modül ağlardan alınan düzenleyici ikilileri doğrulanmıştır. Sonuçları daha güvenilir yapmak için, çıkarılan son modül ağda düzenleyicilerin belli bir *p*-değerine göre yüzde onu alınmıştır. CircuitsDB tarafından sağlanan verilerle düzenleyici ikili motiflerinin doğrulanmış en son hali şekil 6.15 ve 6.16'da görülmektedir.

TF->mRNA düzenleyici motifleri

TBP -> CXADR	ATF6 -> NUP133	ATF6 -> PSMD6	SRF -> HIST2H2AB
AHR -> SCGB2A1	ATF6 -> RBM7	ATF6 -> RNF187	SRF -> HIST1H2AG
YY1 -> MTMR3	ATF6 -> TMEM9	ATF6 -> RAG1AP1	STAT1 -> HLA-E
YY1 -> ACO2	ATF6 -> GNPAT	ATF6 -> ZBTB41	MAZ -> NME4
YY1 -> TRPT1	ATF6 -> TMEM85	ATF6 -> MORF4L1	HSF2 -> PHB
YY1 -> SLC35B2	ATF6 -> DNAJB9	ATF6 -> RBM34	HSF2 -> ZNF607
YY1 -> MRPL40	ATF6 -> C3orf10	ATF6 -> FAM118B	IRF1 -> HLA-J
YY1 -> LYRM5	ATF6 -> ISG20L2	ATF6 -> CLIC1	
BACH2 -> SELE	ATF6 -> ADAR	CRX -> SEC14L3	
BACH2 -> TXNDC13	ATF6 -> PMVK	CRX -> C22orf24	

miRNA->mRNA düzenleyici motifleri

hsa-miR-106a -> CROT	hsa-miR-542-3p -> WDR1
hsa-miR-106a -> DGAT2	hsa-miR-497 -> NUCKS1
hsa-miR-30d -> UCP3	hsa-let-7b -> SMC1A
hsa-miR-27a -> PLEKHJ1	hsa-let-7d -> AP1S1
hsa-miR-125b -> ALDH1A3	hsa-miR-141 -> OPTN
hsa-miR-125b -> C17orf51	hsa-miR-141 -> KHDRBS3
hsa-miR-141 -> PPARA	hsa-miR-26a -> GGA2
hsa-miR-9 -> LDLRAP1	hsa-miR-26a -> RG9MTD2
hsa-miR-155 -> GMNN	hsa-miR-26a -> PDCD6IP
hsa-miR-155 -> SLFN11	hsa-miR-30c -> POLR3E
hsa-miR-23a -> GTF2E2	hsa-miR-142-3p -> SMR3A
hsa-miR-194 -> IL6ST	

Şekil 6.15 Meme Kanseri Veri Kümesi için Çıkarılan Düzenleme Motifleri

TF->mRNA düzenleyici motifleri			
NKX2-5->PTPN21	NKX2-5->TTR	NKX2-5->GPR109B	YY1->EMG1
NKX2-5->MADCAM1	NKX2-5->C1orf61	NKX2-5->COL4A5	YY1->PIAS3
NKX2-5->CTSG	NKX2-5->SAG	NKX2-5->SPRR2A	YY1->DNAJB1
NKX2-5->CHGA	NKX2-5->KRT33B	NKX2-5->ZNF585A	YY1->CCM2
NKX2-5->GUCY2F	NKX2-5->GNAT2	NKX2-5->MYO5A	YY1->CAMLG
NKX2-5->FLT1	NKX2-5->IFI44	NKX2-5->ZNF652	YY1->TUBA1A
NKX2-5->CCL22	NKX2-5->SLC39A8	NKX2-5->CRYZL1	YY1->ZNF613
NKX2-5->CDR2L	NKX2-5->SLC7A1	YY1->NDRG3	YY1->SLC35A4
NKX2-5->CCKBR	NKX2-5->LTC4S	YY1->F3	YY1->PRMT6
NKX2-5->TNNT2	NKX2-5->KRT86	YY1->WDR34	
miRNA->mRNA düzenleyici motifleri			
hsa-miR-103->CYBASC3	hsa-miR-125b->BCL2	hsa-miR-195->KIF1B	hsa-miR-205->YES1
hsa-miR-125b->TJAP1	hsa-miR-125b->POFUT2	hsa-miR-195->ATG4B	hsa-miR-9*->NOS1

Şekil 6.16 Çoklu Kanser Veri Kümesi için Çıkarılan Düzenleme Motifleri

Şekillerde ilk önce düzenleyicinin adı sonrasında düzenlenen genin adı verilmiştir. Bu liste, ıslak labarotuvar deneyleriyle test edilebilir bir hipotez olarak düşünebilir. Bu analizde tüm düzenleyicilerin seçildiği durumda üç adet TF->miRNA ikilisi tespit edilebilmiştir: MAZ->hsa-let-7b, MEIS1->hsa-let-7e, MEIS1->hsa-let-7b.

7. TARTIŞMA VE GELECEK ÇALIŞMALAR

Gen düzenleme analizinde kullanılabilmesi için bütünleştirici bir program geliştirilmiştir. Program biyolojik olarak tutarlı kümeleri ve düzenleyicilerini bulmak için sayısal yöntemler bütününü içerir. Geliştirilen araç bazı sebeplerle avantajlıdır. Eşleştirilmiş mRNA ve miRNA gen ifade örnekleri mevcutsa transkripsiyon sırasında ve transkripsiyon sonrasında düzenlemeyi tek bir programda yapabilmektedir. Bu bazı düzenleme özellikleri, transkripsiyon faktörleri, mikro RNA ve ortak genler üzerinde birleştirilmiş bir görünüm sağlamaktadır. Gen düzenlemesinin ortak bir şekilde olma olasılığı olduğundan son ağda çoka-çok ilişkiler de dikkate alınmaktadır. Genler ve düzenleyiciler bu şekilde çalıştığından, bu yaklaşım gen düzenlemesine daha gerçekçi bir bakış açısı sunmaktadır. Tahmin edilen etkileşimlerin güvenirliliği dizilim verisinin kullanımıyla artırılmıştır. Dizilim verisi tek başına kullanılamaz çünkü dizilim bağlanması tüm durumlarda gen düzenlemesi nedeni olamaz. Gen ifadesi seviyesindeki değişikliğe rastgele bir dalgalanma veya farklı faktörler neden olabilir. Sonuç olarak, dizilim ve gen ifade verisinin bir arada kullanımı daha güvenilir sonuçlar vermiştir. Geliştirilen araç kullanılarak, birçok kanser hücresi içeren veri kümesinden biyolojik olarak kullanılabilecek sonuçlar çıkartılabileceği gösterilmiştir. Bazı sonuçlar ıslak laboratuar doğrulama yöntemi kullanılarak geçerli kılınabilecek şekilde bırakılmıştır.

Geliştirilen araç kullanılarak benzer görevleri yapan genler olasılıksal olarak kümelenebilir, düzenleyicileri belirlenebilir. Bu biyologlara büyük bir avantaj sağlamaktadır. Canlılardaki gen yapısı düşünüldüğünde, en basit canlı genindeki biyolojik bilgi bile büyük bir kütüphaneyi doldurabilecek büyüklüktedir. Bu bilginin insan algısına uyacak küçük parçalara ayrılması gerekmektedir çünkü hepsi bir anda anlaşılamaz. Biyologların her bir gen içindeki bilgiyi ayrı ayrı araştırması düşünülemez çünkü genlerin içinde bir sürü bilgi vardır ve çok uzun zaman almaktadır. Aracı kullanarak biyologlar en olası kümeleri ve düzenleyicileri olasılıksal olarak bulup araştırmalarına o yönde devam edebilirler. Bir nevi yol gösterici olarak düşünülebilir.

Geliştirilen yöntem ve yazılım aracı pek çok hücresel döngü ve mekanizmanın belirlenmesine yardımcı olacak ve bu yöntemlerin kullanımını sağlayacak görsel web araçları, kanser araştırmaları ve ilaç tasarımı gibi biyotıp çalışmalarına katkı sağlayacaktır. Yazılım aracı özellikle biyologların genomik veri analizi, türler arası benzerliklerin araştırılması ve çeşitli hastalıkların tespiti vb. çalışmalarında işlerini kolaylaştırıcı bir araç haline gelecektir.

Yöntem, şu ana kadar geliştirilen düzenleyici atama araçlarından daha kesin sonuçlar vermektedir çünkü gen ifade verisinin yanında dizilim verisini de kullanmaktadır. Gen ifade verisi ve dizilim verisinin aynı araçta kullanılması yeni bir gelişmedir.

Veri tabanlarında aynı örneklere uygulanan mRNA ve miRNA deneylerinin sayısı arttıkça daha fazla karmaşık hesaplamalara dayalı analiz yöntemlerine ihtiyaç duyulacaktır. Böylece bu ve benzeri çalışmaların daha fazla önem kazanacağı düşünülmektedir. Yapılan çalışma kullandığı yöntemler bakımından makine öğrenme, istatistiksel veri analizi ve paralel veri işleme kapsamında sayılabilir.

Gelecek çalışmalar için ise geliştirilen yöntem, benzer veri tipleri üzerinde tanımlı farklı sorunlar için de uygulanabilir olacaktır.

KAYNAKÇA

[1] Ke Y., Stuart L., Yung C., Yan L., Hao Y. Self-assembled water-soluble nucleic acid probe tiles for label-free RNA hybridization assays.// Science – № 5860 (319), 2008 – P.180-183.

[2] Mesajcı RNA, http://tr.wikipedia.org/wiki/Mesajc%C4%B1_RNA, Erişim Tarihi: 28.01.2014

[3] Clustering to Improve Merchandise Allocation, Testing, and Forecasting: An Application of the K-Medians Algorithm, http://espin086.wordpress.com/2011/02/27/clustering-to-improve-merchandise-allocation-testing-and-forecasting-an-application-of-the-k-medians-algorithm/, Erişim Tarihi: 28.01.2014

[4] A. Joshi, Y. Van de Peer, T. Michoel, "Analysis of a Gibbs sampler method for model-based clustering of gene expression data", Bioinformatics 24: 176–183, 2008.

[5] O. Friard, A. Re, D. Taverna, M. De Bortoli, D. Cora, "CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse", BMC Bioinformatics 11, 435, 2010.

[6] Using the Gene Ontology (GO) for analysis of gene expression data, Jane Lomax EMBL-EBI, Slayt 32

[7] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25:25-29.

[8] Friedman N, Linial M, Nachman I, Pe'er D: Using Bayesian Networks to analyze expression data. J Comput Biol 2000, 7:601-620.

[9] Pe'er D, Regev A, Elidan G, Friedman N: Inferring subnetworks from perturbed expression profiles. Bioinformatics 2001, 17(Suppl 1):S215-S224.

[10] Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK: Computational discovery of gene modules and regulatory networks. Nat Biotechnol 2003, 21:1337-1342.

[11] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 2003, 34:166-167.

[12] Beer MA, Tavazoie S: Predicting gene expression from sequence. Cell 2004, 117:185-198.

[13] Friedman N: Inferring cellular networks using probabilistic graphical models. Science 2004, 303:799-805.

[14] Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: Transcriptional regulatory code of a eukaryotic genome. Nature 2004, 431:99-104.

[15] Luscombe NM, Madan Babu M, Yu H, Snyder M, Teichmann SA, Gerstein M: Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 2004, 431:308-312.

[16] Xu X, Wang L, Ding D: Learning module networks from genome-wide location and expression data. FEBS Lett 2004, 578:297-304.

[17] Battle A, Segal E, Koller D: Probabilistic discovery of overlapping cellular processes and their regulation. J Comput Biol 2005, 12:909-927.

[18] Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: Reverse engineering of regulatory networks in human B cells. Nat Genet 2005, 37:382-390.

[19] Garten Y, Kaplan S, Pilpel Y: Extraction of transcription regulatory signals from genome-wide DNA-protein interaction data. Nucleic Acids Res 2005, 33:605-615.

[20] Petti AA, Church GM: A network of transcriptionally coordinated functional modules in Saccharomyces cerevisiae. Genome Res 2005, 15:1298-1306.

[21] Lemmens K, Dhollander T, De Bie T, Monsieurs P, Engelen K, Smets B, Winderickx J, De Moor B, Marchal K: Inferring transcriptional modules from ChIP-chip, motif and microarray data. Genome Biol 2006, 7:R37.

[22] Van den Bulcke T, Lemmens K, Van de Peer Y, Marchal K: Inferring transcriptional networks by mining 'omics' data. Current Bioinformatics 2006, 1:301-313.

[23] Hartwell LH, Hopfield JJ, Leibler S, Murray AW: From molecular to modular cell biology. Nature 1999, 402:C47-C52.

[24] Segal E, Friedman N, Kaminski N, Regev A, Koller D: From signatures to models: understanding cancer using microarrays. Nat Genet 2005, 37:S38-S45.

[25] Cluster Analysis: see it 1st, http://apandre.files.wordpress.com/2011/08/kmeansclustering.jpg, Erişim Tarihi: 28.01.2014

[26] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angeloet al., "Multiclass cancer diagnosis using tumor gene expression signatures", Proc. Natl. Acad. Sci., 98, 15149-15154, 2001.

[27] J. Lu, G. Getz, E.A. Miska et al., "MicroRNA expression profiles classify human cancers", Nature, 435, 834–838, 2005.

[28] R.C. Lee, R.L. Feinbaum, V. Ambros, The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, Cell 75. 843-854, 1993.

[29] Ruvkun G, Molecular biology: Glimpses of a tiny RNA world, Science 294, 797-799, 2001.

[30] Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM., Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs, Nature 17,769-773, 2005.

[31] R.F. Place, Long-Cheng Li, D Pookot, EJ Noonan, R Dahiya, MicroRNA-373 induces expression of genes with complementary promoter sequences, PNAS 105, 1608-1613, 2008.

[32] He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, Powers S, Cordon-Cardo C, Lowe SW, Hannon GJ, Hammond SM, A microRNA polycistron as a potential human oncogene, Nature 435, 828-833, 2005.

[33] O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT, c-Myc-regulated microRNAs modulate E2F1 expression, Nature 435, 839–843, 2005.

[34] Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG, Lost in translation: an assessment and perspective for computational microRNA target identification, Bioinformatics 25, 3049-3055, 2009.

[35] N. D. Mendes, A. T. Freitas, M.-F. Sagot, Current tools for the identification of miRNA genes and their targets, Nucleic Acid Res. 37, 2419-2433, 2009.

[36] Hammell M, Computational methods to identify miRNA targets, Semin. Cell Dev. Biol., 2010 (in press).

[37] Stark A, Brennecke J, Russell RB, Cohen SM, Identification of drosophila microRNA targets, PLOS Biol. 1, E60, 2003.

[38] Lewis BP, Burge CB, Bartel DP, Conserved seed pairing, often flanked by adenisines, indicates that thousands of human gene are microRNA targets, Cell 120, 15-20, 2005.

[39] Friedman RC, Farh KK, Burge CB, Bartel DP, Most mammalian mRNAs are conserved targets of microRNAs, Genome Res. 19, 1-11, 2009.

[40] Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP, MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing, Molecular Cell 27, 91-105, 2007.

[41] Ruan J, Chen H, Kurgan L, Chen K, Kang C, Pu P, HuMiTar: A sequence-based method for prediction of human microRNA targets, Algorithms in Molecular Biology 3,16, 2008.

[42] Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I, A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes, Cell 126, 1203-1217, 2006.

[43] D Gaidatzis, EV Nimwegen, J Hausser, M Zavolan, Inference of miRNA targets using evoluionary conservation and pathway analysis, BMC Bioinformatics 8, 69-79, 2007.

[44] Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Vergoulis T, Koziris N, Sellis T, Tsanakas P, Hatzigeorgiou AG, DIANA-microT web server: elucidating microRNA functions through target prediction, Nucleic Acid Res. 37, W273-W276, 2009.

[45] John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS, Human microRNA targets, PLOS Biol. 2, 1862-1879, 2004.

[46] Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R, Fast and effective prediction of microRNA-target dublexes, RNA 10, 1507-1517, 2004.

[47] Hofacker IL, How microRNAs choose their targets, Nature Genet. 39, 1191-1192, 2007.

[48] M Kertesz, N Iovino, U Unnerstall, U Gaul, E Segal, The role of site accesibility in microRNA target recognition, Nature Genet. 39, 1278-1284, 2007.

[49] Hammell M, Long D, Zhang L, Lee A, Carmack CS, Han M, Ding Y, Ambros V, mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts, Nature Genet. 39, 1278-1284, 2008.

[50] Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N, Combinatorial microRNA target predictions, Nature Genet. 37, 495-500, 2005.

[51] Saetrom O, Snøve O Jr, Saetrom P, Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms, RNA 11, 995-1003, 2005.

[52] Yousef M, Jung S, Kossenkov AV, Showe LC, Showe MK, Naïve Bayes for microRNA target predictions-machine learning for microRNA targets, Bioinformatics 23, 2987-2992, 2007.

[53] V Chandra, R Girijadevi, AS Nair, SS Pillai, RM Pillai, MTar: a computational microRNA target prediction architecture for human transcriptome, BMC Bionformatics 11, S2, 2010.

[54] Bandyopadhyay and Mitra, TargetMiner: MicroRNA target prediction with systematic identification of tissue specific negative examples, Bioinformatics 25, 2625-2631, 2009.

[55] Bartel DP, MicroRNAs:target recognition and regulatory functions, Cell 136, 215-233, 2009.

[56] Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N, Widespread changes in protein synthesis induced by microRNAs, Nature 455, 58-63, 2008.

[57] Slack FJ,Big roles for small RNAs, Nature 463, 616.

[58] Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, Hughes TR, Blencowe BJ, Frey BJ, Morris QD, Using expression profiling data to identify human microRNA targets, Nature Methods 4, 1045-1049, 2007.

[59] Huang JC, Morris QD, Frey BJ, Bayesian inference of microRNA targets from sequence and expression data, J. Comput.Biol. 14, 550–563, 2007.

[60] Wang and Wang, Systematic identification of microRNA functions by combining target prediction and expression profiling, Nucleic Acid Res. 34, 1646-1652, 2006.

[61] Ritchie W, Rajasekhar M, Flamant S, Rasko JEJ, Conserved expression patterns predict microRNA targets, PLOS Comput. Biol. 5, e1000513, 2009.

[62] Wang and Naqa, Prediction of both conserved and nonconserved microRNA targets in animals, Bioinformatics 24, 325-332, 2008.

[63] Cheng and Li, Inferring microRNA activities by combining gene expression with microRNA target prediction, PLoS ONE 3, 1-9, 2008.

[64] Creighton CJ, Nagaraja AK, Hanash SM, Matzuk MM, Gunaratne PH, A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions, RNA 14, 2290–2296, 2008.

[65] Nam S, Kim B, Shin S, Lee S, miRGator: an integrated system for functional annotation of microRNAs,‖ Nucleic Acids Res. 36, D159–D164, 2008.

[66] Joung and Fei, Computational identification of condition-specific miRNA targets based on gene expression profiles and sequence information, BMC Bioinformatics 10, S34, 2009.

[67] Drakaki and Iliopoulos, MicroRNA Gene Networks in Oncogenesis, Current Genomics 35-41, 2009.

[68] Yoon and Micheli, Prediction of regulatory modules comprising microRNAs and target genes, Bioinformatics 21, i93-i100, 2005.

[69] Joung JG, Hwang KB, Nam JW, Kim SJ, Zhang BT, Discovery of microRNA-mRNA modules via population-based probabilistic learning, Bioinformatics 23, 1141-1147, 2007.

[70] Joung and Fei, Identification of microRNA regulatory modules in Arabidopsis via a probabilistic graphical model, Bioinformatics 25, 387-393, 2009.

[71] Tran DH, Satou K, Ho TB, Finding microRNA regulatory modules in human genome using rule induction, BMC Bioinfromatics 9:S5, 2008.

[72] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel- Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25:25-29.

[73] Inoue,K. and Urahama,K. (1999) Sequential fuzzy cluster extraction by a graph spectral method. Pattern Recognit. Lett., 20, 699–705.

[74] Elena Deza & Michel Marie Deza, Encyclopedia of Distances, page 94, Springer, (2009).

[75] Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society, Series B 39 (1): 1–38.

[76] Eugene Lukacs (1942). "A Characterization of the Normal Distribution". The Annals of Mathematical Statistics 13 (1): 91–93.