

**YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**GEN KÜMELEME İŞLEMİNİN ÖZDÜZENLEYİCİ
HARİTALAR KULLANILARAK GEN EKSPRESYONU,
MOTİF SIKLIK VE GEN KONUM VERİLERİNDEN
FAYDALANILARAK GERÇEKLEŞTİRİMİ**

Bilgisayar Müh. Alper ÖZDOĞAN

FBE Bilgisayar Mühendisliği Anabilim Dalı Programında Hazırlanan

YÜKSEK LİSANS TEZİ

Tez Danışmanı : Yrd. Doç. Dr. Songül ALBAYRAK

İSTANBUL, 2009

İÇİNDEKİLER

	Sayfa
SİMGE LİSTESİ	IV
KISALTMA LİSTESİ.....	V
ŞEKİL LİSTESİ.....	VI
ÇİZELGE LİSTESİ	VII
ÖNSÖZ	X
ÖZET	XI
ABSTRACT	XII
1. GİRİŞ.....	1
1.1 Motivasyon.....	2
1.2 Problemin Tanımı	3
1.3 Hipotez.....	4
1.4 Hedefler ve Amaçlar	5
1.5 Tezin Yapısı	6
2. TEMEL KAVRAMLAR	7
2.1 Temel Genetik Kavramları	7
2.1.1 Moleküler Biyolojinin “Temel Dogması”	10
2.1.2 Gen Ekspresyonu Verisi	11
2.1.3 Transkripsiyon Faktörleri, Transkripsiyon Faktörü Bağlanma Alanları ve Motifler	12
2.1.4 Genlerin Kromozomal Konumu.....	13
2.2 Kümeleme Yöntemleri.....	15
2.2.1 SOM Yapay Sinir Ağları: Özdüzenleyici Haritalar	16
2.2.1.1 SOM Mimarisi.....	16
2.2.1.2 SOM Algoritması	19
3. İLGİLİ ÇALIŞMALAR.....	20
3.1 Gen Ekspresyon Verisinin Kümelenmesi	20
3.2 Bilgi Füzyonu Kullanılarak Çoklu Tipte Genetik Verinin Kümelenmesi	21
3.2.1 Algoritma Detayları.....	23
4. KULLANILAN VERİ SETİ VE METODLAR.....	25
4.1 Veri Seti	25
4.2 Veri Tipleri ve İlgili Uzaklık Fonksiyonları	26
4.2.1 Arabidopsis Gen Ekspresyon Verisi.....	26

4.2.2	Arabidopsis Motif Sıklık Verisi	28
4.2.3	Arabidopsis Gen Konum Verisi	29
4.3	Algoritmanın Uygulamaya Dökülüşü	32
4.4	Algoritmanın Genişletilmesi/Geliştirilmesi: SOM Yapay Sinir Ağı Modeli ile Bilgi Füzyonu ve Kümeleme	32
4.5	Kullanılan SOM'un Tasarımı	33
4.6	Sonuçların Değerlendirilmesi	36
5.	ELDE EDİLEN SONUÇLAR VE ANALİZLER	42
5.1	Gen Ekspresyon Verisi ile Elde Edilen Kümeleme Sonuçları	42
5.2	Gen Ekspresyon, Motif Sıklık ve Gen konum Verilerinin Birarada Kullanılması ile Elde Edilen Kümeleme Sonuçları	54
5.2.1	Küme "0,0"a Ait Genler ve Analiz Sonuçları	59
5.2.2	Küme "0,1"e Ait Genler ve Analiz Sonuçları	61
5.2.3	Küme "0,2"ye Ait Genler ve Analiz Sonuçları	63
5.2.4	Küme "0,3"e Ait Genler ve Analiz Sonuçları	64
5.2.5	Küme "0,4"e Ait Genler ve Analiz Sonuçları	65
5.2.6	Küme "1,1"e Ait Genler ve Analiz Sonuçları	66
5.2.7	Küme "1,2"ye Ait Genler ve Analiz Sonuçları	67
5.2.8	Küme "1,3"e Ait Genler ve Analiz Sonuçları	68
5.2.9	Küme "1,4"e Ait Genler ve Analiz Sonuçları	68
5.2.10	Küme "2,2"ye Ait Genler ve Analiz Sonuçları	69
5.2.11	Küme "2,4"e Ait Genler ve Analiz Sonuçları	69
5.2.12	Küme "3,0"a Ait Genler ve Analiz Sonuçları	70
5.2.13	Küme "3,1"e Ait Genler ve Analiz Sonuçları	71
5.2.14	Küme "3,3"e Ait Genler ve Analiz Sonuçları	71
5.2.15	Küme "3,4"e Ait Genler ve Analiz Sonuçları	72
5.2.16	Küme "4,2"ye Ait Genler ve Analiz Sonuçları	72
5.2.17	Küme "4,3"e Ait Genler ve Analiz Sonuçları	73
5.2.18	Küme "4,4"e Ait Genler ve Analiz Sonuçları	74
6.	SONUÇLAR VE ÖNERİLER	76
	KAYNAKLAR	79
	EKLER	83
	Ek 1 Kullanılan Genler	84
	Ek 2 Geliştirilen Uygulama Programı	85
	ÖZGEÇMİŞ	91

SİMGE LİSTESİ

c	Küme merkezi
$D()$	Uzaklık fonksiyonu
d	Uzaklık
g	Gen
k	Kategori
l	Gen g 'ye en yakın küme
max	Maksimum
min	Minimum
P	Kategori ağırlık vektörü
p_i	i . kategori ağırlığı
R	SOM komşuluk yarıçapı
W	Nöron ağırlık vektörü
α	Öğrenme oranı

KISALTMA LİSTESİ

A	Adenin
C	Sitozin
DNA	Deoksiribonükleik Asit
G	Guanin
GLDist	Gene Location Distance
GO	Gen Ontolojisi
ID	Identity (Kimlik)
KL	Kullback-Leibler
mRNA	Mesajcı Ribonükleik Asit
RNA	Ribonükleik Asit
SOM	Self Organizing Map (Özdüzenleyici Harita)
T	Timin
TAIR	The Arabidopsis Information Resource
VAU	Vektör Açı Uzaklığı

ŞEKİL LİSTESİ

	Sayfa
Şekil 2.1 DNA'nın çift sarmallı yapısı ve DNA'yı oluşturan nükleotidler [3].	8
Şekil 2.2 Kromozom, DNA ve genler [4].	8
Şekil 2.3 DNA, gen, kodon ve aminoasitler ile bir proteine ait genetik kodun hiyerarşik biçimde gösterimi [3].	9
Şekil 2.4 Moleküler biyolojinin "Temel Dogması" olarak adlandırılan, hücre içi genetik bilgi akışı (Setubal ve Meidanis, 1997).	10
Şekil 2.5 Transkripsiyon faktörleri, bağlanma alanları ve motifler [5].	13
Şekil 2.6 (a) Uzun ve kısa kolları ile bir kromozomun yapısı [6]. (b) Örnek bir gen için kromozomal lokasyonunun detaylı gösterimi [7].	14
Şekil 2.7 N adet öz niteliğe sahip girdi veri seti için tasarlanmış 3x3 boyutlu SOM'un yapısı.	17
Şekil 2.8 7x7 SOM üzerinde komşuluk ilişkilerinin dikdörtgensel topoloji ve 2, 1, 0 yarıçapları için gösterimi.	18
Şekil 4.1 Gen konum bilgisini kümeleme işleminde kullanabilmek amacı ile, gen ve kümelere (nöronlara) ait kromozom numaralarını karşılaştırabilmek için bulunan çözüm.	31
Şekil 4.2 Kullanılan SOM'un detaylı yapısı.	36
Şekil 4.3 GO:0016597 ID numaralı örnek moleküler fonksiyonunun, moleküler fonksiyon anotasyon ağacı üzerindeki yeri [2].	39
Şekil 4.4 (a) GO:0048863 ID numaralı örnek biyolojik prosesin, biyolojik proses anotasyon ağacı üzerindeki yeri [2]. (b) GO:0044429 ID numaralı örnek hücresel bileşenin, hücresel bileşen anotasyon ağacı üzerindeki yeri [2].	39
Şekil 5.1 Gen ekspresyon verisi için elde edilen kümelerin ve küme eleman sayılarının 2 boyutlu harita üzerinde gösterimi.	44
Şekil 5.2 Kombine genomik verinin (gen ekspresyon verisi, motif sıklık verisi, gen konum verisi) kullanımı ile elde edilen kümelerin ve küme eleman sayılarının 2 boyutlu harita üzerinde gösterimi.	55

ÇİZELGE LİSTESİ

	Sayfa
Çizelge 4.1 Gen ekspresyon verisi örneği.	26
Çizelge 4.2 Motif sıklık verisi örneği.	28
Çizelge 4.3 Gen konum verisi örneği.	29
Çizelge 5.1 Gen ekspresyon verisi için farklı çalıştırmalar sonucu elde edilen toplam küme sayıları ve tek elemana sahip küme sayıları.	43
Çizelge 5.2 Gen ekspresyon verisi için kümeleme sonuçları ve kayda değer gen benzerlikleri.	45
Çizelge 5.3 Küme 0,0'a ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	47
Çizelge 5.4 Küme 0,1'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	47
Çizelge 5.5 Küme 0,2'ye ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	48
Çizelge 5.6 Küme 0,3'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	48
Çizelge 5.7 Küme 0,4'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	48
Çizelge 5.8 Küme 1,0'a ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	49
Çizelge 5.9 Küme 1,1'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	49
Çizelge 5.10 Küme 1,2'ye ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	49
Çizelge 5.11 Küme 1,3'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	50
Çizelge 5.12 Küme 1,4'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	50
Çizelge 5.13 Küme 2,0'a ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	50
Çizelge 5.14 Küme 2,1'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	51
Çizelge 5.15 Küme 2,2'ye ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	51
Çizelge 5.16 Küme 2,3'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	51
Çizelge 5.17 Küme 2,4'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	51
Çizelge 5.18 Küme 3,0'a ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	51
Çizelge 5.19 Küme 3,1'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	52
Çizelge 5.20 Küme 4,0'a ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	53
Çizelge 5.21 Küme 4,2'ye ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	53
Çizelge 5.22 Küme 4,4'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.	53
Çizelge 5.23 Kombine genomik veri (gen ekspresyon, motif sıklık, gen konum) için farklı çalıştırmalar sonucu elde edilen toplam küme sayıları ve tek elemana sahip küme sayıları.	54

Çizelge 5.24 Gen ekspresyon, motif sıklık ve gen konum verisinin bir arada kullanımı ile elde edilen kümeleme sonuçları ve kayda değer gen benzerlikleri.....	56
Çizelge 5.25 Küme 0,0'a ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	60
Çizelge 5.26 Küme 0,1'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	62
Çizelge 5.27 Küme 0,2'ye ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	63
Çizelge 5.28 Küme 0,3'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	64
Çizelge 5.29 Küme 0,4'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	65
Çizelge 5.30 Küme 1,1'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	66
Çizelge 5.31 Küme 1,2'ye ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	67
Çizelge 5.32 Küme 1,3'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	68
Çizelge 5.33 Küme 1,4'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	69
Çizelge 5.34 Küme 2,2'ye ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	69
Çizelge 5.35 Küme 2,4'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda	

üyesi oldukları tespit edilen eski küme bilgileri.....	70
Çizelge 5.36 Küme 3,0'a ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	70
Çizelge 5.37 Küme 3,1'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	71
Çizelge 5.38 Küme 3,3'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	71
Çizelge 5.39 Küme 3,4'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	72
Çizelge 5.40 Küme 4,2'ye ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	73
Çizelge 5.41 Küme 4,3'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	73
Çizelge 5.42 Küme 4,4'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.....	75

ÖNSÖZ

Biyoinformatik, sahip olduğumuz belki de en karmaşık ve en önemli veri tipi olan genetik temelli verilerin anlaşılabilmesi için pek çok farklı disiplinin katkılarıyla gelişimini sürdüren çapraz bir disiplin olarak önemini günden güne arttırmaktadır. Bu disipline ait önemli kilometre taşlarının pek çoğu, ancak son 50 yıl içinde elde edilen bilimsel gelişmelerin ışığında katedilebilmiştir. Kamuoyunda büyük ilgi uyandıran, uluslararası destek ile sürdürülen ve hedefi insan türüne ait tüm genetik bilginin derlenmesi olarak özetlenebilecek “İnsan Genom Projesi” gibi önemli bir proje bile bilindiği üzere ancak 2003 yılında tamamlanabilmiştir. İnsan Genom Projesi ve benzeri projeler neticesinde elde edilen veritabanları hala güncellenerek gelişimlerine devam etmektedir. Bugün geldiğimiz noktada elimizde analiz edilip anlamlandırılmayı bekleyen büyük ve son derece önemli bir veri kaynağı vardır. Söz konusu verinin tam olarak anlaşılması, canlı yaşam süreçlerinin kavranmasını, böyle bir kavrayış ise belki de bu süreçlere gelecekte müdahale edebilmemizi sağlayacaktır. Bu gerçeğin farkında olan ve bilim ile teknolojiye gereken önemi veren ülkelerin çoğunda biyoinformatik çalışmaları gerek üniversite gerekse ulusal araştırma kurumları ölçeğinde yıllar öncesinde başlatılmış ve sürdürülmeye devam edilmektedir.

Biyoinformatiğin en önemli araştırma alanlarından birisi olan gen analizinin en kritik adımlarından birisi olan gen kümelemesi üzerinde yürüttüğüm bu tez çalışmasının, konu ile ilgilenen tüm araştırmacılar için teşvik edici ve yol gösterici bir çalışma olacağını umuyorum. Ülkemizde bu alanda yürütülen çalışmaların günden güne artmasını ve bu alanda tüketen değil üreten bir ülke konumuna gelmemizi diliyorum.

Çalışmam süresince, verimadenciliği alanında bana öğrettikleri ile önümde yeni ufuklar açan ve değerli fikirleriyle bana yol göstereci olan tez danışmanım Yrd. Doç. Dr. Songül Albayrak’a, biyoinformatik alanında öğrendiklerimin çoğunu borçlu olduğum değerli hocalarım Yrd. Doç. Dr. Jane Synnergren ve Doç. Dr. Angelica Lindlöf’e, bana olan inançlarını ve desteklerini her zaman hissettiğim arkadaşlarıma ve hayatım boyunca yanımda olan, desteğini bir an olsun esirgemeyen sevgili aileme en içten teşekkürlerimi ve saygılarımı sunarım. İyi ki varsınız...

ÖZET

GEN KÜMELEME İŞLEMİNİN ÖZDÜZENLEYİCİ HARİTALAR KULLANILARAK GEN EKSPRESYONU, MOTİF SIKLIK VE GEN KONUM VERİLERİNDEN FAYDALANILARAK GERÇEKLEŞTİRİMİ

Alper ÖZDOĞAN

Bilgisayar Mühendisliği, Yüksek Lisans Tezi

Gen analizi biyoinformatiğin en önemli çalışma alanlarından birisi olup gen kümeleme işlemi, gen analizinin en kritik adımlarından birisidir. Gen kümeleme işlemi için literatürde en sık kullanılan veri tipi *gen ekspresyonu* verisidir. Bununla beraber literatürde çok daha az sayıda da olsa, gen kümelemesi için birbirini tamamlayıcı nitelikteki farklı genomik veri tiplerinin birarada kullanılmasını öneren, bu yol ile tek bir genomik veri tipi kullanılması durumunda elde edilemeyecek yeni ve anlamlı gen kümeleri elde edilebileceğini savunan çalışmalar da mevcuttur. Söz konusu çalışmalardan birisi Kasturi ve diğerleri (2005) tarafından sunulan ve *gen ekspresyonu* ile *motif sıklık* verilerinin gen kümelemesi için birarada kullanılmasını öneren çalışmadır.

Bu tez çalışması kapsamında Kasturi ve diğerleri (2005) tarafından sunulan ve temelleri özdüzenleyici haritalara dayanan gen kümeleme algoritması, üçüncü bir genomik veri tipini de gen kümelemesi işleminde kullanabilecek şekilde genişletilmiştir. Bu amacı gerçekleştirmek için seçilen üçüncü genomik veri tipi *gen konum verisidir*. Söz konusu algoritma her veri tipi için farklı ve en uygun uzaklık fonksiyonunun kullanımına izin vermektedir. Bu çalışma kapsamında üçüncü veri tipi olarak ele alınan ve algoritmaya eklenen gen konum verisi ile birlikte kullanılmak üzere GLDist adı verilen yeni bir uzaklık fonksiyonu geliştirilmiş ve kullanılmıştır.

Geliştirilen yöntemin başarımını sınamak için 100 adet Arabidopsis geninden oluşan bir test veri seti oluşturulmuştur. Yöntem ile elde edilen yeni gen kümelerinin biyolojik anlamlılığını değerlendirebilmek için Gen Ontolojisi ve TAIR veritabanlarından faydalanılarak söz konusu genlere ait *moleküler fonksiyon* ve *biyolojik proses* bilgileri elde edilmiş ve bu bilgiler ışığında aynı kümede yer alan genler arasındaki biyolojik benzerlikler incelenmiştir. Elde edilen sonuçların biyolojik değerlendirmesi, geliştirilen yöntemin kullanılması ile yeni ve biyolojik olarak anlamlı gen kümelerinin elde edilebileceğini göstermiştir.

Anahtar kelimeler: Biyoinformatik, Veri Madenciliği, Kümeleme, Gen Kümeleme, Özdüzenleyici Haritalar.

ABSTRACT

CLUSTERING GENES BY USING SELF-ORGANIZING MAPS TOGETHER WITH GENE EXPRESSION, MOTIF FREQUENCY AND GENE LOCATION DATA

Alper ÖZDOĞAN

Computer Engineering, Master Thesis

Gene analysis is one of the most important tasks in bioinformatics and gene clustering is a crucial step that takes part in it. *Gene expression* data is the most commonly-used data type for gene clustering. However, there are some researches in literature in which it is proposed to use different types of genomic data for gene clustering in order to detect new and meaningful gene clusters that may not have been obtained by using only a single type of genomic data. The method proposed by Kasturi et al. (2005) is one of such researches in which *gene expression* and *motif frequency* data are used together to cluster genes.

In this work, the method proposed by Kasturi et al. (2005) is extended by using *gene location* data as a third type of genomic data together with gene expression and motif frequency data. The method proposed by Kasturi et al. (2005) is based on self-organizing maps. Since the algorithm let us use the most appropriate distance functions for each type of genomic data, a new distance function named GLDist is also created to use with gene location data.

In total 100 Arabidopsis genes are chosen as an input dataset to evaluate the results of the proposed method. In order to evaluate the biological significance of the newly obtained clusters by using the proposed method with combined genomic data, Gene Ontology and TAIR databases are used to extract the *molecular function* and *biological process* information of each gene. The biological evaluation of the resulting clusters shows that, the proposed method can be used to obtain new and biologically relevant clusters.

Key words: Bioinformatics, Data Mining, Clustering, Gene Clustering, Self-Organizing Maps.

1. GİRİŞ

Moleküler biyoloji alanındaki bilimsel ve teknolojik gelişmeler sahip olduğumuz genetik bilginin son yıllarda büyük oranda artmasını sağlamıştır. Yeni tekniklerin geliştirilmesi, mevcut teknolojinin ucuzlaması ve özellikle uluslararası genetik araştırma gruplarının çalışmalarıyla beraber söz konusu bilginin büyük bir bölümü bugün herkesin ücretsiz olarak erişebileceği merkezi veritabanlarında ve hergün güncellenerek depolanmaya devam edilmektedir.

Mevcut verinin niteliği göz önünde bulundurulduğunda bu verinin anlaşılabilmesi için yürütülen çalışmaların büyük önem taşıdığı gerçeği artık genel bir kabul haline gelmiştir. Söz konusu verinin boyutları ve karmaşıklığı göz önünde bulundurulduğunda ise analiz için yenilikçi yöntem ve yaklaşımlara ihtiyaç duyulduğu gerçeği görülmektedir. Biyoinformatik, genetik bilgiyi analiz etmek ve anlamak için bilgisayar bilimleri, bilgi teknolojileri ve genetik bilim dallarından faydalanan disiplinler arası bir alandır. Daha geniş bir tanımlama ile ise uygulamalı matematik, enformatik, istatistik, bilgisayar bilimleri, yapay zeka, kimya ve biyokimya gibi disiplinlerden yararlanan ve söz konusu disiplinlere ait tekniklerinin moleküler seviyedeki biyolojik problemlerin çözümünde kullanıldığı disiplinler arası bir alan olarak tanımlanmaktadır [1].

Gen analizi, biyoinformatiğin en önemli çalışma alanlarından birisidir. Genler, tüm canlıların gerçekleşen tüm metabolik olaylarını düzenleyen, dolayısı ile canlı yaşamına yön veren kalıtsal unsurlarıdır. *Gen kümeleme (clustering) işlemi* ise, gen analizinin en kritik adımlarından birisidir. Genel tanımla ile kümeleme işlemindeki amaç, heterojen bir veri kümesinin, kendisini oluşturan elemanların birbirine olan benzerlikleri baz alınarak daha homojen yapıdaki alt gruplara (kümelere) ayrılmasıdır (Berry ve Linoff, 2004). Gen kümeleme işlemi için literatürde en sık kullanılan veri tipi “gen ekspresyonu verisi” adı verilen veri tipidir. İlerleyen bölümlerde detaylı olarak tanımlanacak olan gen ekspresyonu verisi basitleştirilmiş tanımla ile bir genin hangi durum, koşul ve zamanlarda, hangi etkinlikte olduğunu temsil eden sayısal bir veri çeşididir.

Gen analizi kapsamında genlerin tek tek ele alınarak incelenmesi ile oluşacak iş yükü, bu genlerden “benzer ekspresyon profillerine sahip olanların” öncelikle aynı gruba kümelenebilmesi ve bir arada ele alınarak analiz edilmesi ile büyük ölçüde azaltılabilmektedir (He ve diğerleri, 2007). Bu yaklaşım, özellikle karşılaştırmalı genomik çalışmaları için de gereklidir. Uygun

gen kümeleri elde edildikten sonra her küme içerisinde farklı örüntüler (gen benzerlikleri, gen ilişkileri, vb.) bulmak mümkün olabilmektedir. Örneğin, uygun gen kümeleri elde edildikten sonra, fonksiyonel benzerlikler veya gen etkileşimlerine dair ipuçları araştırılarak analiz detaylandırılabilir. Genomik verinin daha etkin bir şekilde kümelenebilmesi amacı ile pek çok farklı kümeleme algoritması ve metodu geliştirilmiştir (Eisen ve diğerleri, 1998; Tamayo ve diğerleri, 1999; D'haeseleer ve diğerleri, 1999; Friedman ve diğerleri, 2000; Holmes ve Bruno, 2000; Jact ve diğerleri, 2001; He ve diğerleri, 2003; Kasturi ve diğerleri, 2005). Bu amaç ile yapılan araştırmalardan bazılarında ise gen kümeleme başarımını arttırmak için genomik bilgi füzyonu (information fusion) önerilmiştir (Holmes ve Bruno, 2000; Kasturi ve diğerleri, 2005).

Söz konusu araştırmalardan Kasturi ve diğerleri (2005) tarafından yayınlananında, temelleri özdüzenleyici haritalara, yada daha yaygın bilinen adı ile SOM (Self Organizing Map) yapay sinir ağı modeline (Kohonen, 1995) dayanan ve bu modelin geliştirilmesi ile elde edilmiş yeni bir gen kümeleme algoritması önerilmiştir. Önerilen bu yeni kümeleme algoritması, elde edilecek gen kümelerinin biyolojik anlamlılığını arttırmak için SOM öğrenme algoritmasının (Kohonen öğrenme kuralının) farklı genomik veri tiplerini bir arada kullanabilecek şekilde geliştirilmesi ile elde edilmiştir. Bu tez çalışmasının amacı, tek bir genomik veri tipinin gen kümelemesi için kullanılması ile elde edilemeyecek fakat biyolojik olarak anlamlı yeni gen kümeleri bulabilmek amacı ile, Kasturi ve diğerleri (2005) tarafından sunulan algoritmayı genişletmek ve uygulamaya dökmektir.

1.1 Motivasyon

Günümüzde, DNA Microarray adı verilen gen ekspresyon seviyesi ölçüm teknolojisinin ulaştığı seviye sayesinde binlerce genin *gen ekspresyon verisi* aynı anda ve detaylı olarak elde edilebilmektedir. Microarray deneyleri sonucunda büyük boyutlarda ve zengin, ancak saklı içeriğe sahip veriler elde edilmektedir. Elde edilen verilerin detaylı olarak analiz edilerek yorumlanması ve bu yol ile biyolojik olarak anlamlı olan bilgilere ulaşılması gerekmektedir. Gen ekspresyon verisi kullanılarak yapılan kümeleme işlemi bu amaca hizmet etmekte ve genler hakkında bilgi edinmemizi sağlamaktadır.

DNA Microarray teknolojisi kullanılarak elde edilmiş verilerden faydalanarak binlerce genin, gen ekspresyonlarındaki değişim eş zamanlı olarak incelenebilmektedir. Ancak bu veri her ne kadar zengin bir içeriğe sahip olursa olsun gen regülasyon mekanizmalarını, yani genlerin

genomik çaptaki ilişkiler ve fonksiyonlar ağını ortaya çıkarabilmek için yeterli bir kaynak değildir (Kasturi ve diğerleri, 2005). Gen regülasyonu, herhangi bir anda bir genin aktif hale gelip hücre içerisinde görev alıp almamasının kontrol edilmesi anlamına gelmekte olup, pek çok farklı dinamiğin ve çeşitli gen etkileşimlerinin sonucunda şekillenmektedir.

Gen ekspresyon verisi, gen kümeleme işlemi için tercih edilen temel ve en yaygın kullanılan genomik veri tipidir. Bununla birlikte bazı araştırmalarda (Holmes ve Bruno, 2000; Barash ve diğerleri, 2002) biyolojik olarak daha anlamlı gen kümeleri elde edebilmek için, kümeleme işleminde gen ekspresyon verileri ile birlikte genlere ait farklı verilerin de, örneğin gen üst-bölgelerinde (upstream regions) yer alan nükleotid dizilim bilgilerinin de, kullanılması önerilmektedir. Bu öneri yeni ve anlamlı gen kümeleri elde edebilmek için, biyolojik olarak birbirini tamamlayıcı nitelikteki kullanılabilir ve anlamlı tüm biyolojik verilerin gen kümeleme işleminde birlikte kullanılabileceği şeklinde de genellenebilir. Bu yol ile sadece tek bir genomik veri tipi kullanılması durumunda tespit edilemeyebilecek, ancak biyolojik olarak ilişkili genlerden oluşan (yani, aynı fonksiyona sahip veya aynı biyolojik prosese katılan genlerden oluşan) daha fazla gen kümesi keşfedilebilir.

Bu tez çalışmasında yukarıda belirtilen gerekçe ve öneri baz alınarak, biyolojik olarak anlamlı genlerden oluşan yeni gen kümeleri bulabilmek adına, “gen ekspresyon verisi”, “motif sıklık verisi” ve “gen konum” verilerinin bir arada kullanımı ile gen kümeleme işleminin gerçekleştirilmesi hedeflenmektedir.

1.2 Problemin Tanımı

Biyolojik sistemlerin sorgulanmasında genler arası etkileşimlerin ve bunların ardında yatan, yani söz konusu etkileşimlere sebep olan biyolojik süreçlerin analiz edilerek anlaşılması son derece önem teşkil etmektedir. Bu etkileşimlerin daha iyi anlaşılması sistem biyolojisinde çok önemli ilerlemeler elde etmemize olanak tanıyacaktır. Gen kümeleme işlemi genellikle, bilinmeyen bir genin fonksiyonunu anlamak, genler arası etkileşimleri analiz etmek veya gen regülasyon mekanizmalarını incelemek gibi son derece önemli çeşitli amaçlar için kullanılmaktadır.

Gen ekspresyon verisi belirtildiği üzere genleri biyolojik olarak anlamlı gruplara ayırmak (kümelemek) için yaygın olarak kullanılmaktadır. Yine belirtildiği üzere anlamlı gen grubu ile kastedilen, grubun aynı moleküler fonksiyona sahip veya aynı biyolojik prosese hizmet eden genlerden oluşmasıdır. Bu yönelimin dayandığı temel hipotez aşağıdaki gibidir:

“Aynı genel fonksiyona sahip genler, farklı koşul ve durumlar altında (örneğin sıcaklık etkisi altında, soğukluk etkisi altında, hücre yaşam döngüsünün çeşitli evrelerinde, vb.) benzer ekspresyon profillerine sahiptirler. Ekspresyon profillerindeki bu benzerliğin temel nedeni, aynı fonksiyonel grupta yer alan genlerin aynı transkripsiyon faktörü tarafından regüle (ko-regülasyon) edilmeleridir (Barash ve diğerleri, 2002).”

Ancak bu genel kabul, bazı gen fonksiyonları için her zaman doğru değildir. Bu sebeple benzer ekspresyon profillerine sahip genlerin yine de farklı fonksiyonlara sahip olabilecekleri gerçeği her zaman göz önünde bulundurulmalıdır. Diğer bir deyişle, birlikte regüle edilmediği (co-regulated) halde, benzer ekspresyon profillerine sahip genler de bulunmaktadır (Barash ve diğerleri, 2002). Bu yüzden sadece gen ekspresyon verisi kullanılarak elde edilen gen kümelerinin biyolojik olarak doğruluğu, fonksiyonel tutarlılıkları açısından ele alındığında her zaman yeterli olmayabilir. Bu ve benzeri nedenlerden ötürü, gen regülasyon mekanizmalarını genomik boyutta ortaya çıkarabilmemiz için gen ekspresyon verisinin yeterli bir kaynak olmadığı söylenebilir (Kasturi ve diğerleri, 2005).

Bu tez çalışması kapsamında geliştirilen kümeleme yöntemi ile, birbirini tamamlayıcı nitelikte farklı genomik veri tipleri bir arada kullanılarak gen ekspresyon verisi temelli kümelemenin sınırlamalarını aşmak hedeflenmektedir. Bu hedefi gerçekleştirmek için önerilen kümeleme algoritması üç farklı genomik veri tipini kullanmaktadır ve Kasturi ve diğerleri (2005) tarafından geliştirilen kümeleme yönteminin genişletilmesi ile elde edilmiştir. Kasturi ve diğerleri (2005)’nin çalışmasında “gen ekspresyon verisi” ile “motif sıklık verisi” bir araya getirilirken, bu tez çalışmasında kümeleme etkinliğini arttırmak için “gen ekspresyon verisi”, “motif sıklık verisi” ve “gen konum verisi” bir arada kullanılmaktadır. Kasturi ve diğerleri (2005) tarafından önerilen algoritmayı geliştirmek amacı ile belirtildiği üzere bir üçüncü genomik veri tipi kullanmanın yanında, yeni bir uzaklık fonksiyonu da bu çalışma kapsamında geliştirilmiş ve kümeleme işleminde kullanılmıştır.

1.3 Hipotez

Ek genomik veriler (örneğin bir gene ait üst-bölge nükleotid dizileri, gen ontolojileri, gen konum bilgisi, vb.) gen ekspresyon verisi ile bir arada kullanıldığında gen kümeleme başarımını arttırabilir (Kasturi ve diğerleri, 2005). Farklı genomik verilerin bir arada kullanılması ile elde edilecek gen kümeleri, tek bir genomik veri tipi kullanılarak elde edilemeyecek, ancak biyolojik olarak anlamlı yeni gen gruplarını keşfetmemizi sağlayabilir.

Bu çalışma kapsamında Kasturi ve diğerleri (2005) tarafından önerilen gen kümeleme algoritmasının genişletilmiş/geliştirilmiş hali kullanılmaktadır. Kasturi ve diğerleri (2005) tarafından sunulan söz konusu algoritma, bu amaçla üçüncü bir genomik veri tipi ve farklı kümeleme metrikleri (uzaklık fonksiyonları) ile çalışabilecek hale getirilmiştir.

1.4 Hedefler ve Amaçlar

Bu tez çalışmasının amaçları ve hedeflenenler aşağıdaki gibidir:

- *Kasturi ve diğerleri (2005) tarafından sunulan algoritmayı genişletmek hedeflenmektedir.*
 - *Mevcut algoritmayı üçüncü bir veri tipini de kullanacak şekilde genişletme/geliştirme hedefi:* Kasturi ve diğerleri (2005) tarafından geliştirilen algoritma, gen kümeleme işlemi için iki farklı genomik veri tipini kullanmaktadır. Bunlar gen ekspresyon verisi ve motif sıklık verisidir. Bu tez çalışması kapsamında mevcut algoritmayı gen lokasyon verisini üçüncü veri tipi olarak kullanacak şekilde geliştirmek hedeflenmektedir.
 - *Mevcut algoritmayı farklı kümeleme metrikleri (uzaklık fonksiyonları) kullanarak genişletme/geliştirme hedefi:* Kümeleme işleminin ilk adımı benzerlik derecesinin ölçülebilmesi için matematiksel bir tanım oluşturmaktır. Bu nedenle kümeleme işleminin performansı seçilen uzaklık fonksiyonuna büyük oranda bağlıdır. Bu bağlamda, Kasturi ve diğerlerinin (2005) çalışmalarında kullandıkları uzaklık fonksiyonlarının (algoritmada her bir veri tipi için farklı bir uzaklık fonksiyonu kullanılmaktadır) yerine daha uygun uzaklık fonksiyonları uygulanıp uygulanamayacağı, bu yol ile mevcut yöntemin geliştirilip/geliştirilemeyeceği sınanacaktır.
- *Algoritmanın yukarıda belirtildiği biçimde genişletilecek/geliştirilecek halini gerçekleştirecek bir yazılım oluşturmak hedeflenmektedir.*
- *Algoritmanın genişletilmiş/geliştirilmiş şeklinin başarımını test edebilmek için uygun bir veri kümesi, yani test veri kümesi oluşturmak hedeflenmektedir.*
- *Geliştirilen yöntemin başarımını elde edilecek gen kümelerinin biyolojik anlamlılığını sınavarak değerlendirmek hedeflenmektedir.* Algoritmanın amacı daha önce de belirtildiği gibi genomik veri füzyonu ile yeni ve biyolojik olarak anlamlı gen kümeleri elde etmektir. Elde edilecek gen kümelerinin biyolojik olarak anlamlılığı Gen Ontolojisi [2] anotasyonlarından faydalanılarak değerlendirilecek, böylelikle yöntemin başarısı sınanacaktır.

1.5 Tezin Yapısı

Bu tez çalışmasının ilk bölümünde tez konusu, konunun özündeki problem, bu konunun seçilmesinin ardında yatan motivasyon nedenleri, çalışmanın amaçları ve hedefleri açıklanmaktadır. İkinci bölümde konunun anlaşılması için gerekli olan temel kavramlar açıklanmaktadır. Gen ekspresyon verisi, motif sıklık verisi gibi bu çalışma kapsamında gen kümeleme işlemi için kullanılan tüm genomik veri tiplerine ve gerekli diğer temel genetik konularına bu bölümde değinilmektedir. Kullanılan kümeleme tekniğinin temellerinin dayandığı SOM yapay sinir ağı modeli de yine bu bölümde açıklanmaktadır. Üçüncü bölümde gen kümeleme konusu üzerinde yapılmış önceki çalışmalar hakkında genel bilgi verilmektedir. Bu tez çalışması kapsamında kullanılan ve geliştirilen Kasturi ve diğerleri (2005)'ne ait bilgi füzyonuna dayalı gen kümeleme algoritması da detaylı olarak Bölüm 3'te açıklanmaktadır. Bölüm 4'te çalışma kapsamında kullanılan materyaller ve metodlar açıklanmaktadır. Geliştirilen algoritmanın test edilmesi için kullanılan veri kümesi detaylı özellikleri ile birlikte bu bölümde açıklanmaktadır. Mevcut algoritmanın ne şekilde geliştirildiği/genişletildiği ve uygulama yazılımının ne şekilde oluşturulduğu yine bölüm dördte anlatılmaktadır. Kümeleme işlemi sonuçları ve sonuçların analizi Bölüm 5'te yer almaktadır. Tez çalışmasının sonuçları, değerlendirmeler ve öneriler ile birlikte Bölüm 6'da yer almaktadır.

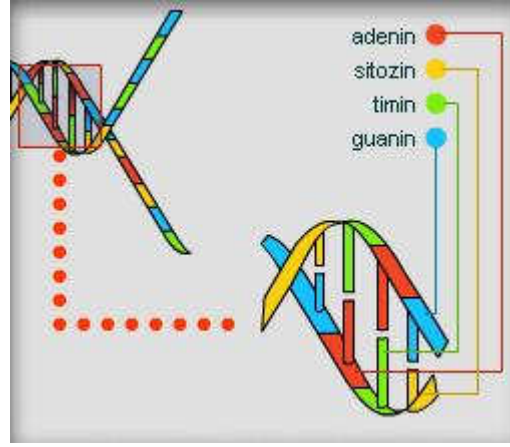
2. TEMEL KAVRAMLAR

Çalışma kapsamında ele alınan konunun daha iyi anlaşılabilmesi için gerek duyulan temel kavramlar bu bölümde sunulmaktadır. Bu amaç ile ilk olarak gerek duyulan temel genetik kavramları açıklanmaktadır. Çalışma kapsamında kullanılan üç farklı genomik veri tipinin, yani gen ekspresyon verisinin, motif sıklık verisinin ve gen konum verisinin nitelikleri bu bölümde açıklanmaktadır. Veri madenciliğinin en bilinen uygulama alanlarından biri olan ve bu tez çalışması kapsamında genler üzerinde uygulanan kümeleme işlemine yine bu bölümde değinilmektedir. Teze konu olan ve bu bağlamda genişletilen/geliştirilen Kasturi ve diğerleri (2005)'ne ait gen kümeleme algoritmasının temelini teşkil eden SOM (Kohonen, 1995) yapay sinir ağı modeli de yine bu bölümde ayrıntıları ile açıklanmaktadır.

2.1 Temel Genetik Kavramları

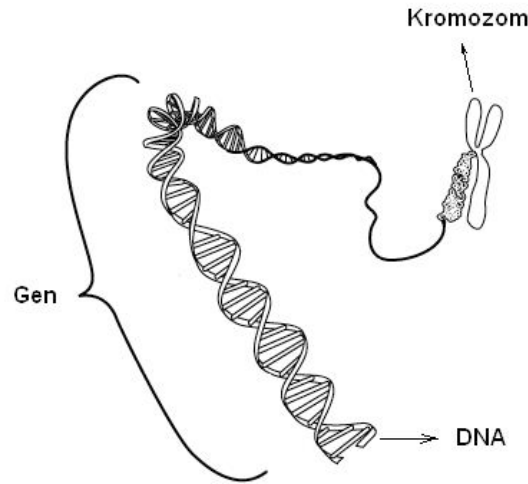
Biyoloji, temelleri çok eskilere dayanan bir bilim dalı olmasına rağmen, günümüzde biyoloji alanında yürütülen pek çok ilgi çekici çalışmanın temelleri son zamanlarda bulunan yeni teknik ve fikirlere dayanmaktadır. Modern biyoloji çalışmalarında son derece önemli yere sahip genetik biliminin temelleri günümüzden sadece 100 yıl kadar önce, Mendel'in çalışmalarıyla atılmıştır (Tisdall, 2001). Modern genetik çalışmaları için son derece önemli bir kilometre taşı olan DNA'nın (deoksiribonükleik asit) yapısının açıklanması ise günümüzden sadece 50 yıl kadar önceye dayanmaktadır (Tisdall, 2001).

DNA, tüm hücreli canlılarda (ve bazı virüslerde) yer alan, canlının gelişimi için gerekli olan biyolojik bilgiyi taşıyan ve bu bilginin nesilden nesile aktarılmasında görev alan, bu nedenle de kalıtım molekülü olarak da adlandırılan nükleik asittir. DNA, baz veya nükleotid olarak adlandırılan 4 molekülün birleşiminden oluşmuştur ve çift sarmallı yapıdadır. (Tisdall, 2001). DNA'yı oluşturan nükleotidler Adenin (A), Sitozin (C), Guanin (G), ve Timin (T) olarak adlandırılmaktadır (bkz. Şekil 2.1). DNA'yı oluşturan nükleotidlerin her biri, sarmalın diğer iplikindeki tek bir nükleotid ile bağ oluşturabilmektedir. Adenin (A) ve Timin (T) sadece birbirleri arasında bağ kurabilirken, Sitozin (C) ve Guanin (G) de sadece birbirlerine bağlanabilmektedir. Dolayısıyla sarmalın bir iplikçiğinde A yer alıyor ise, diğer iplikçiğinde T yer almakta; bir iplikçikte C yer alıyor ise, diğer iplikçikte G yer almaktadır. Birbirlerine bağlanan bu nükleotidlere, nükleotid çifti (baz çifti) adı verilmektedir (Tisdall, 2001).



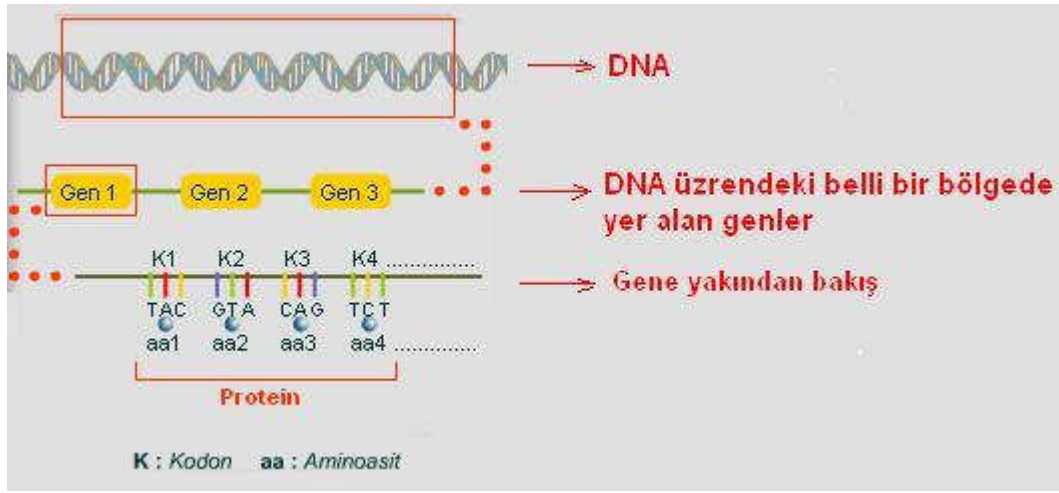
Şekil 2.1 DNA'nın çift sarmallı yapısı ve DNA'yı oluşturan nükleotidler [3].

Gelişmiş organizmalarda DNA, kromozom adı verilen yapılar üzerinde yer almaktadır (Lesk, 2002) (bkz. Şekil 2.2). DNA'nın üzerinde ise gen adı verilen özel bölgeler bulunmaktadır. Kromozom sayısı türler arasında farklılık gösterebilmektedir. Örneğin, insan hücrelerinde 23 kromozom çifti yani toplamda 46 kromozom bulunmaktadır. Farklı canlı türleri arasında eşit sayıda kromozoma sahip olan türler de bulunmaktadır. Kromozom sayılarındaki bu eşitlik, söz konusu canlı türlerinin birbirine benzer olduğunu göstermemektedir. Çünkü türler arası benzerlik kromozom sayısına değil, nükleotid dizilişine bağlıdır. Benzer şekilde canlıların gelişmişlik dereceleri de kromozom sayılarına değil, nükleotid dizilişlerine bağlıdır. Peki nükleotid dizilişleri bunu nasıl belirlemektedir?



Şekil 2.2 Kromozom, DNA ve genler [4].

Nükleotidler, üç tanesi bir araya gelmek suretiyle kodon adı verilen birimleri oluştururlar. Her bir kodon ise aminoasit adı verilen birimlerin kodunu içerir. Aminoasitler çeşitli sayı ve sıra kombinasyonlarında bir araya gelerek, protein adı verilen üst düzey organik birimleri oluştururlar (bkz. Şekil 2.3). Bir canlı organizmasında gerçekleşen tüm olayların kontrolü büyük ölçüde, tüm doku ve organların yapı taşları olan proteinler tarafından düzenlenmektedir. Dolayısı ile canlıların gelişmişliği DNA'ları üzerinde saklı bulunan genetik bilgi sayesinde üretebildikleri gen ürünlerinin (protein ve RNA) zenginliğine ve gelişmişliğine bağlıdır. Doğada 20 adet aminoasit bulunmaktadır ve örneğin insan bedeninde söz konusu 20 aminoasit ile 200.000 kadar farklı protein oluşturulabilmektedir [3].

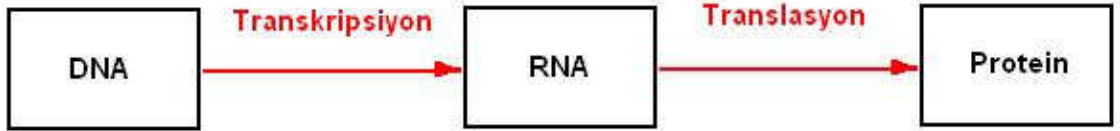


Şekil 2.3 DNA, gen, kodon ve aminoasitler ile bir proteine ait genetik kodun hiyerarşik biçimde gösterimi [3].

Daha önce de belirtildiği gibi DNA üzerinde gen adı verilen özel bölgeler yer almaktadır. Bu tanımdan da anlaşılacağı gibi aslında genler, DNA üzerinde yer alan nükleotidlerin belirli bir grubunu teşkil eden segmentlerdir. Her bir gen bir veya genellikle işlevsel açıdan ilintili birden çok proteine ait genetik kodu içermektedir. Diğer bir deyiş ile canlıların ihtiyaç duydukları proteinler, gen adı verdiğimiz DNA segmentlerinde yer alan genetik bilgi sayesinde üretilirler. Gen üzerinde yer alan genetik veriden hareketle, protein gibi bir gen ürününün oluşturulması sürecine “gen ekspresyonu” adı verilmektedir (Setubal ve Meidanis, 1997). Gen ekspresyonu adı verilen bu süreç iki temel evreden oluşmakta olup, moleküler biyolojinin temel dogması olarak adlandırılmaktadır.

2.1.1 Moleküler Biyolojinin “Temel Dogması”

DNA üzerinde yer alan genetik bilgiden hareketle bir gen ürününün, yani genellikle bir proteinin veya bir RNA'nın oluşması süreci moleküler biyolojinin temel dogması olarak kabul edilen iki aşamalı bir süreçtir. Bu aşamalar transkripsiyon ve translasyon olarak adlandırılmaktadır (bkz. Şekil 2.4).



Şekil 2.4 Moleküler biyolojinin “Temel Dogması” olarak adlandırılan, hücre içi genetik bilgi akışı (Setubal ve Meidanis, 1997).

Sürecin ilk aşaması olan transkripsiyon evresinde, DNA'daki genetik veriden hareketle bir mesajcı RNA (mRNA) oluşturulmaktadır. Bu işlem, RNA polimeraz enzimi (RNAP) adı verilen enzimin devreye girerek DNA sarmalını aralaması ile gerçekleşir. RNAP, sentezlenecek olan gen ürününe ait genetik bilgiyi barındıran genin promotör bölge olarak adlandırılan bölgesine bağlanır. Bağlandığı genin başından sonuna kadar DNA sarmalını aralar. Aralanan sarmalın bir kolundaki nükleotidlere karşı düşen nükleotidler (A için T veya tersi, C için G veya tersi) bir araya gelecek şekilde bir RNA iplikçığı oluşturulur. Yani mRNA üzerine kopyalanan veri, aslında tam olarak DNA sarmalının bir kolundaki verinin aynısıdır. Böylelikle DNA'daki genetik verinin o an hücre tarafından ihtiyaç duyulan ilgili bölümü mRNA üzerine kopyalanmış olur.

Hücrenin protein üretiminden sorumlu olan organeli ribozom adı verilen organelidir. Tüm proteinler, ribozomda üretilmektedir. Transkripsiyon işlemi sonucu oluşan mesajcı RNA'nın görevi, taşıdığı genetik veriyi protein sentezinin tamamlanması için ribozoma taşımaktır. Ribozomların kendilerine ulaşan mesajcı RNA'lar tarafından taşınan genetik veriden hareketle ilgili proteinleri sentezlemesi işlemine translasyon denilmektedir.

Böylelikle, iki aşamadan oluşan bir süreç sonunda DNA üzerinde kodlu bulunan genetik veriden faydalanılarak, hücrenin o an ihtiyaç duyduğu gen ürünlerinin oluşması sağlanmış olur. Bu işlem “gen ekspresyonu” olarak adlandırılmaktadır.

Gen ekspresyonu, üzerinde hassasiyetle durulması gereken bir süreçtir. Çünkü canlıların yaşamlarını sürdürebilmeleri için gerek duyduğu tüm bileşenlerin üretimi ve tüm fonksiyonların kontrolü genler tarafından gerçekleştirilmektedir. Büyüme, hücre bölünmesi, hücre farklılaşması, çeşitli adaptasyonlar, vb. tüm gerekli yaşamsal süreçler gen ekspresyonu sayesinde gerçekleşebilmektedir. Gen ekspresyonu süreç dinamiklerinin ve gen ekspresyonunun nasıl düzenlendiğinin (regüle edildiğinin), diğer bir deyiş ile kontrol edildiğinin anlaşılması bu nedenle son derece önem teşkil etmektedir. Gen ekspresyonunun dolayısıyla gen etkileşimlerinin, gen fonksiyonlarının ve bu fonksiyonların hangi şartlar altında, ne şekilde devreye girdiğinin anlaşılması yaşamsal süreçlerin nasıl şekillendiğini anlamamızı sağlayacaktır. Canlı doğasına ve yaşam süreçlerine dair bu denli kapsamlı bir bilgi belki de bu süreçlere müdahale edebilmemizi olanaklı kılacaktır. Örneğin, herhangi bir hastalıktan sorumlu olan bir genin tespit edilmesi ve soruna yol açan geni doğrudan hedef alacak ilaçların geliştirilmesi ile söz konusu hastalığa temelden çözüm geliştirilmesi mümkün olabilecektir. Örneğin, soğuğa ve sususluğa karşı direnç gösteren bitkilerin bu direci göstermelerini sağlayan genlerinin tespiti ve genetik mühendisliği ile daha dayanıklı tarım ürünlerinin geliştirilerek yaygınlaştırılması belki de küresel anlamda açlık sorununu çözebilecektir. Bugün bilimkurgu senaryosu gibi gözükebilecek bu ve benzeri pek çok örnek gen fonksiyonlarının anlaşılması ve genetik mühendisliği ile belki de yakın gelecekte gerçeğe dönüşebilecektir.

2.1.2 Gen Ekspresyonu Verisi

Belirtildiği üzere gen ekspresyonu işlemi bir gende saklı bulunan kalıtsal bilgiden yararlanılarak bir gen ürününün, örneğin bir protein veya RNA'nın üretildiği bir işlemdir (Hunter, 1993). Gen ekspresyonu verisi en basit tanımı ile, herhangi bir anda bir genin hücre içi aktiflik durumunu ifade eden sayısal bir veridir. Bu sayısal veri, hücre içinde söz konusu anda söz konusu gene ait gen ürünlerinin miktarının tespitine dayanmaktadır. Söz konusu miktar ne kadar fazla ise genin etkinlik durumu da o denli yüksek anlamına gelmektedir. Bugün sahip olunan teknoloji ile pek çok genin ekspresyon seviyelerini eş zamanlı olarak ölçebilmek ve böylelikle karşılaştırmak mümkün hale gelmiştir.

Geleneksel moleküler biyolojide bir genin fonksiyonunu belirleyebilmek için uzun yıllar boyunca kullanılan genel yaklaşım, her geni tekil olarak ele almak ve izolasyon altında incelemektir. Ancak kompleks gen etkileşimlerini ve kompleks biyolojik süreçlerin doğasını doğru olarak anlayabilmek için büyük miktarda gene ait ekspresyon bilgilerinin eş zamanlı

olarak incelenmesi gerekmektedir (Michaels ve diğerleri, 1998). Çünkü yürütülen çalışmalar göstermiştir ki, biyolojik süreçler büyük ölçüde bir değil birden fazla genin komplike etkileşimleri sonucunda gerçekleşebilmektedir. Genlerin fonksiyonel olarak ve birbirini tamamlayıcı ilişkileri üst düzeydedir. Bu ilişkilerin tespit edilebilmesi için aynı koşullar altında elde edilmiş birden çok gene ait gen ekspresyon verilerinin bir bütün halinde incelenmesi ve karşılaştırılması gerekmektedir.

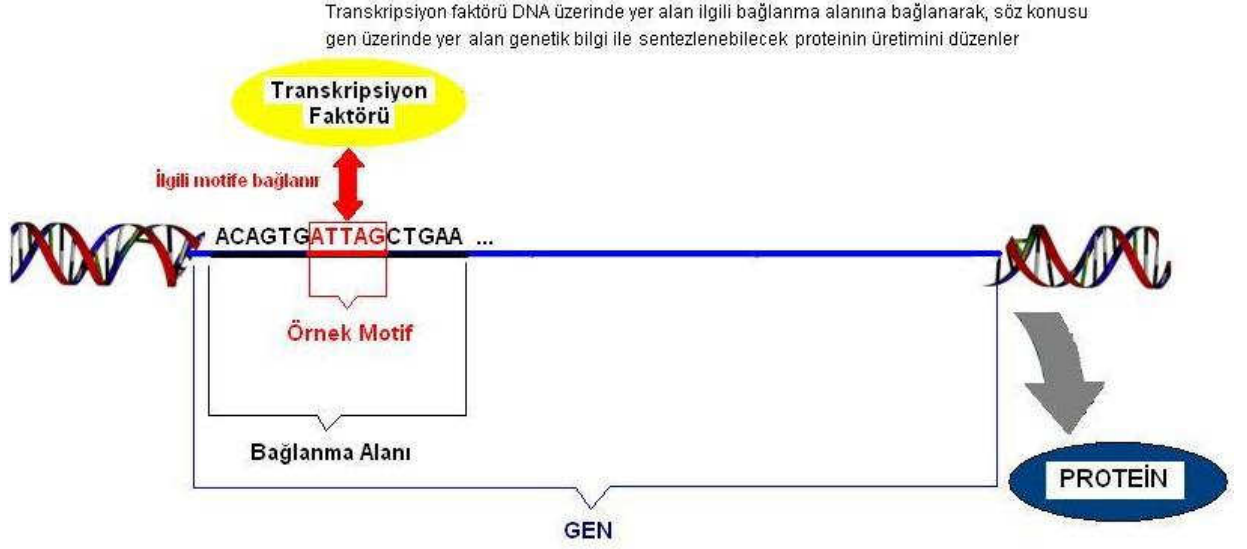
DNA Microarray adı verilen teknolojinin günümüzde ulaştığı seviye sayesinde binlerce gene ait ekspresyon seviyelerinin eş zamanlı olarak tespit edilmesi olanaklı hale gelmiş ve söz konusu teknoloji günümüzde gen ekspresyon analizinin en önemli araçlarından biri haline gelmiştir (Speer ve diğerleri, 2004). Hücre yaşam döngüsünün farklı evrelerinde veya farklı çevresel koşullar altında (örneğin soğukluk etkisi altında, sıcaklık etkisi altında, hastalık durumunda, bölünme evresinde, vb.) aynı ekspresyon seviyelerine sahip genleri tespit edebilmek son derece önemlidir. Çünkü eş-ekspresyona sahip genlerin (co-expressed genes), diğer bir deyiş ile benzer ekspresyon profillerine sahip olan genlerin muhtemelen birbiriyle ilişkili fonksiyonlara sahip olduğuna inanılmaktadır (Mount, 2001). Bu nedenle gen ekspresyonu verisi, gen kümeleme işlemi için sıklıkla kullanılan temel veri tipi konumuna gelmiştir.

2.1.3 Transkripsiyon Faktörleri, Transkripsiyon Faktörü Bağlanma Alanları ve Motifler

Moleküler biyolojinin en zorlu problemlerinden biri gen ekspresyonunun nasıl düzenlendiğinin anlaşılmasıdır (Jact ve diğerleri, 2001). Bir diğer deyiş ile genlerin hangi koşullar altında ve zamanlarda, ne şekilde ve hangi düzenleyici öğelerin etkisi altında devreye girerek hangi amaca hizmet ettiğinin anlaşılmasıdır. Gen ekspresyonunu düzenleyici öğelerin, örneğin transkripsiyon faktörlerinin bağlanma alanlarının tespit edilmesi ise bu problemin çözülebilmesi için gerek duyulan en önemli adımlardan birisidir (Tompas ve diğerleri, 2005). Transkripsiyon faktörleri, isimlerinden de anlaşılabilceği gibi transkripsiyon işlemini düzenleyen en önemli moleküllerdendir.

Transkripsiyon faktörleri, DNA'ya bağlanan ve gen ekspresyonunu kontrol eden özel proteinlerdir. Transkripsiyon faktörleri DNA üzerinde belli alanlara bağlanmaktadır. Transkripsiyon faktörlerinin ve ilişkili oldukları bağlantı alanlarının (binding sites) tespiti, "motif" adı verilen kısa uzunluktaki baz dizilerinin tespit edilmesi ile gerçekleştirilmektedir.

Motif adı verilen söz konusu alanlar, transkripsiyon faktörlerinin bağlantı noktalarında yer almaktadır (bkz. Şekil 2.5). Bu anlamda transkripsiyon faktörleri ve motifler arasındaki ilişkiyi kilit ve anahtar ilişkisine benzetmek mümkündür.



Şekil 2.5 Transkripsiyon faktörleri, bağlanma alanları ve motifler [5].

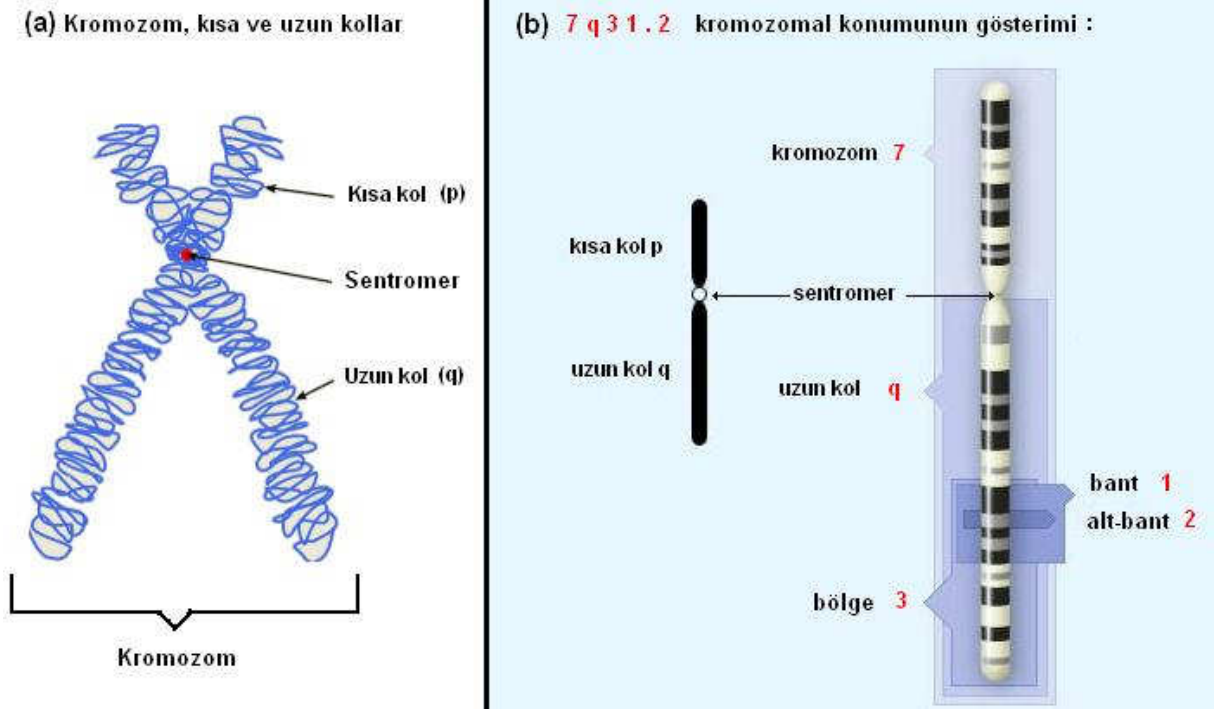
Aynı transkripsiyon faktörü tarafından düzenlenen genlerin benzer fonksiyonlara sahip olduklarına ve/veya aynı biyolojik süreçlerde görev aldıklarına inanılmaktadır. Motif adı verilen bu kısa diziler genler ve transkripsiyon faktörleri arasındaki ilişkiye işaret ettikleri için, aynı motiflere sahip olan genlerin potansiyel benzerlikler taşıyan genler olduklarına inanılmaktadır. Bu nedenle genler ve sahip oldukları motiflere ilişkin veriler (yani bir motifin bir gen üzerinde yer alıp almadığı, alıyor ise kaç kez yer aldığı bilgisi), gen kümeleme işlemi için önemli bir kaynak veri olarak görülmektedir.

Motifleri tespit edebilmek için ko-regüle edildiğine inanılan, yani benzer şekilde düzenlendiğine inanılan bir grup genin, regülasyon işleminden sorumlu olduğuna inanılan bölgelerine (motiflerin bağlandığı bölgelere) ait nükleotid dizilim bilgilerine gereklidir. Elde edilen bu bölgeler, motifleri tekrar etme sıklıklarının istatistiksel kayda değerliğine dayanarak tespit eden araçlar ile incelenmekte ve bu yol ile motifler tespit edilmektedir (Tomba ve diğerleri, 2005).

2.1.4 Genlerin Kromozomal Konumu

Her genin üzerinde yer aldığı kromozomdaki konumu kendisine özgü olup, bu bilgi her gen için farklıdır. Söz konusu konum bilgisine lokus (locus) adı verilmektedir. Bir gene ait

kromozomal lokus bilgisi genel olarak iki farklı notasyon ile ifade edilmektedir. Bu notasyonlardan ilki “5p12.3” örneğinde görülen biçimdir. Bu notasyonda harften önce gelen ilk sayısal değer olan “5” değeri kromozom numarasını ifade etmektedir. “p” değeri ise kromozomun kısa kolunu ifade etmektedir. Kromozomlar bir kısa ve bir de uzun olmak üzere iki kola sahiptir ve bu notasyona göre “q” değeri uzun kolu ifade etmektedir (bkz. Şekil 2.6 (a)). “5p12.3” örneğindeki noktadan önce gelen “12” değeri bant numarasını (bölge-bant) ifade etmektedir. Noktadan sonra gelen “3” değeri ise alt-bant numarasını ifade etmektedir. Bantlar, alt-bantlardan oluşmakta olup ikisi birlikte bir kromozomun kısa veya uzun kolu üzerindeki belli bir bölgeye işaret etmektedir (bkz. Şekil 2.6 (b)). Alt-bantlar daha spesifik bölgeleri belirtmek için, yani daha detaylı konum bilgisi verebilmek için kullanılmaktadır.



Şekil 2.6 (a) Uzun ve kısa kolları ile bir kromozomun yapısı [6]. (b) Örnek bir gen için kromozomal lokasyonunun detaylı gösterimi [7].

Bir gene ait lokus bilgisini belirtmek için kullanılan bir diğer notasyon ise sıra dizilim notasyonudur. Bu notasyon, yukarıda belirtilen notasyonda olduğu gibi hiyerarşik seviyelerden oluşmayıp basitçe “kromozom numarası” ve “geni oluşturan bazların başlangıç indeksi” ile “geni oluşturan bazların bitiş indeksi” bilgilerinden oluşmaktadır. Bu notasyon moleküler lokasyon notasyonu olarak da bilinmektedir.

2.2 Kümeleme Yöntemleri

Veri madenciliğinin en bilinen ve en yaygın şekilde kullanılan yöntemlerinden birisi de *kümeleme* işlemidir. Kümeleme işleminde amaç, heterojen bir veri kümesinin, kendisini oluşturan elemanların birbirlerine olan benzerlikleri göz önünde bulundurularak daha homojen yapıdaki alt kümelere ayrılmasıdır (Berry ve Linoff, 2004). Bu tanımdan da anlaşılabilceği gibi kümeleme işlemi genellikle daha önce var olduğu bilinmeyen ancak veri setlerinde saklı olan benzerlik ilişkilerini ortaya çıkarmak için kullanılan bir yöntemdir.

İdeal bir kümeleme işlemi sonucunda hedeflenen küme içi elemanlarının benzerliği yüksek, kümeler arası benzerliğin ise düşük olduğu kümeler elde etmektir. Kümeleme işleminin yapılabilmesi için öncelikle “benzerlik” kavramının net bir şekilde ortaya konması gerekmektedir. Kümelemeye tabi tutulan elemanlar arası benzerlik seviyeleri genellikle uzaklık fonksiyonu olarak adlandırılan matematiksel fonksiyonlar ile ölçülmektedir. Böylelikle birbirine en yakın, yani benzer elemanlar aynı kümenin elemanları olarak gruplanabilmektedir. Bu tanımdan da anlaşılabilceği gibi kümeleme işleminin başarımı, uzaklık fonksiyonu seçimi ile büyük oranda ilişkilidir.

Farklı veri tipleri farklı karakteristik özelliklere sahip olabilmektedir. Bu nedenle bir uzaklık fonksiyonu bir veri tipi için benzerliği başarılı bir şekilde ifade edebilirken, bir başka veri tipi için ise yetersiz kalabilmektedir. Literatürde çok sayıda farklı uzaklık fonksiyonu/benzerlik ölçütü bulunmaktadır (Öklid uzaklığı, Manhattan uzaklığı, Minkowski, vb.). Kümeleme işleminde dikkat edilmesi gereken en önemli noktalardan birisi, kümelenmek istenilen veri setinin karakteristiğine en uygun uzaklık fonksiyonunun seçilmesi veya gerek duyuluyor ise geliştirilmesidir.

Literatürde çok sayıda kümeleme metodu bulunmaktadır. Kümeleme metodlarını temel olarak hiyerarşik ve hiyerarşik olmayanlar olmak üzere ikiye ayırmak mümkündür (Giudici, 2003). Hiyerarşik metodlar, her bir elemanın başlangıçta bağımsız bir küme olarak kabul edilip kademeli olarak birleştirildiği (agglomerative) veya tüm elemanların başlangıçta tek bir büyük kümenin elemanı olarak kabul edilip daha sonra kademeli olarak alt kümelere ayrıldığı (divisive) metodlardır. Hiyerarşik olmayan metodlar ise (örneğin k-means yöntemi), n adet farklı kümeyi önceden tanımlı olmak kaydı ile doğrudan elde etmemizi sağlayan metodlardır. Her iki tür kümeleme çeşidi de bazı avantajlara ve dezavantajlara sahiptir. Örneğin hiyerarşik metodlarda elde edilmek istenen küme adedini önceden bilmek gerekmemektedir. Çeşitli

hiyerarşik seviyelerde elde edilmiş kümeler incelenerek söz konusu analiz için en uygun olduğu tespit edilen seviyedeki sonuçlar dikkate alınabilir. Bununla beraber bu avantaj beraberinde çok fazla kaynak (zaman, bellek, vb.) kullanımı dezavantajını getirir. Yine örneğin, hiyerarşik olmayan metodlar hiyerarşik metodlara kıyasla daha hızlıdır (Giudici, 2003). Bu da özellikle büyük veri setleri için önemli bir alternatif olmalarını sağlar. Ancak, hiyerarşik olmayan bir kümeleme yöntemi kullanılacak ise elde edilmek istenen nihai küme adedinin önceden bilinmesi gerekmektedir. Bu de genellikle bir ön analiz yapılmasını gerektirir.

Bu çalışma kapsamında kullanılan kümeleme yöntemi olan özdüzenleyici haritalar (SOM) hiyerarşik olmayan kümeleme yöntemlerinden biri olup, bu yönteme ait detaylı bilgi Bölüm 2.2.1’de yer almaktadır.

2.2.1 SOM Yapay Sinir Ağları: Özdüzenleyici Haritalar

Özdüzenleyici haritalar, yani SOM (Self-Organizing Maps), Fin profesör Tuevo Kohonen tarafından bulunan rekabet-temelli (competition) bir çeşit yapay sinir ağı modelidir (Kohonen, 1995). Özdüzenleyici haritalar bazı kaynaklarda Kohonen Ağları olarak da anılmaktadır. Özdüzenleyici haritalar kümeleme işlemi ve gen ekspresyon örüntülerinin analizi için uygun bir tekniktir (Tamayo ve diğerleri, 1999). Özdüzenleyici haritalara ait detaylı bilgi bu bölümde yer almaktadır.

2.2.1.1 SOM Mimarisi

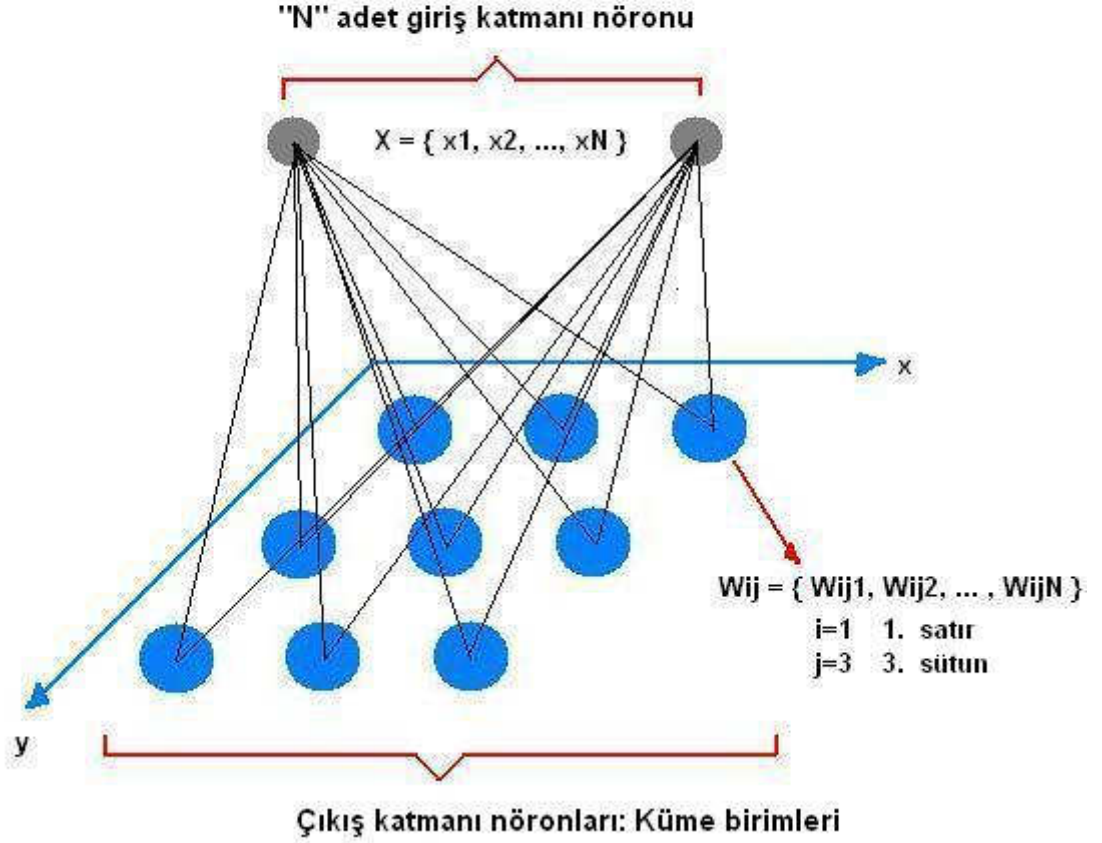
Özdüzenleyici haritalar, eğitici olmaksızın (unsupervised) veri sınıflamayı öğrenebilen bir yapay sinir ağı modelidir. Bu nedenle gen kümeleme işlemi için son derece uygun bir alternatiftir. Çünkü gen kümeleme işlemi yapılırken, kullanılabilecek eğitim veri setleri (genler ve ait oldukları gen grupları bilgilerini içeren bir eğitim seti) bulunmamaktadır.

SOM sinir ağı iki temel bileşenden oluşmaktadır. Bunlar girdi birimleri ve küme merkezleridir (bkz. Şekil 2.7). SOM ağlarının girdisi, girdi verisinin öznitelikleridir (features) ve girdi birimlerinin adedi söz konusu özniteliklerin adedine eşittir. Rekabet eden nöronlar, kümeleme birimleridir. Kümeleme birimlerinin sayısı, diğer bir deyiş ile nöronların sayısı probleme ve kullanıcının seçimine göre değişmektedir. SOM ağlarında kümeleme birimleri arasında topolojik bir komşuluk ilişkisi olduğu varsayılmaktadır.

"N" adet özniteliğe sahip örnek veri

$X_i = \{ x_{i1}, x_{i2}, \dots, x_{iN} \}$, Örneğin: $X_1 = \{ 0.9, 0.3, \dots, 0.7 \}$

"N" adet özniteliğe sahip girdi verisi için tasarlanmış 3 x 3 SOM



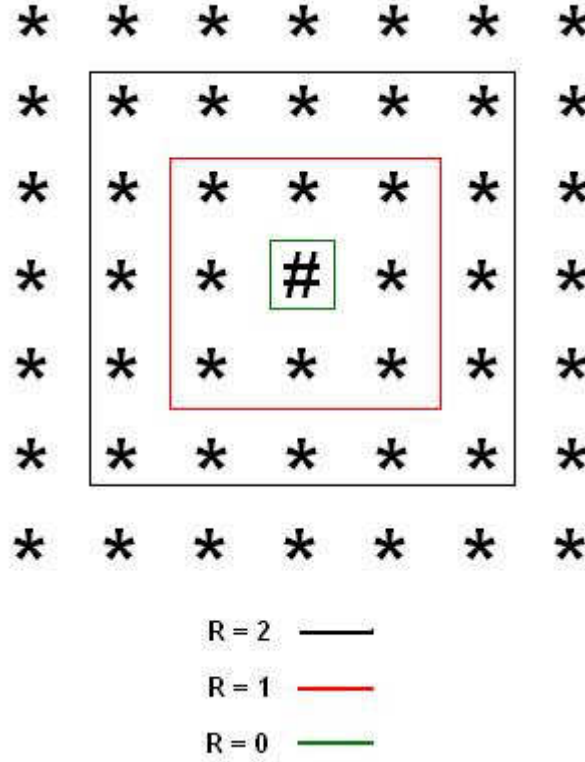
Şekil 2.7 N adet özniteliğe sahip girdi veri seti için tasarlanmış 3x3 boyutlu SOM'un yapısı.

Örnekte görülen 3x3 boyutlu SOM'da yer alan rekabet halindeki nöronların her biri, iki boyutlu uzayda belli bir pozisyona (x ve y koordinat bilgisine) sahiptir. Her nöronun bir ağırlık vektörü bulunmaktadır (W), ve bu vektör girdi vektörleri ile aynı boyuttadır. Örneğin, girdi verisi N boyutlu X vektörlerinden oluşuyor ise ($x_1, x_2, \dots, x_N$), her nöronun buna karşılık düşecek N boyutlu bir W ağırlık vektörü olacaktır ($w_1, w_2, \dots, w_N$).

Nöron ağırlık vektörleri, temsil ettikleri küme ile ilişkilendirilmiş girdi verileri için bir model teşkil etmektedir. Özdüzenleme (self-organization) sürecinde, her girdi örneği için kendisine en çok benzeyen ağırlık vektörüne sahip nöron bulunur ve bu nöron kazanan nöron olarak seçilir. Bu seçim uzaklık fonksiyonuna göre yapılır. Bu nedenle uzaklık fonksiyonu seçimi

kümeleme başarımını doğrudan etkiler. Kazanan nörona ve kullanılan komşuluk topolojisi ile yarıçapına göre belirlenen komşu nöronlara ait ağırlık vektörleri güncellenir (Laurene, 1993).

2, 1 ve 0 yarıçap değerleri ile dikdörtgensel topoloji için örnek komşuluk ilişkileri Şekil 2.8’de gösterildiği gibidir. Şekilde kazanan nöron “#” sembolü ile gösterilirken diğer nöronlar “*” sembolü ile gösterilmektedir. SOM sinir ağlarının komşuluk ilişkilerini belirtmek için dairesel, altıgen, vb. gibi farklı topolojiler de kullanılmaktadır.



Şekil 2.8 7x7 SOM üzerinde komşuluk ilişkilerinin dikdörtgensel topoloji ve 2, 1, 0 yarıçapları için gösterimi.

2.2.1.2 SOM Algoritması

SOM yapay sinir ağı modeline ait algoritma ve Kohonen öğrenme kuralı aşağıdaki gibidir (Laurene, 1993):

1 boyutlu uzayda tanımlı SOM için,

N adet özniteliğe sahip X girdi vektörleri $X_i=(X_{i1}, X_{i2}, \dots, X_{iN})$ olmak üzere ve

W ağırlık vektörleri: $W_j=(W_{1j}, W_{2j}, \dots, W_{Nj})$ olmak üzere;

Adım 0: Ağırlıklara ilk değerlerini rastgele ata.

Komşuluk parametrelerini (komşuluk topolojisi ve yarıçapını) ata.

Öğrenim oranı (α) parametrelerini ata.

Adım 1: Sonlandırma koşulu sağlanmadığı sürece Adım 2-8 i gerçekleştir.

Adım 2: Her girdi vektörü X için Adım 3-5 i gerçekleştir.

Adım 3: Her j için, aşağıdaki hesaplamayı gerçekleştir:

(Seçilen uzaklık fonksiyonunu kullanarak girdi verisi ile her nöron arasındaki mesafeyi hesapla. Bu örnekte Öklid uzaklığı kullanılmıştır.)

$$D(j) = \sum_{i=1}^N (w_{ij} - x_i)^2$$

Adım 4: $D(j)$ nin minimum olduğu J indeksini (yani kazanan nöronu) bul.

Adım 5: J 'nin tanımlanan komşuluğu içerisinde kalan tüm j birimleri ve tüm i 'ler için:

$$w_{ij}(\text{yeni}) = w_{ij}(\text{eski}) + \alpha [x_i - w_{ij}(\text{eski})]. \text{ (Kohonen öğrenme kuralı)}$$

Adım 6: Öğrenim oranını güncelle.

Adım 7: Belirlenen zamanlarda topolojik komşuluk yarıçapını azalt.

Adım 8: Sonlandırma koşulunu kontrol et. (İterasyon sayısına ulaşıldığında veya α eşik değere ulaştığında sona erdir.)

Öğrenme oranı α , zamana veya devir sayısına göre yavaş yavaş azalan bir fonksiyondur. Laurene (1993)'de belirtildiği üzere öğrenme oranının lineer olarak azalan bir fonksiyon olması (örneğin geometrik azaltım gibi) pratik hesaplamalar için yeterlidir. Kazanan nörona komşu olan nöronların tespit edilmesi için kullanılan komşuluk yarıçapı değeri de kümeleme işleme süresince gerekli anlarda azaltılmalıdır (Laurene, 1993).

3. İLGİLİ ÇALIŞMALAR

Bu bölümde, gen kümeleme konusu ile ilgili yürütülmüş önceki çalışmalara ilişkin bilgiler verilmektedir. Bu bağlamda, Bölüm 3.1’de gen kümeleme işlemi için günümüze dek kullanılagelmiş çeşitli algoritma ve metodlara değinilen genel bir gözden geçirme yer almaktadır. Bölüm 3.2’de ise, bu çalışma kapsamında genişletilen/geliştirilen Kasturi ve diğerleri (2005)’ne ait gen kümeleme algoritması detayları ile açıklanmaktadır.

3.1 Gen Ekspresyon Verisinin Kümelenmesi

Biyolojik olarak ilişkili gen grupları elde edebilmek için genleri ekspresyon profillerine göre kümelemek yaygın bir yaklaşım olup, Microarray teknolojisi ile elde edilen verilerin analizinde kullanılan temel tekniklerden biridir. Bu yaklaşımın dayandığı temel hipotez, gen ekspresyon profillerindeki benzerliğin (yani ko-ekspresyonun), genler arasındaki fonksiyonel veya düzenleyici bir benzerliğe işaret ettiğidir (Azuaje ve Dopazo, 2005). Bu nedenle, fonksiyonel olarak benzer veya aynı biyolojik proseste yer alan genleri tespit etme işlemi, genleri, ekspresyon profillerine dayanarak gruplara ayırma işlemine dönüşmüştür (Azuaje ve Dopazo, 2005). Bununla beraber unutulmaması gereken önemli bir nokta ise, Microarray veri analizi tekniklerinin henüz gelişimlerinin erken evrelerinde olduğu ve hala gelişimlerine devam ettiğidir (Baldi ve Brunak, 2001). Bu nedenle gerek Microarray teknolojisi gerekse de bu teknoloji ile elde edilecek gen ekspresyon verilerinin analizi için kullanılan metodların geliştirilmesine hala ihtiyaç duyulmaktadır.

Gen ekspresyon verisinin kümelenmesinde günümüze kadar pek çok farklı algoritma ve metod kullanılmıştır. Hiyerarşik kümeleme (Eisen ve diğerleri, 1998; Wen ve diğerleri, 1998), k-means kümeleme (Tavazoie ve diğerleri, 1999), özdüzenleyici haritalar (SOM) (Tamayo ve diğerleri, 1999; Törönen ve diğerleri, 1999; He ve diğerleri, 2003), destek vektör makineleri (support vector machines) (Brown ve diğerleri, 2000), Bayes ağları (Bayesian networks) (Friedman ve diğerleri, 2000), bulanık mantık yaklaşımı (Woolf ve Wang, 2000) bu amaçla kullanılmış metodlardandır.

Yukarıda belirtilen çalışmaların yanında, geliştirilen bazı metodlarda ise kümeleme başarımını arttırmak için sadece gen ekspresyon verisi ile yetinmeyip, bir araya getirilmiş farklı tipte ancak birbiri ile ilişkili genomik verilerin kullanılması önerilmektedir. Wang ve diğerleri (2005)’nin geliştirdiği ontolojiye-dayalı kümeleme metodu bu grup metodlara

verilebilecek örneklerden biridir. Wang ve diğerleri (2005) bu amaçla geliştirdiği metoddan gen kümelemesi için gen ekspresyon verisi ve gen ontolojisi verilerini bir arada kullanmıştır. Bu yaklaşıma bir diğer örnekte ise, bazı araştırmacılar çeşitli çalışmalarında (Holmes ve Bruno, 2000; Barash ve diğerleri, 2002; Kasturi ve diğerleri, 2005) genleri biyolojik olarak daha anlamlı gruplara ayırabilmek amacıyla, gen ekspresyon profillerini ve genlerin promotör bölgelerindeki nükleotid dizilim bilgilerini bir arada kullanarak kümeleme işlemini gerçekleştirmeyi önermiştir.

Söz konusu çalışmalardan Kasturi ve diğerleri (2005) tarafından sunulanında, SOM yapay sinir ağı modelinin öğrenme algoritması (Kohonen öğrenme kuralı) farklı genomik veri tiplerini bir arada kullanmayı olanaklı kılacak şekilde geliştirilmiştir. Kasturi ve diğerleri (2005) tarafından geliştirilen bu metodun en büyük artılarından biri, gen kümelemesi için birbirini tamamlayıcı nitelikte farklı veri tiplerini bir arada kullanmaya olanak tanıyan, model-bağımsız yani uyarlanabilir yapıda bir metod oluşudur. Kasturi ve diğerleri (2005)'nin söz konusu çalışmasında gen kümelemesi için iki farklı tipte genomik veri kullanılmıştır. Bu tez çalışması kapsamında Kasturi ve diğerleri (2005) tarafından sunulan metod üç farklı genomik veri tipini (gen ekspresyon verisi, motif sıklık verisi ve gen konum verisi) ve farklı uzaklık fonksiyonlarını kullanarak gen kümeleme işlemini gerçekleştirecek şekilde genişletilmiş/geliştirilmiş ve uygulamaya dökülmüştür. Kasturi ve diğerleri (2005)'ne ait söz konusu çalışmanın ve metodun detayları Bölüm 3.2'de yer almaktadır.

3.2 Bilgi Füzyonu Kullanılarak Çoklu Tipte Genetik Verinin Kümelenmesi

Bu tez çalışmasında kullanılan genetik veri füzyonu algoritması Kasturi ve diğerleri (2005) tarafından geliştirilmiştir. Söz konusu algoritmanın amacı, gen ekspresyon verisi kullanılarak elde edilen gen kümeleme başarımını, farklı tipte genomik verileri bir arada kullanarak arttırmaktır. Böylelikle sadece gen ekspresyon verisinin kullanılması ile elde edilememiş ancak biyolojik olarak ilişkili genlerden oluşan yeni gen kümeleri elde etmek hedeflenmiştir. Algoritma bu amaç ile gen ekspresyon verisi ve motif sıklık verilerini bir arada kullanmaktadır. Bölüm 1'de de değinildiği gibi bu tez çalışmasının ilk ve temel amacı mevcut algoritmayı genişletmek/geliştirmektir. Bu amaç ile algoritma üçüncü bir veri tipi ve yeni kümeleme metrikleri (yani uzaklık fonksiyonları) kullanacak şekilde genişletilmiş/geliştirilmiştir. Sonraki adımda ise çalışmanın ikinci amacı olarak, genişletilen algoritma uygulamaya dökülmüştür.

Kasturi ve diğerleri (2005) tarafından sunulan algoritmanın temelleri öz düzenleyici haritalara (SOM) dayanmaktadır. Bölüm 2’de de açıklandığı gibi bir çeşit yapay sinir ağı olan SOM’lar rekabete (competition) dayanan yapıdadır. Eğiticişiz (unsupervised) yapıda olan SOM’lar, girdi-çıkı çiftlerini içeren bir eğitim setine ihtiyaç olmaksızın kümeleme işlemini gerçekleştirebilirler.

Diğer kümeleme algoritmaları gibi SOM’lar da girdi verisi ve küme merkezleri arasındaki mesafeleri hesaplayıp en yakın küme merkezini (kazanan nöronu) bulabilmek için bir uzaklık fonksiyonuna ihtiyaç duymaktadır. Geleneksel SOM’lardan farklı olarak Kasturi ve diğerleri (2005) tarafından geliştirilen algoritma, birden fazla uzaklık fonksiyonunu kullanabilmektedir. Bu, SOM tekniğine getirilen önemli bir yeniliktir. Söz konusu algoritma farklı tipte verileri bir arada kullanabilmekte ve kullanılan her veri tipine özel farklı bir uzaklık fonksiyonunu kullanmayı olanaklı kılmaktadır. Böylelikle her bir veri tipi için kendi karakteristiğine en uygun uzaklık fonksiyonu kullanılabilir. Başarımın uzaklık fonksiyonu seçimi ile son derece ilintili olduğu kümeleme işlemi için Kasturi ve diğerleri (2005)’nin getirdiği bu özellik son derece önemli bir artıdır. Bu tekniğe göre kullanılan her bir veri tipi (gen ekspresyon verisi, motif sıklık verisi gibi) bir kategori olarak adlandırılmaktadır. Kategori sayısı k olmak üzere, her kategorinin kendisine ait bir uzaklık fonksiyonu D_i ($i = 1, 2, \dots, k$) olmaktadır.

Her kategorinin kendisi için en uygun uzaklık fonksiyonu kategoriler arasında değişiklik göstermektedir. Çünkü her kategorinin kendine has karakteristik özellikleri bulunmaktadır. Örneğin, Pearson korelasyon katsayısı (Pearson correlation coefficient) gen ekspresyon verisi için uygun bir uzaklık ölçütüyken, motif sıklık verisi için ise uygun bir uzaklık ölçütü değildir. Bu nedenle, birden fazla veri tipini bir arada kullanabilen bir kümeleme algoritmasında söz konusu veri tiplerinin tümü için tek bir uzaklık ölçütü (uzaklık fonksiyonu) kullanmak anlamlı değildir. Kasturi ve diğerleri (2005) tarafından sunulan algoritma, her kategoriye kendisine en uygun uzaklık fonksiyonu ile birlikte kullanabilmektedir. Önceki bölümlerde de belirtildiği gibi Kasturi ve diğerleri (2005) iki farklı tipte genetik veriyi (kategoriye) kullanarak kümeleme işlemini gerçekleştirmiştir. Bunlar gen ekspresyon verisi ve motif sıklık verisidir. Teze konu olan Kasturi ve diğerleri (2005) çalışmasından önce yayınladıkları Kasturi ve diğerleri (2003) çalışmasında aynı araştırmacılar, Kullback-Leibler (KL) ıraksaklığını, gen ekspresyonu verisi için genellikle kullanılagelmiş uzaklık ölçütü olan Pearson Korelasyonu ile karşılaştırmış ve gen ekspresyon verisi için Kullback-Leibler (KL) ıraksaklığının daha iyi sonuç verdiğini göstermiştir. Bu

nedenle, bu teze konu olan çalışmalarında da gen ekspresyon veri tipi için uzaklık ölçütü olarak Kullback-Leibler (KL) ıraksaklığını kullanmışlardır. Motif sıklık verisi için ise S-Dist adını verdikleri yeni bir uzaklık fonksiyonunu geliştirmiş ve kullanmışlardır. (Söz konusu uzaklık fonksiyonlarının ve başarımlarının detayları için bkz. Kasturi ve diğerleri (2005).)

Algoritma, farklı veri tiplerinin bir arada kullanımını her veri tipi ile “ağırlık” olarak adlandırılan bir değeri ilişkilendirmek suretiyle gerçekleştirir. Ağırlık değerleri, kümeleme işleminde her bir kategorinin ne kadar etkiye sahip olacağını belirlemek için kullanılır. Toplam kategori sayısı k olmak üzere, tüm kategoriler için atanan ağırlıklar $P=\{p_1, p_2, \dots, p_k\}$ vektörü ile ve $p_1+p_2+\dots+p_k=1$ olmak üzere ifade edilir (Kasturi ve diğerleri, 2005). Ağırlık vektörünün kullanım detayları, algoritma detaylarının sunulduğu takip eden bölümde yer almaktadır.

3.2.1 Algoritma Detayları

Kasturi ve diğerleri (2005) tarafından tanımlanan algoritmanın adımları aşağıdaki gibidir:

N adet küme merkezinin ($c_1, c_2, \dots, c_N$) ilk değerlerini ata. Girdi verisini her kategori için ayrı ayrı olmak üzere normalize et.

Her bir iterasyon n için aşağıdaki adımları uygula (her iterasyonda veri setindeki satırların yerini rastgele değiştir):

1. Bir gen “ g ” seç : X_g
 2. P ağırlık vektörünü kullanarak rastgele bir kategori r seç.
 3. Gen g ile her bir küme merkezi c_i arasındaki d_i uzaklığını, D_r uzaklık fonksiyonunu kullanarak hesapla (sadece kategori r 'ye ait X_g kolonlarını, yani özniteliklerini, kullan).
 4. Gen g 'ye en yakın küme l 'yi hesapla.
 5. Küme l 'ye ve komşularına ait ağırlıkları (tüm kategorilere ait tüm öznitelikleri), Kohonen öğrenme kuralına göre güncelle.
 6. Tüm kategoriler için öğrenme oranlarını güncelle.
- Yakınsama koşulu sağlanana kadar iterasyona devam et.

Algoritma, farklı veri tiplerinin birlikte kullanımını destekleyebilmek için Kohonen öğrenme kuralı olarak da bilinen SOM öğrenim algoritmasını birlikte-öğrenime olanak verecek şekilde genişletmektedir. SOM'lar, küme merkezlerinin nihai pozisyonlarını bulabilmek için iteratif bir yöntem kullanırlar. SOM algoritmasına dayalı olan bu metodda, iterasyon sürecinin her bir devirinde ağırlık vektörü P 'den faydalanılarak bir kategori rastgele seçilir. Seçilen kategori r

ve kendisine ait uzaklık fonksiyonu D_r kullanılarak sinir ağını oluşturan nöronlar arasından kazanan nöron belirlenir. Sonraki adımda ise uzaklık değerleri her ne kadar girdi verisinin belirli bir segmenti (seçilen kategoriye ait kolonlar) ve ilgili uzaklık fonksiyonu (seçilen kategoriye ait uzaklık fonksiyonu) kullanılarak hesaplanmış olsa da, tüm ağırlıklar Kohonen öğrenim kuralına göre güncellenir (Kasturi ve diğerleri, 2005).

Eğitim safhasının tamamlanması ile birlikte küme merkezlerinin nihai pozisyonları belirlenmiş olur. Böylelikle sıra her bir girdinin (genin), N adet küme arasından elemanı olduğu kümenin tespit edilmesine gelir. Bu aşama son bir iterasyon ile yapılır. Kasturi ve diğerleri (2005) söz konusu üyelik belirleme aşaması için, önceden belirlenen P kategori ağırlık dağılımlarının kullanılmasına dayanan yeni bir metod önermektedir. Bu metoda göre öncelikle her gen g için kendisine en yakın küme merkezi, her bir kategori için ayrı ayrı olmak üzere hesaplanır. Bu hesaplama, sadece söz konusu kategoriye oluşturan öznelilikler ve ilgili uzaklık fonksiyonu kullanılarak yapılır. Bu yol ile her bir kategoriye göre ait olunan kümeler (kazanan kümeler) bulunduktan sonra, gen g , ağırlık değeri en yüksek olan küme ile ilişkilendirilir. Örneğin, 3 adet kategori ve 10 adet küme olduğunu, 3 kategoriye ait ağırlık dağılımlarının ise $P=\{p_1=0.6, p_2=0.2, p_3=0.2\}$ olduğunu kabul edelim. Gen g için, her bir kategoriye göre hesaplanan en yakın küme merkezlerinin ise sırası ile küme-5, küme-3 ve küme-3 olduğunu varsayalım. Bu örneğe göre gen g 'nin küme-5'e ait olmasının toplam ağırlığı " $p_1=0.6$ " dır. Gen g 'nin küme-3'e ait olma toplam ağırlığı ise " $p_2+p_3=0.4$ " tür. Bu nedenle gen g , maksimum ağırlığa sahip küme-5 ile ilişkilendirilir. Ağırlık değerlerindeki olası beraberlik durumlarında ise seçim isteğe bağlı olarak yapılır.

4. KULLANILAN VERİ SETİ VE METODLAR

Bu bölümde, çalışma kapsamında kullanılan materyaller ve metodlar ayrıntılı biçimde açıklanmaktadır. Bölüm 4.1.'de ilk olarak kullanılan veri seti özellikleri detayları ile birlikte açıklanmaktadır. Mevcut algoritmanın ne şekilde genişletildiğine ve genişletilen haliyle algoritmanın uygulamaya ne şekilde döküldüğüne de yine bu bölümde değinilmektedir. Kullanılan SOM'a ait tasarım detayları Bölüm 4.5'te, elde edilen sonuçların ne şekilde değerlendirmeye tabi tutulacağı ise Bölüm 4.6'da açıklanmaktadır.

4.1 Veri Seti

Algoritma belirtildiği üzere yeni ve anlamlı gen kümeleri bulabilmek adına üç farklı tipte genomik veriyi bir araya getirmektedir. Bu veri tipleri;

- gen ekspresyon verisi,
- motif sıklık verisi,
- gen konum verisi,

olmak üzere üç tanedir. Bu nedenle, çalışma kapsamında kullanılmak üzere öncelikle söz konusu veri tiplerine ait bilgileri içeren uygun bir veri setine ihtiyaç duyulmaktadır.

Arabidopsis thaliana bitkisine ait veri seti geliştirilen algoritmayı test etmek için seçilmiştir. Arabidopsis, bitki türleri arasında yer alan önemli bir model organizmadır. Arabidopsis, daha üst düzey bitkiler ile aynı büyüme, gelişme, çiçek verme ve üreme süreçlerine sahip olup, DNA'sı ise mısır bitkisinden 30 kat daha küçüktür (Hunter, 1993). Bu özellikleri, Arabidopsis'i biyoinformatik çalışmalarında sıklıkla kullanılan bir tür haline getirmiştir. Arabidopsis bitkisine ait gen ekspresyon, motif sıklık ve gen konum bilgileri bu tez çalışması kapsamında kullanılmaktadır.

Kullanılan test veri seti 100 adet Arabidopsis genine ait gen ekspresyon, motif sıklık ve konum bilgilerini içermektedir. Bu veri seti, farklı 7 zaman dilimi için gen ekspresyon verisi, 614 motif için motif sıklık verisi ve gen konum verisinden oluşmaktadır. Gen konum verisi kromozom numarası, baz çiftlerinin başlangıç ve bitiş indeks numaraları olmak üzere 3 kolonda toplanan bilgidir. Kümeleme başarımını daha iyi ölçebilmek için, soğuk şartları altında önemli rol üstlendikleri bilinen 3 gen (AT4G25470, AT4G25480, AT4G25490) özellikle seçilen 100 gen arasına dahil edilmiştir. Diğer genler, her bir

Arabidopsis kromozomundan eşit adette olacak şekilde rastgele seçilmiştir. Bu amaçla, Arabidopsis'e ait 5 kromozomun her birinden 20'şer adet olmak üzere genlerin rastgele seçimi tamamlanmıştır. Her kromozomdan eşit miktarda gen seçilmesinin ardında yatan temel neden, her Arabidopsis kromozomundan genler içeren daha homojen yapıda bir veri seti oluşturabilmektir. Buna ek olarak gen konum bilgisinin kümeleme işlemindeki etkisini daha iyi görebilmek için, seçilen genlerin kromozomal dağılım kalitesi hakkında da emin olmak gerektiği söylenebilir. Tüm bu nedenlerden ötürü test veri seti oluşturulurken Arabidopsis kromozomlarının her birinden eşit sayıda gen seçilmiştir.

4.2 Veri Tipleri ve İlgili Uzaklık Fonksiyonları

Çalışma kapsamında kullanılan veri tipleri ve uzaklık fonksiyonları çiftleri aşağıdaki gibidir.

4.2.1 Arabidopsis Gen Ekspresyon Verisi

Gen ekspresyon verisi bu çalışma kapsamında kullanılan genomik veri tiplerinden biridir. Veri setini oluşturmak için kullanılan Arabidopsis gen ekspresyonu verisi soğuk şartları altında elde edilmiş olup Arabidopsis türü için uluslararası bir veri kaynağı olan TAIR (The Arabidopsis Information Resource) web sitesinden temin edilmiştir [8]. Bu veri, soğuk etkisinin erken safhalarındaki 7 farklı zaman noktasında elde edilmiş ekspresyon değerlerini içermektedir. Soğuk etkisinin başlatılmasından sonraki 30. dakika, 1. saat, 3. saat, 6. saat, 12. saat ve 24. saat için elde edilmiş ekspresyon verilerinden oluşmaktadır. “AT1G03590” genine ait örnek gen ekspresyon verisi Çizelge 4.1'deki gibidir:

Çizelge 4.1 Gen ekspresyon verisi örneği.

Gen Adı	0. dk	30.dk	1 sa	3 sa	6 sa	12 sa	24 sa
AT1G03590	0,999	1,011	0,699	0,711	0,62	0,223	0,249

Ekspresyon verisi kullanılmadan önce 0-1 aralığına normalize edilmiştir. Diğer bir deyiş ile, maksimum-minimum normalizasyonu adı ile de bilinen yöntem ile, ekspresyon değerleri en küçük değer 0 ve en büyük değer ise 1 olacak şekilde normalize edilmiştir. Normalizasyon için bu tez çalışması kapsamında geliştirilen programın ilgili özelliği kullanılmıştır (detaylar için bkz. Ek 2).

Bölüm 3.2’de açıklandığı üzere Kasturi ve diğerleri (2005) tarafından önerilen algoritma her veri tipi için farklı bir uzaklık fonksiyonunun kullanılmasına olanak tanımaktadır. Bu nedenle, gen ekspresyon veri ile kullanılmak üzere uygun bir uzaklık fonksiyonunun seçilmesi gerekmektedir. Knudsen (2002), gen ekspresyon verisi için vektör-açı-uzaklığının (vector angle distance), Pearson korelasyon uzaklığı (Pearson correlation distance) ve Öklid uzaklığından (Euclidean distance) daha iyi sonuç verdiğini göstermiştir. Kasturi ve diğerleri (2005) ise çalışmasında bu veri tipi için KL ıraksaklığını (KL divergence) kullanmıştır. Bununla beraber literatürde vektör-açı-uzaklığı ile KL ıraksaklıklarından hangisinin gen ekspresyon verisi için daha iyi sonuç verdiği üzerine bir araştırma bulunmamaktadır. Bu tez çalışmasında, Knudsen (2002)’in önerisinden hareketle, gen ekspresyon verisi için “vektör-açı-uzaklığı” uzaklık fonksiyonu olarak kullanılmıştır.

N boyutlu uzayda a ve b noktaları arasındaki vektör-açısı α aşağıdaki gibi hesaplanmaktadır:

$$\cos \alpha = \frac{\sum_{i=1}^N a_i b_i}{\sqrt{\sum_{i=1}^N a_i^2} \sqrt{\sum_{i=1}^N b_i^2}} \quad (4.1)$$

Daha sonra a ve b noktaları arasındaki vektör-açı-uzaklığı (VAU) aşağıdaki gibi hesaplanmaktadır (Knudsen, 2002):

$$VAU(a, b) = (1 - \cos \alpha) \quad (4.2)$$

ve aşağıdaki özelliklere sahiptir:

$$1. \quad 0 \leq VAU(a, b) \leq 1 \quad (4.3)$$

$$2. \quad VAU(a, a) = 0 \quad (4.4)$$

$$3. \quad VAU(a, b) = VAU(b, a) \quad (4.5)$$

4.2.2 Arabidopsis Motif Sıklık Verisi

Çalışma kapsamında kullanılan motif sıklık verisi, her gen için ayrı ayrı oluşturulmuş her motifin kaç kez tekrar ettiği bilgisini taşıyan bir vektör biçimindedir. Kolonların her biri bir motife karşılık düşmektedir ve her kolon için söz konusu motifin ilgili gende kaç kez tekrar ettiği bilgisi hesaplanmıştır. Eğer söz konusu motife ilgili gen üzerinde hiç rastlanmadı ise karşılık düşen değer 0'dır.

Arabidopsis'e ait motif sıklık verisi oluşturulurken 614 farklı motiftten faydalanılmıştır. Her motifin genler üzerinde kaç kez tekrar ettiğini hesaplayabilmek için öncelikle motiflere ve genlerin promotör bölgelerine ait dizilim (sequence) bilgileri elde edilmiştir. Daha sonra ise geliştirilen bir programcık ile her motifin her gen üzerinde kaç kez tekrar ettiği bilgisi elde edilmiştir. "AT1G03590" genine ait örnek motif sıklık verisi Çizelge 4.2'deki gibidir:

Çizelge 4.2 Motif sıklık verisi örneği.

Gen Adı	Motif "A000001"	Motif "A000002"	Motif "A000003"	Motif "A000004"	...
AT1G03590	16	2	0	0	...

Kasturi ve diğerleri (2005) tarafından geliştirilen metod her veri tipi için farklı bir uzaklık fonksiyonu kullandığı için, motif sıklık verisi için de uygun bir uzaklık fonksiyonuna gereksinim duyulmaktadır. Bu amaçla Kasturi ve diğerleri (2005) çalışmasında kullanılan *SDist* uzaklık fonksiyonu tercih edilmiştir. Motif sıklık verisi için kullanılacak uzaklık fonksiyonu seçilirken, seçilen fonksiyonun çok sayıda 0 içeren veri setleri ile düzgün olarak çalışabilmesine dikkat edilmiştir. Çünkü motif sıklık verisi çok sayıda 0'dan oluşmaktadır. *SDist* uzaklık fonksiyonu bu ihtiyacı karşılayabilmektedir. *SDist* uzaklık fonksiyonu iki motif sıklık vektörü arasındaki uzaklığı hesaplar, her motif için ayrı ayrı olmak üzere minimum ve maksimum motif tekrar etme bilgilerini kullanmaktadır. Yöntemin bu karakteristiği, toplam motif sayısından bağımsız olarak uzaklığın ölçülebilmesini olanaklı kılmaktadır (Kasturi ve diğerleri, 2005).

Motif sıklık verisi de diğer veri tipleri gibi kullanılmadan önce 0-1 aralığına normalize edilmiştir. Diğer bir deyiş ile, maksimum-minimum normalizasyonu adı ile de bilinen yöntem ile, motif sıklık değerleri en küçük değer 0 ve en büyük değer ise 1 olacak şekilde normalize

edilmiştir. Normalizasyon için bu tez çalışması kapsamında geliştirilen programın ilgili özelliği kullanılmıştır (detaylar için bkz. Ek 2).

N adet motif için elde edilmiş x ve y motif sıklık vektörleri arasındaki SDist uzaklığı aşağıdaki gibidir (Kasturi ve diğerleri, 2005):

$$SDist(x, y) = 1 - \frac{\sum_{i=1}^N \min(x_i, y_i)}{\sum_{i=1}^N \max(x_i, y_i)} \quad (4.6)$$

ve aşağıdaki özelliklere sahiptir:

$$1. \quad 0 \leq SDist(x, y) \leq 1 \quad (4.7)$$

$$2. \quad SDist(x, x) = 0 \quad (4.8)$$

$$3. \quad SDist(x, y) = SDist(y, x) \quad (4.9)$$

4.2.3 Arabidopsis Gen Konum Verisi

Bölüm 1’de belirtildiği üzere bu tez çalışmasının hedeflerinden biri Kasturi ve diğerleri (2005) tarafından bulunan metodu üçüncü bir genomik veri tipi ekleyerek geliştirmektir. Bu amacı gerçekleştirmek için “genlere ait kromozomal konum bilgileri” kullanılmıştır. Arabidopsis genlerine ait kromozomal lokasyon verisi de diğer veri tipleri gibi TAIR [8] kaynağından edinilmiştir. Bu verinin formatı, kromozom numarası, nükleotid çiftlerinin başlangıç indeksi ve bitiş indeksi olmak üzere üç bileşenlidir. “AT1G03590” ve “AT1G75860” genlerine ait örnek gen konum verileri Çizelge 4.3’teki gibidir:

Çizelge 4.3 Gen konum verisi örneği.

Gen Adı	Kromozom Numarası	Baz Çiftlerinin Başlangıç İndeksi	Baz Çiftlerinin Bitiş İndeksi
AT1G03590	1	894254	896439
AT1G75860	1	28487824	28489612

Gen lokasyon verisi için de diğer veri tiplerinde olduğu gibi uygun bir uzaklık fonksiyonu gerekmektedir. Yapılan literatür taraması sonucunda bu amaca hizmet eden bir fonksiyonun literatürde yer almadığı görülmüştür. Bu çalışma kapsamında gen konum bilgileri arasındaki

mesafeyi ölçebilmek amacı ile yeni bir lineer uzaklık fonksiyonu geliştirilmiştir. Farklı iki kromozomda yer alan genler arasındaki mesafenin baz çiftlerinin indeks bilgilerine bakılmaksızın maksimum olması gerektiğinden hareketle fonksiyon ilk olarak genlerin kromozom bilgilerini kontrol etmektedir. Eğer birbirlerine uzaklıkları ölçülmek istenen genlerin kromozom numaraları farklı ise, söz konusu iki gen arasındaki uzaklık maksimum uzaklığa, yani “1”e eşittir. Eğer kromozom numaraları eşit ise, iki gen arasındaki uzaklık baz çiftlerinin başlangıç konumları arasındaki mesafeye eşittir.

if “veri-1’in kromozom numarası” equals “veri-2’nin kromozom numarası” then

Uzaklık = Genlerin başlangıç konumları arasındaki lineer uzaklık

else

Uzaklık = 1; // maksimum uzaklık (4.10)

Genler arasındaki lineer uzaklık bulunurken genlerin başlangıç pozisyonları arasındaki mesafenin ölçülmesinin yeterli hassasiyeti sağlayacağına karar verilmiştir. Bu kararın ardında yatan temel neden her Arabidopsis kromozomunda on milyonlarca baz çifti bulunması ve her Arabidopsis geninin ise yaklaşık olarak 2000-3000 baz çiftinden oluşmasıdır. Bu nedenle, aynı kromozom üzerinde yer alan iki gen arasındaki mesafenin ölçümünde genlerin başlangıç indekslerinin kullanılmasının yeterli ölçüm duyarlılığını sağlayacağı görülmektedir.

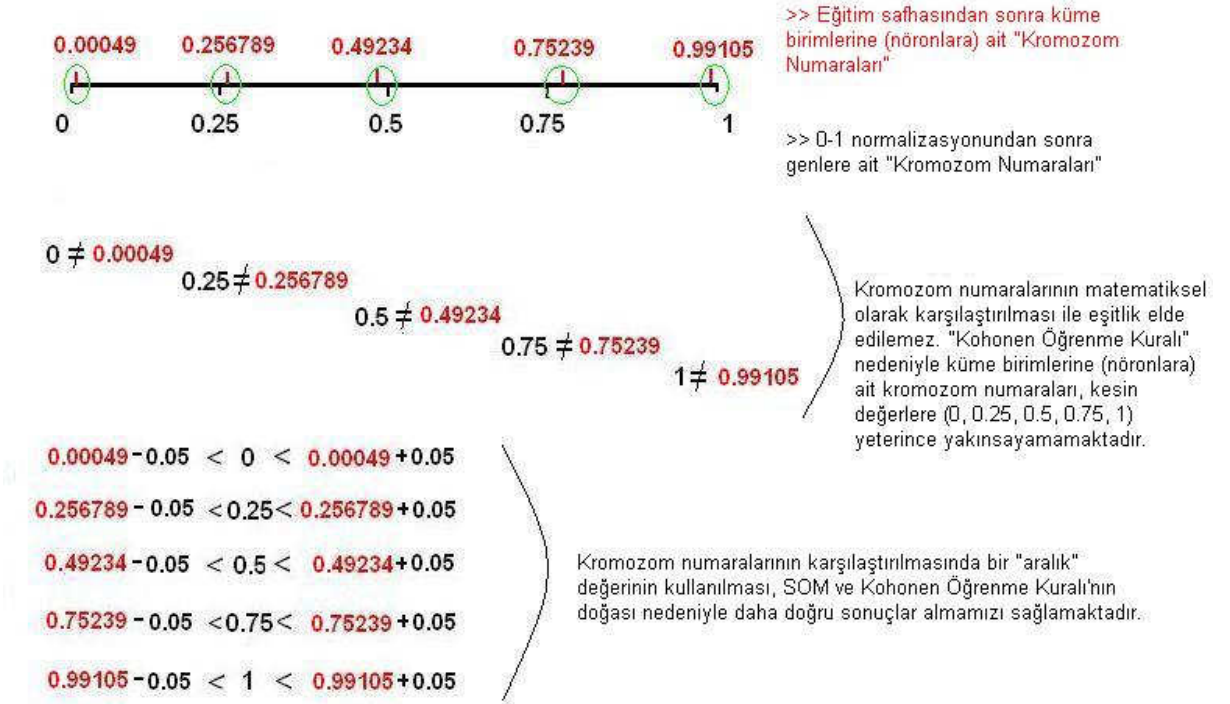
Geliştirilen algoritma her genomik veri tipini normalize edilmiş halde kullanmaktadır. Gen konum verisi de bu nedenle maksimum-minimum normalizasyonu ile 0-1 aralığına normalize edilmiştir. Böylelikle iki gen arasındaki maksimum uzaklığın 1, minimum uzaklığın ise 0 olması garantilenmiştir.

Geliştirilen bu uzaklık fonksiyonu ile ilgili vurgulanması gereken bir diğer önemli nokta da kromozom numaralarının karşılaştırılması hakkındadır. Kümeleme işleminde gen-gen arası uzaklıklar yerine, genler ile kümeler arasındaki uzaklığı ölçtüğümüz için, ve kümelere ait ilk değerler rastgele atanıp Kohonen öğrenim kuralına göre (bkz. bölüm 2.4.2) güncellendiği için, kümelere, dolayısı ile nöronlara ait kromozom numaraları her zaman ondalık değerli bir sayı olacaktır. Oysa, Arabidopsis kromozom bilgileri 0-1 aralığına normalize edilecek ve her kromozom spesifik bir değere sahip olacaktır. Diğer bir deyişle, 5 kromozoma sahip Arabidopsis veri setinin kromozom numaraları (1, 2, 3, 4, 5) normalizasyon işleminden sonra 0; 0,25; 0,5; 0,75 ve 1 olacaktır. Bu nedenle eğer kromozom numaralarının eşit olup olmadığını matematiksel eşitlik ile kontrol edersek, genler ve küme noktaları arasındaki uzaklık her zaman maksimum uzaklığa eşit olacaktır. Çünkü küme merkezlerine yani

nöronlara ait kromozom bilgisi hiçbir zaman gen kromozom numaralarına eşit olacak kadar yakınsayamayacaktır (örneğin, $0,25 \neq 0,2455999999$). Bu nedenle, gen kromozom numaralarının küme noktalarına ait kromozom numarasına eşit olup olmadığını tespit edebilmek için bir aralık değerine gerek duyulmaktadır. Bunun için geliştirilen fonksiyon aşağıdaki gibidir:

if küme kromozom değeri - 0.05 < gen kromozom numarası < küme kromozom değeri + 0.05 then kromozom numaraları eşittir!

else kromozom numaraları eşit değildir! (4.11)



Şekil 4.1 Gen konum bilgisini kümeleme işleminde kullanabilmek amacı ile, gen ve kümelere (nöronlara) ait kromozom numaralarını karşılaştırabilmek için bulunan çözüm.

Geliştirilen uzaklık fonksiyonu GLDist (Gene Location Distance) olarak adlandırılmıştır ve aşağıdaki özelliklere sahiptir:

$$1. \quad 0 \leq GLDist(x, y) \leq 1 \quad (4.12)$$

$$2. \quad GLDist(x, x) = 0 \quad (4.13)$$

$$3. \quad GLDist(x, y) = GLDist(y, x) \quad (4.14)$$

4.3 Algoritmanın Uygulamaya Dökülüşü

Giriş bölümünde de belirtildiği üzere bu tez çalışmasının hedeflerinden birisi de Kasturi ve diğerleri (2005) tarafından önerilen algoritmanın, bu tez çalışması kapsamında genişetilecek/geliştirilecek halini gerçekleştiren bir program hazırlamaktır. Program, nesneye yönelik bir programlama dili olan C# ile, MS .NET Framework V.2 teknolojisi kullanılarak geliştirilmiştir. Bu nedenle, programın kullanılmak istendiği bilgisayarlarda .NET Framework V.2 kurulu olmalıdır. (Geliştirilen uygulama hakkında daha fazla bilgi bkz. Ek-2)

4.4 Algoritmanın Genişletilmesi/Geliştirilmesi: SOM Yapay Sinir Ağı Modeli ile Bilgi Füzyonu ve Kümeleme

Kasturi ve diğerleri (2005) tarafından sunulan algoritma bu tez çalışması kapsamında iki farklı anlamda geliştirilmiştir:

- *İlk olarak algoritmayı gen ekspresyon ve motif sıklık veri tiplerinin yanında üçüncü bir genomik veri tipiyle de çalışacak şekile getirmek hedeflenmiştir.* Bu amaç ile kullanılacak üçüncü genomik veri tipi olarak gen konum verisi seçilmiş ve algoritma bu veriyi kullanabilecek şekilde genişletilmiştir. (Gen konum verisinin detayları için bkz. Bölüm 4.2.3)

Kasturi ve diğerleri (2005) tarafından geliştirilen algoritma her genomik veri tipi için farklı bir uzaklık fonksiyonu (en uygunu) kullanmaya olanak tanıdığı için (bkz. Bölüm 3.2), eklenen bu üçüncü veri tipi için de uygun bir uzaklık fonksiyonu belirlemek ve kullanmak gerekmektedir. Bu amaç ile, gen konumları arasındaki uzaklığı ölçebilmek için yeni bir lineer uzaklık fonksiyonu geliştirilmiştir (detaylı bilgi için bkz. Bölüm 4.2.3).

- *Algoritmayı farklı kümeleme metrikleri (uzaklık fonksiyonları) kullanarak genişletmek hedeflenmiştir.* Kümeleme işleminde ilk adım, benzerlik kavramı (dolayısıyla uzaklık kavramı) için matematiksel bir tanım oluşturmaktır. Kümeleme işleminin performansı büyük ölçüde seçilecek uzaklık fonksiyonuna bağlıdır. Bu nedenle, kümeleme işleminin yapılmak istendiği veri seti için, söz konusu veri setinin karakteristik özelliklerine uygun bir uzaklık fonksiyonu seçilmesi hayati önem taşımaktadır. Bu çalışma kapsamında Kasturi ve diğerleri (2005) tarafından kullanılan iki uzaklık fonksiyonunun alternatifleri araştırılmış ve kullanılan fonksiyonlardan biri değiştirilmiştir. Eklenen üçüncü genomik veri tipi için ise yeni bir uzaklık fonksiyonu oluşturulmuştur.

Knudsen (2002), gen ekspresyon verisinin kümelenmesinde vektör-açı-uzaklığının, Pearson korelasyonu ve Öklid uzaklığından daha iyi sonuç verdiğini göstermiştir. Bununla beraber gen ekspresyon verisinin kümelenmesinde, vektör-açı-uzaklığı ile Kasturi ve diğerleri (2005)'nin çalışmasında kullandığı KL ıraksaklığından hangisinin daha iyi sonuç verdiğini irdeleyen bir çalışma literatürde bulunmamaktadır. Bu tez çalışmasında, Knudsen (2002)'in çalışmasındaki öneriden hareketle gen ekspresyon veri tipi ile birlikte kullanılmak üzere uzaklık fonksiyonu olarak vektör-açı-uzaklığı seçilmiştir.

Kasturi ve diğerleri (2005) tarafından motif sıklık veri tipiyle kullanılmak üzere geliştirilen SDist uzaklık fonksiyonu ise değiştirilmeden bu çalışmada da kullanılmıştır. Motif sıklık verisi çok sayıda 0 değeri içerdiği için, bu veri tipi için kullanılacak uzaklık fonksiyonunun çok sayıda 0 değeri içeren vektörler arasında da sorunsuz şekilde çalışıyor olması gerekmektedir. SDist uzaklık fonksiyonu, Genişletilmiş Jaccard Benzerlik Katsayısı'na (Extended Jaccard Similarity Coefficient) dayanarak oluşturulmuştur ve bu uzaklık fonksiyonu, veri setinde yer alan toplam motif sayısından bağımsız olarak çalışabilmektedir (Kasturi ve diğerleri, 2005). Kasturi ve diğerleri (2005), söz konusu gerekliliği şu şekilde vurgulamıştır: “Motif sıklık verisi için bir uzaklık ölçüsü oluştururken dikkat edilmesi gereken son derece önemli bir nokta, genellikle tek bir genin, motif sıklık verisi oluşturmak için kullanılan tüm motifleri içermediğinin göz önünde bulundurulmasıdır.” SDist uzaklık fonksiyonu yukarıda tanımlanan ihtiyaçları karşılayabilmektedir. SDist uzaklık fonksiyonu hakkında detaylı bilgi bölüm 4.2.2’de yer almaktadır.

Gen konum verileri arasındaki uzaklığı ölçebilmek için ise, bölüm 4.2.3’te açıklanan lineer uzaklık fonksiyonu kullanılmıştır.

4.5 Kullanılan SOM’un Tasarımı

Bu çalışma kapsamında kullanılan SOM yapay sinir ağı, iki boyutlu ve 5x5 topolojisinde tasarlanmıştır. Bu nedenle, tasarlanan bu SOM ile elde edilebilecek maksimum küme sayısı 25’tir. Kasturi ve diğerleri (2005)'nin çalışmasında kullandığı 100 adet genden oluşan veri setinin geliştirdikleri metod ile her kümelenişinde, 16 ile 20 arasında değişen sayıda gen kümesi elde edildiği belirtilmiştir. Bu tez çalışması kapsamında kullanılan veri seti de Kasturi ve diğerleri (2005)'nin kullandığı ile aynı boyutta olduğu için, gerçekleştirilecek kümeleme işlemi sonucunda yaklaşık 20 kümenin elde edileceğini kabul etmek ve kullanılacak SOM’un

boyutunu bu kabulü göz önünde bulundurarak tasarlamak makuldür. Kasturi ve diğerleri (2005)'nin bu tez çalışmasına konu olan uygulamasında elde ettiği bu tecrübeye dayanarak, kullanılan SOM 5x5 topolojisinde tasarlanmış ve beklenen küme sayısını destekleyecek bir SOM oluşturulmuştur (Bu SOM ile 25 adet küme elde edilebilir). Seçilen 5x5'lik topolojinin kullanılan veri seti için uygunluğu sonuçlar bölümünde de görülmüştür. Sonuçlar bölümünde görüleceği üzere metod her çalıştırıldığında 16-22 arasında değişen sayıda gen kümesi elde edilmiştir (bkz. Bölüm 5).

Bölüm 2'de de belirtildiği gibi SOM yapay sinir ağlarının ön tanımlı iki temel parametresi bulunmaktadır. Bunlar öğrenme oranı (α) ve komşuluk yarı çapı (R)'dir. Bu çalışmada öğrenme oranı (α)'nın ilk "0,9" olarak atanmış ve Laurene (1993)'teki öneriden hareketle geometrik olarak azaltılmıştır. Öğrenme oranı (α)'nın geometrik olarak azaltımı aşağıdaki denklemde görüldüğü gibi gerçekleştirilmiştir:

$$\alpha_{t+1} = \frac{\alpha_t}{2} \quad (4.15)$$

Tasarlanan SOM'un komşuluk ilişkileri *diktörtgensel topolojiye* göre belirlenmiştir (detaylı bilgi için bkz. Bölüm 2). Komşuluk yarıçapı (R)'nin ilk değeri 2 olarak atanmıştır. Toplam iterasyon sayısı ise 1050 olarak belirlenmiştir. Yarıçap değerinin $R=2$, $R=1$ ve $R=0$ olduğu durumlarda eşit sayıda iterasyon yapılarak öğrenme sürecinin gerçekleşmesi için 350. iterasyon ile 700. iterasyonlarda (R) değeri 1 azaltılmıştır.

Komşuluk yarıçapı (R)'nin ilk değerinin ne olacağına ve söz konusu değerin SOM öğrenim süreci boyunca ne şekilde azaltılacağına yapılan çeşitli testler sonucunda karar verilmiştir. Komşuluk yarıçapı ile ilgili belirlenecek söz konusu iki kriter, elde edilecek nihai küme adetlerini ve kalitelerini doğrudan etkilemektedir. Bu nedenle bu kriterler kullanılan SOM haritasının boyutları da göz önünde bulundurularak dikkatli bir şekilde seçilmelidir. Bu çalışma kapsamında kullanılan 5x5 SOM haritası için (R)'nin tüm harita elemanlarını kapsamayı sağlayacak maksimum değeri 4'tür. Yapılan çeşitli denemeler sonucunda yarıçap ilk değerinin 4 veya 3 olarak atanması durumunda, elde edilen nihai küme sayılarının beklenilenden çok daha az sayıda olduğu tespit edilmiştir (küme sayılarının 1 ile 7 aralığında değiştiği görülmüştür). Elde kümelerle yakından bakıldığında ise genellikle 1 veya 2 adet çok büyük küme oluştuğu ve kümeleme işlemine tabi tutulan genlerin büyük çoğunluğunun da bu az sayıdaki kümelerde gruplandığı tespit edilmiştir. Dolayısıyla kümeleme kalitesi düşüktür.

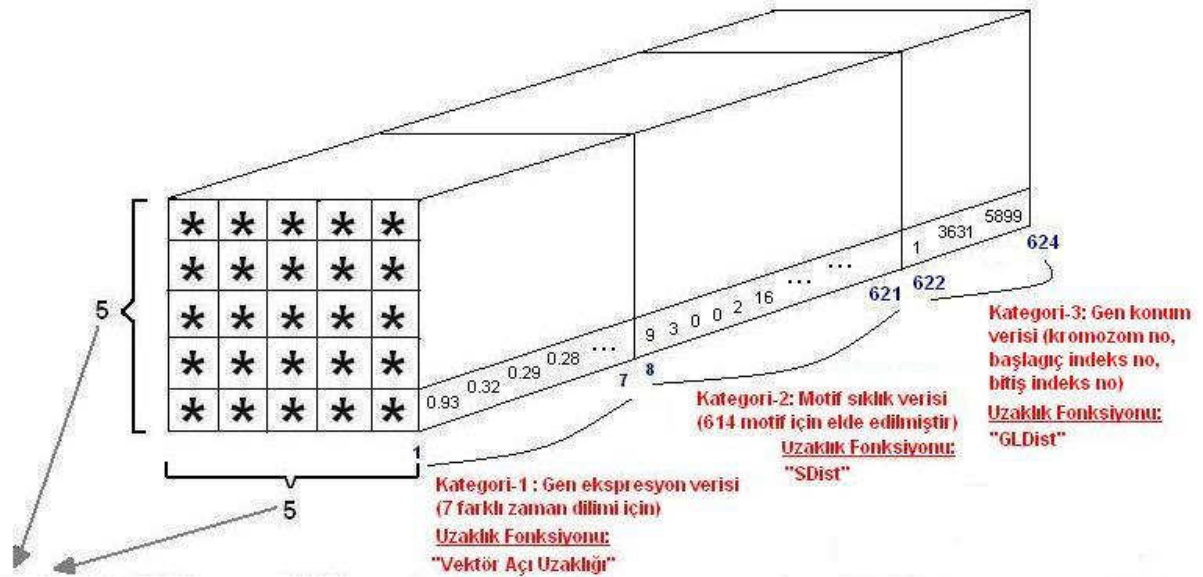
İterasyon sayısının arttırılması durumunda mevcut durumun daha da kötüleştiği ve kümeleme işlemi sonucunda tek bir büyük küme elde edildiği tespit edilmiştir.

Yarıçap ilk değerinin 4 veya 3 olduğu durumlarda az sayıda küme elde edilmesinin ve iterasyon sayısının arttırılması ile söz konusu sayının daha da azalarak 1'e kadar düşmesinin ardında yatan nedenin, SOM öğrenme sürecinin büyük kısmında yarıçap büyüklüğü nedeniyle belirli nöronları kapsaması gereken komşuluğun tüm SOM'u kapsaması temeline dayandığı görülmüştür. Diğer bir deyiş ile iterasyonların büyük çoğunluğunda neredeyse tüm nöronlar kazanan nöron ile birlikte güncellenmekte bunun sonucunda da nöronlar (küme merkezleri) farklı konumlara sahip olmak yerine aynı noktaya yakınsamakta ve farklı gen gruplarını temsil etmekten uzak final konumlarına sahip olmaktadır. Bu tespitten sonra yarıçap ilk değerinin kullanılan 5x5 boyutlarındaki SOM için 4 veya 3 gibi büyük bir değer olarak atanabilmesi için, iterasyon sayısının eşit aralıklarla azaltılması yerine (R)'nin büyük değerleri için az iterasyon, küçük değerleri için ise daha fazla iterasyon yapılarak çözülüp çözilemeyeceği sınanmıştır. Bu amaçla, R 'nin ilk değeri 4 için yapılan iterasyon sayısı " n " olmak üzere, $R=3$ için " $2n$ " adet iterasyon, $R=2$ için " $4n$ " adet iterasyon, $R=1$ için " $8n$ " adet iterasyon, $R=0$ için " $16n$ " adet iterasyon yapılması gibi 2'nin kuvvetlerini takip edecek yarıçap/iterasyon-adedi dağılımı denenmiştir. Benzer şekilde 3'ün kuvvetleri için de söz konusu yarıçap/iterasyon-adedi dağılımları denenmiştir. Bu yaklaşım ile iterasyon sayısının artması durumunda nihai küme sayısının 1'e yaklaştığı durumlar bir ölçüde azaltıldı ise de, elde edilen nihai küme sayıları ve kaliteleri yine beklendiği düzeyde olmamış ve büyük kümeler oluştuğu gözlenmiştir.

Yukarıda özetlenen denemeler sonucunda bu tez çalışmasında kullanılan veri seti ve 5x5 SOM haritası için komşuluk yarıçapı R 'nin ilk değerinin 2 olduğu, toplam iterasyon sayısının ~1000 civarında olduğu ve R değerinin toplam iterasyon sayısının eş aralıklarına ulaşıldığında 1 azaltıldığı durumda en makul sayıda küme elde edildiği ve elde edilen kümelerin gen dağılımlarının kaliteli olduğu tespit edilmiştir.

Tasarlanan SOM'un detaylı yapısı şekil 4.2'de gösterilmektedir. Kullanılan veri yapısı da (gen ekspresyon verisi, motif sıklık verisi ve gen lokasyon verisinden oluşan vektörel yapı) yine şekil 4.2'de gösterilmektedir. Geleneksel SOM yapay sinir ağı modeli, sadece tek bir tip veri ve tek bir uzaklık fonksiyonu ile kullanılabilirken, SOM'un Kasturi ve diğerleri (2005) tarafından geliştirilen versiyonu farklı veri tipleri ve bu tiplere ait farklı uzaklık fonksiyonları ile birlikte kullanılabilir.

İki Boyutlu, 5 x 5 Izgara Yapısında SOM



5 x 5 SOM = "25 " nöron = "25 " potansiyel küme

- 5x5 SOM, 25 nörona sahiptir
- Her nöron, potansiyel bir küme merkezini temsil eder. Bu neden ile, bu SOM ile elde edilebilecek maksimum küme sayısı 25'tir.
- Tasarlanan SOM, 3 farklı kategoride (tip) veriye sahiptir: gen ekspresyon verisi, motif sıklık verisi ve gen konum verisi.
- Her kategori, kendine ait bir "uzaklık fonksiyonuna" sahiptir.
- En yakın küme (kazanan nöron), her seferinde sadece tek bir veri tipi (yani tek bir kategori ve sadece o kategoriye ait kolonlar kullanılarak) ve o veri tipine ait uzaklık fonksiyonu kullanılarak bulunur.
- Kazanan nörona ve bu nörona komşu olan diğer nöronlara ait ağırlıklar (tüm kategorilere karşı gelen tüm kolonlar), Kohonen Öğrenme Kuralı'na göre güncellenir.

Şekil 4.2 Kullanılan SOM'un detaylı yapısı.

4.6 Sonuçların Değerlendirilmesi

Geliştirilen yöntemin başarısını değerlendirmek için kümeleme işleminde sadece gen ekspresyon verisinin kullanılması ile elde edilen gen kümeleri, her üç veri tipinin bir arada kullanılması ile elde edilen gen kümeleri ile karşılaştırılacaktır. Uygulanan yöntemin başarısı, üç farklı genomik veri tipinin kümeleme işleminde birlikte kullanılması sonucunda “yeni ve biyolojik olarak ilişkili gen kümelerinin” tespit edilip edilemediği değerlendirilerek ölçülecektir. Eğer bu nitelikte gen kümeleri tespit edilebilir ise, bu çalışma kapsamında geliştirilen metodun yeni ve biyolojik olarak ilişkili gen kümelerini tespit edebildiği sonucuna varılmış olacaktır. Böylelikle gen kümeleme işlemi için genetik veri füzyonunun önemi ve uygulanan yöntemin başarımı değerlendirilmiş olacaktır.

Elde edilecek gen kümelerinin biyolojik olarak ilişkili genlerden oluşup oluşmadığı söz konusu genlerin *moleküler fonksiyon* ve *biyolojik proses* bilgileri değerlendirilerek

gerçekleştirilecektir. Söz konusu bilgileri elde etmek için Gen Ontolojisi [2] ve TAIR [8] veri tabanları kullanılacaktır.

TAIR (The Arabidopsis Information Resource), üst seviye bir bitki modeli olup genetik çalışmalarında sıklıkla kullanılan *Arabidopsis thaliana* bitkisine ait genetik ve moleküler biyolojik bilgilerin saklı tutulduğu, çeşitli genetik araştırma gruplarının söz konusu tür üzerinde yürüttüğü kabul görmüş çalışmalar doğrultusunda her iki haftada bir güncellenen, bu yönüyle de söz konusu türe ait temel bilgi kaynağı olma özelliğini koruyan bir veritabanıdır. TAIR aynı zamanda Gen Ontolojisi Projesi' nin katılımcılarından.

Gen Ontolojisi Projesi, farklı canlı türlerine ait genlerin ve gen ürünlerinin niteliklerini tanımlamak için kontrollü bir isimlendirme (anotasyon) sistemi oluşturmayı hedef alan, uluslararası katılımı sürdürülen son derece önemli bir projedir. Genetik çalışmalarında karşılaşılan en önemli problemlerden birisi, veri kaynakları arasındaki isimlendirme standartı eksiklikliğinden kaynaklı farklılıklardır. Aynı bilgiler ve kavramlar, farklı kaynaklarda farklı şekilde ifade edilebilmektedir. Söz gelimi, bir araştırma sonucunda tespit edilen bir gen özelliği bir başka çalışma sonucunda da tespit edilmiş, ancak söz konusu sonuçlar farklı genetik veritabanlarına ve farklı isimler altında kaydedilmiş olabilir. Böylelikle özünde aynı olan ancak farklı şekilde ifade edilen bilgiler gelecekte bu bilgilerden faydalanarak çalışmalarını sürdürecek araştırmacılar için sorun oluşturmaktadır. Örneğin literatürde “glikoz sentezi”, “glikoz biyosentezi”, “glikoz formasyonu”, “glikogenez” gibi isimlerle adlandırılan işlemlerin tümü aslında aynı olup, daha basit bileşenlerden hareketle glikoz üretiminin gerçekleştirilmesini ifade etmek için kullanılmaktadır [9]. Söz konusu istenmeyen durumun beraberinde getirdiği en büyük dezavantaj, özellikle farklı türler ve veritabanları arasında karşılaştırma yapmanın oldukça zor hale gelmesidir. Tüm bu nedenlerden ötürü, tüm türleri kapsayacak kontrollü bir isimlendirme standartına gereksinim duyulmaktadır.

Gen Ontolojisi Projesi ile söz konusu problemleri ortadan kaldırmak için genlere/gen ürünlerine ait üç farklı bilgi türü için belirli bir isimlendirme standardı getirmek hedeflemiştir.

Bu bağlamda bir gen ürününün;

- ne yaptığını (işlevini),
- bu aktiviteleri ne için yaptığını (amacını) ve
- nerede yaptığını (görev yerini),

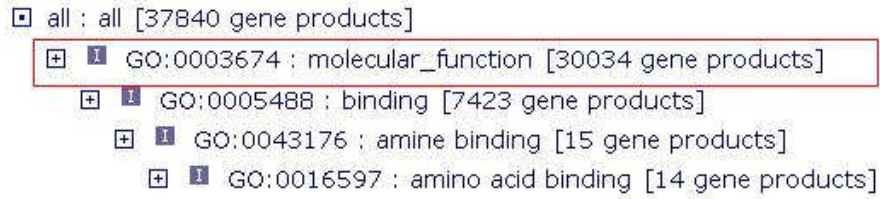
ifade edecek bilgiler düzenlenmiştir. Söz konusu bilgiler sırası ile *moleküler fonksiyon*, *biyolojik proses* ve *hücreyel bileşen* başlıkları altında toplanan anotasyonlar ile ifade edilmiştir. Her anotasyonun kendisine has ve eşsiz bir GO ID'si bulunmaktadır.

Moleküler fonksiyonlar, moleküler seviyede gerçekleşen aktiviteleri ifade etmek için kullanılmaktadırlar [2]. Örneğin, bağlanma aktivitesi (binding activity, GO:0005488) ve katalitik aktivitesi (catalytic activity, GO:0003824) GO veritabanında tanımlı moleküler fonksiyonlar arasında yer almaktadır. Kavramın daha iyi anlaşılabilmesi için bir diğer örnek olarak “çekiç” aletinin bir gen ürünü olduğunu düşünürsek, bu gen ürünün moleküler fonksiyonunun “çivi çakmak” olduğunu söyleyebiliriz.

Biyolojik prosesler, bir veya birden çok sıralı biçimde yürütülen moleküler fonksiyon grubu tarafından gerçekleştirilen işlemlerdir [2]. Örneğin organ gelişimi/büyümesi (organ growth, GO:0035265) gibi. Biyolojik prosesleri ve moleküler fonksiyonları birbirinden ayıran temel özellik, bir biyolojik prosesin birden çok ve farklı adımın gerçekleşmesi ile tamamlanmasıdır. Çekiç aracını bir gen ürünü olarak kabul ettiğimiz önceki örneğe göre bu gen ürünün biyolojik prosesinin örneğin “marangozluk” olduğunu söyleyebiliriz.

Hücreyel bileşen kavramı adından da anlaşılabilceği gibi hücrede yer alan bir birimi ifade etmekte olup, genellikle gen ürününün görev aldığı yeri ifade etmek için kullanılmaktadır [2]. Çekicinin bir gen ürünü olarak kabul edildiği örneğe göre, bu gen ürünün hücreyel bileşen bilgisinin “marangoz atölyesi” olduğunu söyleyebiliriz.

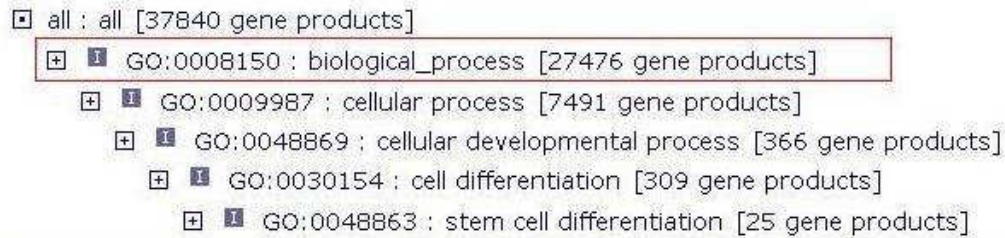
Moleküler fonksiyon, biyolojik proses ve hücreyel bileşen olmak üzere üç ayrı başlık altında toplanan anotasyonlar, genelden-özele doğru olmak üzere hiyerarşik bir ağaç yapısı meydana getirecek şekilde organize olmuştur. Ağacın üst kademelerinde yer alan anotasyonlar daha fazla gen ürünü tarafından paylaşılan daha genel anotasyonlar iken, ağacın alt hiyerarşik birimlerine inildikçe daha az gen ürünü tarafından paylaşılan daha özel anotasyonlara ulaşılmaktadır. Örneğin moleküler fonksiyon ağacı üzerinde alt hiyerarşik birimlere inildikçe daha spesifik bir moleküler fonksiyona ulaşılmaktadır. Bununla beraber ağaç üzerinde yer alan bir moleküler fonksiyon bağlı olduğu üst-hiyerarşik moleküler fonksiyonun bir parçası iken, kendisine bağlı olan alt-hiyerarşik moleküler fonksiyonlar ise kendisinin parçalarıdır (bkz. Şekil 4.3). Aynı durum biyolojik proses ve hücreyel bileşen başlıkları altında toplanan GO anotasyonları için de geçerlidir.



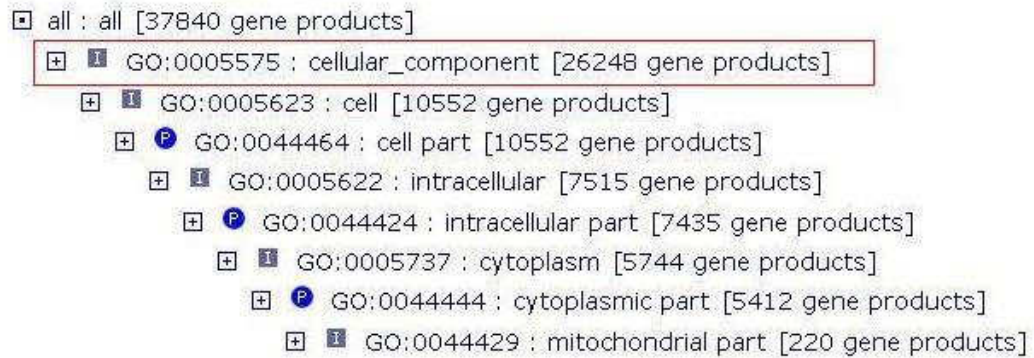
Şekil 4.3 GO:0016597 ID numaralı örnek moleküler fonksiyonunun, moleküler fonksiyon anotasyon ağacı üzerindeki yeri [2].

GO:0016597 ID numaralı “amino acid binding” adlı örnek moleküler fonksiyonun GO moleküler fonksiyon anotasyon ağacı üzerindeki yeri Şekil 4.3’te gösterildiği gibidir. Şekilde kırmızı kutu ile belirtilen anotasyon ağacının kökünü tüm diğer moleküler fonksiyonları kapsayan “moleküler fonksiyon, GO:0003674” adlı anotasyon oluşturmaktadır. Anotasyon adının hemen yanında köşeli parantezler arasında yer alan rakam, söz konusu anotasyonun kaç adet gen ürünü tarafından paylaşıldığını göstermektedir. Görüleceği üzere ağaç üzerinde alt hiyerarşik birimlere inildikçe daha spesifik anotasyonlara ulaşılmakta ve söz konusu anotasyonları paylaşan gen sayısı azalmaktadır. Aynı durum biyolojik proses ve hücresel bileşen başlıkları altında toplanan GO anotasyonları için de geçerlidir (bkz. Şekil 4.4).

(a) "Biyolojik Proses" ağacı



(b) "Hücresel Bileşen" ağacı



Şekil 4.4 (a) GO:0048863 ID numaralı örnek biyolojik prosesin, biyolojik proses anotasyon ağacı üzerindeki yeri [2]. (b) GO:0044429 ID numaralı örnek hücresel bileşenin, hücresel bileşen anotasyon ağacı üzerindeki yeri [2].

Böylelikle GO projesi ile birlikte bir isimlendirme standartının sağlanmasının yanında, söz konusu bilgilerin hiyerarşik sınıflandırması da gerçekleştirilmiştir. Özellikle genetik temelli veriler gibi yorum yapabilmenin ve mevcut ilişkileri görebilmenin çoğu zaman önemli bir bilgi birikimi ve hatta uzmanlık gerektirebildiği veri kaynakları üzerinde çalışan araştırmacılar için bu önemli bir kazanımdır. Örneğin, GO anotasyonlarının olmadığı bir ortamda analize tabi tuttuğumuz iki gen arasında bir benzerlik olup olmadığını araştırdığımızı düşünelim. Genlerden birinin son derece spesifik bir fonksiyona sahip olduğunu literatürde yaptığımız araştırma sonucu tespit ettiğimizi varsayalım. Yine literatür de yaptığımız araştırma sonucunda diğer genin daha az derecede spesifik bir moleküler fonksiyona sahip olduğunu tespit ettiğimizi düşünelim. GO olmadan iki farklı detay seviyesinde moleküler fonksiyon bilgilerine eriştiğimiz genler arasında bir ilişki olup olmadığını görebilmek ciddi bir bilgi birikimi gerektirebilir. Oysa GO anotasyonları ve anotasyonlar arası hiyerarşik ilişkilere bakılarak söz konusu ilişkinin varlığını sınamak önemli ölçüde kolaylaşmaktadır.

Daha önce de belirtildiği üzere bu çalışma kapsamında elde edilen gen kümelerinin biyolojik anlamlılığı, genlerin moleküler fonksiyon ve biyolojik proses benzerliklerinden faydalanılarak değerlendirilmektedir. Anotasyonlar arası hiyerarşik ilişkileri de göz önüne alarak gen benzerliklerini analiz etmemize yardımcı olan “Term Enrichment” adlı Gen Ontolojisi’nin web üzerinde hizmet veren analiz aracı da bu bağlamda kullanılmış, elde edilen gen kümelerinin homojenitesi ve küme elemanlarının biyolojik ilişkileri sayısal olarak da ölçülmüştür. Gen Ontolojisi web sitesinde on-line olarak kullanıma sunulan “Term Enrichment” aracı, Stanford Üniversitesi araştırmacıları tarafından geliştirilen “GO-TermFinder” [10] adlı aracın web temelli versiyonudur.

“Term Enrichment” aracı, basit tanımı ile kendisine girdi olarak aktarılan gen listesinin elemanları ile ilişkili anotasyonlardan hareketle, tespit edilen her benzerlik için bir p-değeri üretir. Üretilen her p-değeri tespit edilen benzerliğin kayda değerliğini göstermektedir. p-değeri ne kadar küçük ise tespit edilen benzerlik de o derece kayda değer demektir. p-değeri hesaplanırken analize tabi tutulan gen listesinde yer alan genlerden kaçının söz konusu anotasyonu paylaştığı (anotasyonun grup içindeki sıklığı), tüm genom göz önünde bulundurulduğunda kaç genin söz konusu anotasyonu paylaştığı (yani anotasyonun tür içindeki genel yaygınlığı) gibi bilgiler de göz önünde bulundurularak elde edilen benzerliğin kayda değerlik seviyesi tespit edilmektedir [10]. “Term Enrichment” aracı bunun yanında “artalan frekansı” olarak adlandırılan ek bir bilgi ile, paylaşıldığı tespit edilen bir anotasyonun genom ölçeğindeki yaygınlık değerini ifade etmektedir. Artalan frekansı kesirli bir sayı olarak

ifade edilmekte olup, pay değeri tüm genomda kaç genin ilgili anotasyonu paylaştığına, payda değeri ise tüm genomda kaç genin bulunduğu bilgisine karşı düşmektedir. Artalan frekansı, bu tanımdan da anlaşılacağı gibi tespit edilen benzerliğe konu olan anotasyonun paylaşım yaygınlığını daha iyi görebilmeyi sağlamaktadır. Böylelikle küme elemanlarının biyolojik olarak ilişkileri daha iyi analiz edilebilmektedir. (Bu tez çalışması kapsamında Gen Ontolojisi'nin 17 Mayıs 2009 tarihli veritabanı sürümü ve araçları kullanılmıştır.)

5. ELDE EDİLEN SONUÇLAR VE ANALİZLER

Bu tez çalışması kapsamında elde edilen tüm sonuçlar ve sonuçların detaylı analizi bu bölümde yer almaktadır. Bu çalışma kapsamında geliştirilen metod ile amaçlanan, farklı genomik veri tiplerini gen kümeleme işleminde bir arada kullanarak yeni ve biyolojik olarak anlamlı gen kümeleri tespit etmek olduğu için, elde edilen sonuçların analizi, tek bir genomik veri tipinin (gen ekspresyon verisi) kullanılması ile elde edilen kümelerin, farklı tipte genomik verilerin (gen ekspresyon verisi, motif sıklık verisi ve gen konum verisi) bir arada kullanılması ile elde edilen kümeler ile karşılaştırılmasına dayanmaktadır. Öne sürülen hipotezin ve geliştirilen algoritmanın başarımlı değerlendirimi, farklı genomik veri tiplerinin bir arada kullanılması durumunda yeni ve biyolojik olarak anlamlı gen kümelerinin tespit edilip edilmediğine dayanmaktadır.

Bölüm 5.1’de ilk olarak sadece gen ekspresyon verisinin kullanılması ile elde edilen kümeleme sonuçları sunulmaktadır. Bölüm 5.2’de ise gen ekspresyon verisi, motif sıklık verisi ve gen konum verilerinin bir arada kullanılması ile elde edilen kümeleme sonuçları yer almaktadır. Her iki yöntem ile elde edilen kümelerin karşılaştırımı ve geliştirilen metodun başarımlı değerlendirmesi Bölüm 5.2’de yapılmaktadır.

5.1 Gen Ekspresyon Verisi ile Elde Edilen Kümeleme Sonuçları

Geliştirilen uygulama programı ilk olarak sadece gen ekspresyon verisinin kullanımı ile elde edilecek kümeleri tespit etmek için çalıştırılmıştır. Bu amaç ile gen ekspresyon veri tipi için ilgili ağırlık değeri “1” olarak, motif sıklık ve gen konum veri tipleri için ise ilgili ağırlık değerleri “0” olarak atanmıştır. Algoritma, elde edilecek küme adetlerindeki değişimi görebilmek amacı ile bir çok kez çalıştırılmıştır. Algoritma SOM tekniğine dayandığı için, SOM tekniğinin doğası gereği her çalıştırmada farklı sayıda küme ve yeni üyelik ilişkileri elde edilmektedir. Bunun temel nedeni, nöron ağırlıklarının ilk değerlerinin her eğitim fazının (yani çalıştırımın) başında rastgele atanmasıdır. Söz konusu ilk değerler her çalıştırmada değiştiği için, küme merkezleri de her çalıştırmada farklı nihai konumlara yakınsayabilmektedir. Geliştirilen program gen ekspresyon verisi için 10 kez çalıştırılmış ve elde edilen küme adetlerinin 19-21 aralığında değiştiği tespit edilmiştir (bkz. Çizelge 5.1). Bu sonuca dayanarak elde edilen küme adetlerinin farklı çalıştırmalarda büyük oranda değişmediği söylenebilir. İrdelenmesi gereken bir diğer önemli nokta ise genlerin elde edilen kümeler arasındaki dağılımıdır. Bu bağlamda tek bir genden oluşan küme sayısı önemli bir

kriterdir. Algoritma 10 kez çalıştırılmış ve tek elemanlı küme sayısının 1 ile 8 arasında değiştiği belirlenmiştir (bkz. Çizelge 5.1).

Çizelge 5.1 Gen ekspresyon verisi için farklı çalıştırmalar sonucu elde edilen toplam küme sayıları ve tek elemana sahip küme sayıları.

Çalıştırım	Toplam küme sayısı	Tek elemanlı küme sayısı
1	20	8
2	20	3
3	20	4
4	21	6
5	20	5
6	19	6
7	19	4
8	19	5
9	20	5
10	19	1
	Ortalama: 19,7	Ortalama: 4,7

Elde edilen kümeleme sonuçlarından birisi ayrıntılı biyolojik analiz ve değerlendirmelerde kullanılmak üzere seçilmiştir. Bu amaçla seçilen sonuç seti, 3. çalıştırmada elde edilen ve 4'ü tek elemanlı olmak üzere toplam 20 adet nihai kümeden oluşan sonuç setidir.

Algoritma SOM tekniğine dayandığı için elde edilen gen kümelerinin 2 boyutlu 5x5 SOM haritası üzerindeki nihai konumları son derece önemlidir. Çünkü SOM tekniğine göre komşu kümeler, birbirine en çok benzeyen kümelerdir. Söz konusu konumlara bağlı olarak oluşan komşuluklar, birbirleri ile ilişkili gen kümelerini işaret etmektedir. Bu nedenle SOM üzerinde yer alan her kümenin konumunu göz önünde bulundurmak önemlidir. Elde edilen kümelerin iki boyutlu harita üzerindeki konumları ve üye adetleri Şekil 5.1'de gösterildiği gibidir.

küme 0,0 2	küme 0,1 2	küme 0,2 1	küme 0,3 1	küme 0,4 8
küme 1,0 7	küme 1,1 4	küme 1,2 10	küme 1,3 6	küme 1,4 5
küme 2,0 3	küme 2,1 2	küme 2,2 1	küme 2,3 2	küme 2,4 3
küme 3,0 25	küme 3,1 7			
küme 4,0 8		küme 4,2 1		küme 4,4 2

Şekil 5.1 Gen ekspresyon verisi için elde edilen kümelerin ve küme eleman sayılarının 2 boyutlu harita üzerinde gösterimi.

Detaylı biyolojik değerlendirmeler için Gen Ontolojisi (GO) [2] ve TAIR [8] veritabanları kullanılarak her bir gene ait *moleküler fonksiyon* ve *biyolojik proses* bilgileri elde edilmiştir. Gen ekspresyon verisi kullanılarak elde edilmiş kümeler ve GO Term Enrichment aracı kullanılarak tespit edilmiş söz konusu kümeleri oluşturan genler arasındaki biyolojik proses ve moleküler fonksiyon benzerlikleri Çizelge 5.2’de gösterildiği gibidir.

Çizelge 5.2 Gen ekspresyon verisi için kümeleme sonuçları ve kayda değer gen benzerlikleri.

Küme	Biyolojik Proses	p-değeri & Artalan frekansı	Moleküler Fonksiyon	p-değeri & Artalan frekansı
Küme 0,0 (2 gen)	GO:0009628, response to abiotic stimulus (2 gen) GO:0006950, response to stress (2 gen)	5.57e-02 (1188/36620) 1.31e-01 (1822/36620)	Kayda değer bir benzerlik yok	
Küme 0,1 (2 gen)	Kayda değer bir benzerlik yok		Kayda değer bir benzerlik yok	
Küme 0,2 (1 gen)	Bilinmiyor		GO:0004722, has protein serine/threonine phosphatase activity	
Küme 0,3 (1 gen)	Bilinmiyor		GO:0004445, has inositol-polyphosphate 5-phosphatase activity	
Küme 0,4 (8 gen)	Kayda değer bir benzerlik yok		GO:0015171, amino acid transmembrane transporter activity (2 gen) GO:0005342, organic acid transmembrane transporter activity (2 gen)	9.33e-03 (54/36620) 1.23e-02 (62/36620)
Küme 1,0 (7 gen)	GO:0009628, response to abiotic stimulus (2 gen)	1.72e+00 (1188/36620)	GO:0003824, catalytic activity (4 gen)	1.98e+00 (6559/36620)
Küme 1,1 (4 gen)	Kayda değer bir benzerlik yok		Kayda değer bir benzerlik yok	
Küme 1,2 (10 gen)	GO:0006952, defense response (2 gen)	1.14e+00 (684/36620)	GO:0004857, enzyme inhibitor activity (2 gen)	3.09e-02 (109/36620)
Küme 1,3 (6 gen)	Kayda değer bir benzerlik yok		GO:0016740, transferase activity (3 gen) GO:0003824, catalytic activity (4 gen) GO:0016301, kinase activity (2 gen)	4.79e-01 (2435/36620) 1.08e+00 (6559/36620) 1.57e+00 (1274/36620)

Küme 1,4 (5 gen)	Kayda değer bir benzerlik yok		GO:0043169, cation binding (2 gen) GO:0046872, metal ion binding (2 gen)	3.22e-01 (786/36620) 3.79e-01 (855/36620)
Küme 2,0 (3 gen)	Kayda değer bir benzerlik yok		Kayda değer bir benzerlik yok	
Küme 2,1 (2 gen)	Kayda değer bir benzerlik yok		Kayda değer bir benzerlik yok	
Küme 2,2 (1 gen)	Bilinmiyor		GO:0004428, has inositol or phosphatidylinositol kinase activity	
Küme 2,3 (2 gen)	Kayda değer bir benzerlik yok		Kayda değer bir benzerlik yok	
Küme 2,4 (3 gen)	Kayda değer bir benzerlik yok		Kayda değer bir benzerlik yok	
Küme 3,0 (25 gen)	GO:0009887, organ morphogenesis (3 gen) GO:0035315, hair cell differentiation (2 gen) GO:0010026, trichome differentiation (2 gen) GO:0009965, leaf morphogenesis (2 gen) GO:0007169, transmembrane receptor protein tyrosine kinase signaling pathway (2 gen) GO:0009913, epidermal cell differentiation (2 gen) GO:0007398, ectoderm development (2 gen)	2.07e-01 (281/36620) 2.20e-01 (67/36620) 2.20e-01 (67/36620) 7.87e-01 (128/36620) 8.11e-01 (130/36620) 9.50e-01 (141/36620) 9.76e-01 (143/36620)	GO:0003824, catalytic activity (12 gen) GO:0016740, transferase activity (7 gen) GO:0016757, transferase activity, transferring glycosyl groups (3 gen) GO:0008194, UDP-glycosyltransferase activity (2 gen) GO:0019825, oxygen binding (2 gen)	1.25e-01 (6559/36620) 2.15e-01 (2435/36620) 7.19e-01 (435/36620) 1.66e+00 (188/36620) 2.46e+00 (231/36620)
Küme 3,1 (7 gen)	Kayda değer bir benzerlik yok		GO:0016787, hydrolase activity (2 gen)	4.55e+00 (2261/36620)
Küme 4,0 (8 gen)	GO:0009414, response to water deprivation (2 gen)	4.02e-02 (161/36620)	GO:0003700, transcription factor activity (3 gen)	3.34e-01 (1662/36620)

	GO:0031323, regulation of cellular metabolic process (3 gen)	2.00e-01 (1387/36620)	GO:0005488, binding (5 gen)	4.71e-01 (6520/36620)
	GO:0045449, regulation of transcription (2 gen)	2.14e+00 (1245/36620)	GO:0003677, DNA binding (3 gen)	6.32e-01 (2085/36620)
			GO:0005515, protein binding (2 gen)	3.02e+00 (1497/36620)
Küme 4,2 (1 gen)	GO:0009631, cold acclimation		GO:0003677, DNA binding	
	GO:0006355, regulation of transcription, DNA-dependent		GO:0000130, transcription factor activity	
Küme 4,4 (2 gen)	Kayda değer bir benzerlik yok		Kayda değer bir benzerlik yok	

Gen ekspresyon verisinin kullanımı ile elde edilen kümeler ve söz konusu kümeleri oluşturan genlere ait detaylı bilgiler, aşağıdaki çizelgelerde yer almaktadır. Her çizelge üç adet kolondan oluşmaktadır. İlk kolon, kümeyi oluşturan her bir genin ID bilgisini (söz konusu Arabidopsis geninin TAIR veritabanı ID'sini) göstermektedir. İkinci kolon söz konusu genin ilintili olduğu *biyolojik proses* bilgilerini, üçüncü kolon ise söz konusu gene ait *moleküler fonksiyon* bilgilerini göstermektedir. Küme adları/numaraları, kümelerin 5x5 SOM haritası üzerindeki konumunu gösterebilmek adına iki boyutlu koordinat notasyonu ile belirtilmiştir.

Çizelge 5.3 Küme 0,0'a ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 0,0		
AT1G29395	involved in cold acclimation, response to abscisic acid stimulus, cellular response to water deprivation	<Bilinmiyor>
AT1G16850	involved in response to salt stress	<Bilinmiyor>

Çizelge 5.4 Küme 0,1'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 0,1		
AT5G39410	involved in metabolic process	functions in binding; has catalytic activity
AT2G38465	<Bilinmiyor>	<Bilinmiyor>

Çizelge 5.5 Küme 0,2'ye ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 0,2		
AT5G27930	<Bilinmiyor>	has protein serine/threonine phosphatase activity

Çizelge 5.6 Küme 0,3'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 0,3		
AT1G71710	<Bilinmiyor>	has inositol-polyphosphate 5-phosphatase activity

Çizelge 5.7 Küme 0,4'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 0,4		
AT2G39130	<Bilinmiyor>	has amino acid transmembrane transporter activity
AT5G44620	<Bilinmiyor>	functions in oxygen binding
AT2G46620	<Bilinmiyor>	functions in ATP binding; has ATPase activity
AT2G37580	<Bilinmiyor>	functions in protein binding, zinc ion binding; has transferase activity, transferring glycosyl groups
AT2G21380	involved in microtubule-based movement	has microtubule motor activity
AT2G34650	involved in response to auxin stimulus, auxin mediated signaling pathway, auxin polar transport, root hair elongation, root hair initiation, cotyledon development, phyllome development	has kinase activity, protein kinase activity, protein serine/threonine kinase activity
AT4G36930	involved in response to cold, flower development, fruit development, response to red light, negative regulation of seed germination, carpel development	functions in DNA binding, has transcription factor activity
AT5G02180	involved in amino acid transport	has amino acid transmembrane transporter activity

Çizelge 5.8 Küme 1,0'a ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 1,0		
AT3G15357	<Bilinmiyor>	<Bilinmiyor>
AT5G66460	involved in carbohydrate metabolic process	has hydrolase activity, hydrolyzing O-glycosyl compounds
AT3G61890	involved in response to water deprivation, response to osmotic stress, response to salt stress, regulation of transcription, DNA-dependent, response to abscisic acid stimulus, positive regulation of transcription	has transcription factor activity, transcription activator activity
AT4G03430	involved in response to cold, response to abiotic stimulus	has RNA splicing factor activity, transesterification mechanism
AT2G17280	involved in metabolic process	has catalytic activity
AT3G09350	<Bilinmiyor>	functions in binding
AT5G15950	involved in spermidine biosynthetic process, spermine biosynthetic process; related to polyamine biosynthetic process	has adenosylmethionine decarboxylase activity

Çizelge 5.9 Küme 1,1'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 1,1		
AT1G24575	<Bilinmiyor>	<Bilinmiyor>
AT2G22080	involved in response to oxidative stress	<Bilinmiyor>
AT5G41580	<Bilinmiyor>	functions in zinc ion binding
AT5G40390	involved in sucrose biosynthetic process, mannitol biosynthetic process, raffinose family oligosaccharide biosynthetic process	has hydrolase activity, hydrolyzing O-glycosyl compounds, galactinol-sucrose galactosyltransferase activity

Çizelge 5.10 Küme 1,2'ye ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 1,2		
AT4G15420	involved in proteolysis, ubiquitin-dependent protein catabolic process	functions in zinc ion binding; has peptidase activity
AT5G55990	involved in calcium-mediated signaling	functions in calcium ion binding
AT1G34220	<Bilinmiyor>	<Bilinmiyor>
AT3G05130	<Bilinmiyor>	<Bilinmiyor>
AT3G07640	<Bilinmiyor>	<Bilinmiyor>
AT4G39090	involved in response to desiccation, response to water deprivation, response to osmotic stress, proteolysis, response to salt stress, defense response to bacterium	has cysteine-type endopeptidase activity, cysteine-type peptidase activity
AT2G02100	involved in defense response	has peptidase inhibitor activity

AT5G17050	involved in metabolic process	has UDP-glycosyltransferase activity, transferase activity, transferring glycosyl groups, anthocyanidin 3-O-glucosyltransferase activity, quercetin 3-O-glucosyltransferase activity
AT5G62350	<Bilinmiyor>	has pectinesterase inhibitor activity
AT4G22590	involved in trehalose biosynthetic process	has trehalose-phosphatase activity

Çizelge 5.11 Küme 1,3'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 1,3		
AT4G26270	involved in glycolysis	has 6-phosphofructokinase activity
AT3G60980	involved in response to oxidative stress	<Bilinmiyor>
AT4G11600	involved in response to salt stress, response to cadmium ion	has glutathione peroxidase activity
AT1G20030	involved in response to other organism	<Bilinmiyor>
AT3G55450	involved in protein amino acid phosphorylation	has kinase activity
AT3G12980	involved in protein amino acid acetylation, flower development	has H3/H4 histone acetyltransferase activity, histone acetyltransferase activity

Çizelge 5.12 Küme 1,4'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 1,4		
AT4G23040	<Bilinmiyor>	<Bilinmiyor>
AT5G48250	<Bilinmiyor>	functions in zinc ion binding; has transcription factor activity
AT5G26340	is downregulated by phosphorylation	has carbohydrate transmembrane transporter activity, hexose:hydrogen symporter activity, high-affinity hydrogen:glucose symporter activity, sugar:hydrogen symporter activity
AT5G51050	involved in transport	functions in calcium ion binding
AT2G22830	involved in sterol biosynthetic process	has oxidoreductase activity

Çizelge 5.13 Küme 2,0'a ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 2,0		
AT5G22250	involved in RNA modification	has ribonuclease activity
AT2G18050	involved in response to water deprivation, nucleosome assembly	functions in DNA binding, nucleosomal DNA binding
AT1G21050	<Bilinmiyor>	<Bilinmiyor>

Çizelge 5.14 Küme 2,1'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 2,1		
AT3G55430	involved in carbohydrate metabolic process	has hydrolase activity, hydrolyzing O-glycosyl compounds
AT3G54040	<Bilinmiyor>	<Bilinmiyor>

Çizelge 5.15 Küme 2,2'ye ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 2,2		
AT1G64460	<Bilinmiyor>	has inositol or phosphatidylinositol kinase activity

Çizelge 5.16 Küme 2,3'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 2,3		
AT4G34150	involved in response to cold	<Bilinmiyor>
AT2G29670	<Bilinmiyor>	functions in binding

Çizelge 5.17 Küme 2,4'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 2,4		
AT4G02350	involved in pollen germination, pollen tube growth	<Bilinmiyor>
AT1G34780	involved in cell redox homeostasis	<Bilinmiyor>
AT3G50310	involved in protein amino acid phosphorylation	has kinase activity

Çizelge 5.18 Küme 3,0'a ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 3,0		
AT1G28560	involved in snRNA transcription, organ morphogenesis	functions in DNA binding
AT1G15260	<Bilinmiyor>	<Bilinmiyor>
AT3G21790	involved in metabolic process	has UDP-glycosyltransferase activity
AT3G28040	involved in protein amino acid phosphorylation, transmembrane receptor	functions in ATP binding; has kinase activity, protein serine/threonine kinase

	protein tyrosine kinase signaling pathway	activity
AT3G14620	<Bilinmiyor>	functions in oxygen binding
AT5G55120	involved in L-ascorbic acid biosynthetic process, response to jasmonic acid stimulus, response to ozone	has galactose-1-phosphate guanylyltransferase (GDP) activity, GDP-D-glucose phosphorylase activity
AT1G63260	involved in aging	<Bilinmiyor>
AT3G26960	<Bilinmiyor>	<Bilinmiyor>
AT1G07350	involved in RNA splicing	functions in RNA binding
AT2G41940	involved in regulation of transcription, trichome differentiation	functions in nucleic acid binding, zinc ion binding; has transcription factor activity
AT1G29530	<Bilinmiyor>	<Bilinmiyor>
AT5G58770	involved in dolichol biosynthetic process	has dehydrodolichyl diphosphate synthase activity
AT1G12520	involved in cellular copper ion homeostasis	has superoxide dismutase activity, superoxide dismutase copper chaperone activity
AT3G26320	<Bilinmiyor>	functions in oxygen binding
AT5G49730	involved in response to light stimulus	does not have ferric-chelate reductase activity; has oxidoreductase activity, ferric-chelate reductase activity
AT4G23300	involved in protein amino acid phosphorylation	has kinase activity
AT2G45340	involved in protein amino acid phosphorylation, transmembrane receptor protein tyrosine kinase signaling pathway	functions in ATP binding; has protein serine/threonine kinase activity
AT1G55360	<Bilinmiyor>	<Bilinmiyor>
AT2G01670	<Bilinmiyor>	has hydrolase activity
AT1G67830	involved in lipid metabolic process	has alpha-L-fucosidase activity, carboxylesterase activity
AT2G46225	involved in actin nucleation, trichome morphogenesis	<Bilinmiyor>
AT3G47430	involved in peroxisome fission, peroxisome organization	<Bilinmiyor>
AT4G39770	involved in trehalose biosynthetic process	has trehalose-phosphatase activity
AT1G32900	involved in biosynthetic process, glucan biosynthetic process	has transferase activity, transferring glycosyl groups
AT4G14540	involved in regulation of transcription, DNA-dependent	has transcription factor activity

Çizelge 5.19 Küme 3,1'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 3,1		
AT1G08390	<Bilinmiyor>	functions in nucleic acid binding
AT1G75460	<Bilinmiyor>	has ATP-dependent peptidase activity
AT5G58375	<Bilinmiyor>	<Bilinmiyor>
AT2G26740	involved in response to water deprivation, response to auxin stimulus	has epoxide hydrolase activity
AT4G17670	<Bilinmiyor>	<Bilinmiyor>
AT5G28770	involved in regulation of transcription,	functions in DNA binding, has

	DNA-dependent	transcription factor activity, protein heterodimerization activity
AT3G13780	<Bilinmiyor>	<Bilinmiyor>

Çizelge 5.20 Küme 4,0'a ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 4,0		
AT2G25900	involved in regulation of transcription	has transcription factor activity
AT2G41010	involved in response to water deprivation, response to salt stress, regulation of salicylic acid metabolic process	functions in calmodulin binding
AT2G25250	<Bilinmiyor>	<Bilinmiyor>
AT4G25490	involved in cold acclimation, response to cold, response to water deprivation	functions in DNA binding; has transcription factor activity, transcription activator activity
AT4G33666	<Bilinmiyor>	<Bilinmiyor>
AT3G55980	involved in regulation of transcription, response to chitin	has transcription factor activity
AT4G29780	<Bilinmiyor>	<Bilinmiyor>
AT4G28270	<Bilinmiyor>	functions in protein binding, zinc ion binding

Çizelge 5.21 Küme 4,2'ye ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 4,2		
AT4G25470	involved in cold acclimation, response to cold, regulation of transcription, DNA-dependent	functions in DNA binding; has transcription factor activity, transcription activator activity

Çizelge 5.22 Küme 4,4'e ait genlerin biyolojik proses ve moleküler fonksiyon bilgileri.

Gen	Biyolojik Proses	Moleküler Fonksiyon
Küme 4,4		
AT3G16860	<Bilinmiyor>	<Bilinmiyor>
AT4G25480	involved in cold acclimation, response to cold, response to water deprivation, regulation of transcription, DNA-dependent	functions in DNA binding; has transcription factor activity, transcription activator activity

5.2 Gen Ekspresyon, Motif Sıklık ve Gen konum Verilerinin Birarada Kullanılması ile Elde Edilen Kümeleme Sonuçları

Gen ekspresyonu verisi, motif sıklık verisi ve gen konum verisinin bir arada kullanılması ile elde edilen kümeleme sonuçları bu bölümde yer almaktadır. Elde edilen sonuçların biyolojik kayda değerliği, dolayısı ile anlamlılığı da yine bu bölüm de irdelenmektedir.

Her üç genomik veri tipini kümeleme işleminde bir arada kullanmak için, geliştirilen uygulama programında ilgili ağırlık değerleri gen ekspresyon verisi için “0,4”, motif sıklık verisi için “0,3” ve gen konum verisi için ise “0,3” olacak şekilde atanmış ve program çalıştırılmıştır. Elde edilecek küme adetlerindeki değişimi görebilmek amacı ile, geliştirilen uygulama programı gen ekspresyon, motif sıklık ve gen konum verilerinin kombinasyonu için 10 kez çalıştırılmış ve elde edilen küme adetlerinin 16-22 aralığında değiştiği tespit edilmiştir (bkz. Çizelge 5.23). Bu tespit sayesinde, algoritmanın tutarlı sonuçlar ürettiği ve elde edilen küme adetlerinin farklı çalıştırmalarda büyük oranda değişmediği görülmüştür. Kombine veri seti için elde edilen küme adedi ortalaması 18,5 olup, gen ekspresyon verisi için elde edilen ortalama olan 19,7 değerinden daha azdır. Bir diğer önemli nokta ise genlerin elde edilen kümeler arasındaki dağılımıdır. Bu bağlamda tek bir genden oluşan küme sayısındaki değişim kombine genomik veri kullanılan bu bölümde de gözlemlenmiştir. Algoritma 10 kez çalıştırılmış ve tek elemanlı küme sayısının 1 ile 9 arasında değiştiği belirlenmiştir (bkz. Çizelge 5.23).

Çizelge 5.23 Kombine genomik veri (gen ekspresyon, motif sıklık, gen konum) için farklı çalıştırmalar sonucu elde edilen toplam küme sayıları ve tek elemana sahip küme sayıları.

Çalıştırım	Toplam küme sayısı	Tek elemanlı küme sayısı
1	21	5
2	18	4
3	20	1
4	18	6
5	19	5
6	18	5
7	16	4
8	16	3
9	22	9
10	17	3
	Ortalama: 18,5	Ortalama: 4,5

Elde edilen kümeleme sonuçlarından birisi ayrıntılı biyolojik analiz ve değerlendirmelerde kullanılmak üzere seçilmiştir. Bu amaçla seçilen sonuç seti, 6. çalıştırmada elde edilen ve 5'i tek elemanlı olmak üzere toplam 18 adet kümeden oluşan sonuç setidir.

Algoritma SOM tekniğine dayandığı için elde edilen kümelerin 2 boyutlu 5x5 SOM haritası üzerindeki nihai konumları da önem taşımaktadır. Önceki bölümlerde de belirtildiği gibi SOM haritası üzerinde yer alan komşu kümeler, birbirine benzer özellikte kümelerdir. Bu nedenle elde edilen her kümenin SOM haritası üzerindeki konumunu göz önünde bulundurmak ve komşu kümelerde yer alan genler arasındaki olası ilişkileri de irdelemek önemlidir. Şekil 5.2'de elde edilen her kümenin iki boyutlu harita üzerindeki konumları, üye adetleri ile birlikte gösterilmektedir.

küme 0,0 7	küme 0,1 21	küme 0,2 8	küme 0,3 4	küme 0,4 14
	küme 1,1 3	küme 1,2 5	küme 1,3 1	küme 1,4 2
		küme 2,2 2		küme 2,4 5
küme 3,0 1	küme 3,1 1		küme 3,3 3	küme 3,4 7
		küme 4,2 1	küme 4,3 1	küme 4,4 14

Şekil 5.2 Kombine genomik verinin (gen ekspresyon verisi, motif sıklık verisi, gen konum verisi) kullanımı ile elde edilen kümelerin ve küme eleman sayılarının 2 boyutlu harita üzerinde gösterimi.

Gen Ontolojisi [2] ve TAIR [8] veritabanları kullanılarak her bir kümede yer alan her bir gene ait *moleküler fonksiyon* ve *biyolojik proses* bilgileri elde edilmiştir. Söz konusu bilgiler, elde edilen kümelerin biyolojik anlamlılığının değerlendirilmesinde kullanılmıştır. Geliştirilen yöntem ile gen ekspresyon, motif sıklık ve gen konum verilerinin bir arada kullanılması sonucunda elde edilmiş kümeler, GO Term Enrichment aracı kullanılarak tespit edilmiş genler

arası biyolojik proses ve moleküler fonksiyon benzerlikleri ile birlikte Çizelge 5.24'te listelenmiştir.

Çizelge 5.24 Gen ekspresyon, motif sıklık ve gen konum verisinin bir arada kullanımı ile elde edilen kümeleme sonuçları ve kayda değer gen benzerlikleri.

Küme	Biyolojik Proses	p-değeri & Artalan frekansı	Moleküler Fonksiyon	p-değeri & Artalan frekansı
Küme 0,0 (7 gen)	GO:0009631, cold acclimation (2 gen)	7.83e-04 (23/36620)	GO:0016563, transcription activator activity (2 gen)	2.38e-02 (125/36620)
	GO:0050896, response to stimulus (5 gen)	1.30e-02 (3445/36620)	GO:0003677, DNA binding (3 gen)	5.37e-01 (2085/36620)
	GO:0009414, response to water deprivation (2 gen)	3.94e-02 (161/36620)	GO:0005488, binding (4 gen)	2.20e+00 (6520/36620)
	GO:0009628, response to abiotic stimulus (3 gen)	1.07e-01 (1188/36620)		
	GO:0019752, carboxylic acid metabolic process (2 gen)	4.64e-01 (562/36620)		
Küme 0,1 (21 gen)	GO:0007169, transmembrane receptor protein tyrosine kinase signaling pathway (2 gen)	4.85e-01 (130/36620)	GO:0004674, protein serine/threonine kinase activity (2 gen)	2.31e+00 (291/36620)
	GO:0050794, regulation of cellular process (5 gen)	1.46e+00 (2239/36620)	GO:0003700, transcription factor activity (4 gen)	2.63e+00 (1662/36620)
	GO:0006468, protein amino acid phosphorylation (2 gen)	2.07e+00 (275/36620)	GO:0005488, binding (8 gen)	4.38e+00 (6520/36620)
Küme 0,2 (8 gen)	GO:0016070, RNA metabolic process (2 gen)	1.35e+00 (1024/36620)	Kayda değer bir benzerlik yok.	
	GO:0009628, response to abiotic stimulus (2 gen)	1.78e+00 (1188/36620)		
Küme 0,3 (4 gen)	Kayda değer bir benzerlik yok.		Kayda değer bir benzerlik yok.	
Küme 0,4 (14 gen)	GO:0009414, response to water deprivation (2 gen)	2.23e-01 (161/36620)	GO:0016757, transferase activity, transferring glycosyl groups (2 gen)	1.54e+00 (435/36620)

	GO:0006950, response to stress (4 gen)	5.40e-01 (1822/36620)		
	GO:0009737, response to abscisic acid stimulus (2 gen)	6.31e-01 (274/36620)		
Küme 1,1 (3 gen)	Kayda değer bir benzerlik yok.		GO:0003824, catalytic activity (2 gen)	1.53e+00 (6559/36620)
Küme 1,2 (5 gen)	Kayda değer bir benzerlik yok.		GO:0046914, transition metal ion binding (2 gen)	4.21e-01 (642/36620)
Küme 1,3 (1 gen)	GO:0005975, carbohydrate metabolic process		GO:0004553, hydrolase activity, hydrolyzing O-glycosyl compounds	
Küme 1,4 (2 gen)	Kayda değer bir benzerlik yok.		GO:0003824, catalytic activity (2 gen)	7.38e-01 (6559/36620)
Küme 2,2 (2 gen)	Kayda değer bir benzerlik yok.		Kayda değer bir benzerlik yok.	
Küme 2,4 (5 gen)	Kayda değer bir benzerlik yok.		GO:0004857, enzyme inhibitor activity (2 gen)	2.05e-03 (109/36620)
			GO:0016791, phosphatase activity (2 gen)	1.09e-02 (252/36620)
Küme 3,0 (1 gen)	GO:0008295, spermidine biosynthetic process GO:0006597, spermine biosynthetic process		GO:0004014, adenosylmethionine decarboxylase activity	
Küme 3,1 (1 gen)	Bilinmiyor.		GO:0008270, zinc ion binding	
Küme 3,3 (3 gen)	Kayda değer bir benzerlik yok.		Kayda değer bir benzerlik yok.	
Küme 3,4 (7 gen)	Kayda değer bir benzerlik yok.		GO:0016301, kinase activity (2 gen)	1.97e+00 (1274/36620)
			GO:0003824, catalytic activity (4 gen)	1.98e+00 (6559/36620)
Küme 4,2 (1 gen)	GO:0008152, metabolic process		GO:0008194, UDP-glycosyltransferase activity GO:0016757, transferase activity, transferring glycosyl groups	

Küme 4,3 (1 gen)	Bilinmiyor.		Bilinmiyor.	
Küme 4,4 (14 gen)	GO:0003006, reproductive developmental process (3 gen)	9.16e-01 (896/36620)	GO:0015171, amino acid transmembrane transporter activity (2 gen)	4.05e-02 (54/36620)
	GO:0048608, reproductive structure development (3 gen)	9.16e-01 (896/36620)	GO:0022804, active transmembrane transporter activity (3 gen)	1.79e-01 (505/36620)
	GO:0009908, flower development (2 gen)	1.12e+00 (289/36620)	GO:0022891, substrate-specific transmembrane transporter activity (3 gen)	4.15e-01 (677/36620)

Bu tez çalışması kapsamında geliştirilen yöntem ile elde edilmiş gen kümelerinin detaylı analizi, takip eden bölümlerde ve elde edilen her gen kümesi için ayrı ayrı olmak üzere yapılmaktadır. Her bölümde söz konusu analizler bölüm kapsamında irdelenen kümeyi temsil eden bir çizelgeden faydalanılarak yapılmaktadır. Her çizelge dört adet kolondan oluşmaktadır. Daha önce de belirtildiği üzere bu tez çalışması kapsamında geliştirilen yöntemin başarımı, yöntemin uygulanması sonucunda yeni ve anlamlı gen kümeleri elde edilip edilmediği değerlendirilerek yapılmaktadır. Elde edilen kümelerin ve tespit edilen gen ilişkilerinin yeni olup olmadığını anlamak için ise, elde edilen kümelerde yer alan genlerin sadece gen ekspresyonu verisi kullanılarak kümelendiği bir önceki adımda hangi kümede yer aldıkları bilgisine ihtiyaç duyulmaktadır. Bu nedenle söz konusu çizelgelerin ilk kolonu, genlerin sadece gen ekspresyon verisinden faydalanılarak kümelendiği bir önceki bölümde (bkz. Bölüm 5.1) üyesi olduğu *eski küme* bilgisini göstermektedir. İkinci kolon, kümeyi oluşturan her bir genin ad/ID bilgisini (söz konusu Arabidopsis geninin TAIR veritabanı ID'sini) göstermektedir. Üçüncü kolon söz konusu genin ilintili olduğu *biyolojik proses* bilgilerini, dördüncü kolon ise söz konusu gene ait *moleküler fonksiyon* bilgilerini göstermektedir. Küme adları/numaraları, kümelerin 5x5 SOM haritası üzerindeki konumunu gösterebilmek adına iki boyutlu koordinat notasyonu ile belirtilmiştir.

5.2.1 Küme “0,0”a Ait Genler ve Analiz Sonuçları

“Küme 0,0”a ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.25’te gösterilmektedir. “Küme 0,0” 7 adet genden oluşmaktadır. Söz konusu 7 gen önceki adımda, yani sadece gen ekspresyon verisi kullanılarak yapılan kümeleme işleminde, 3 farklı küme ile ilişkilendirilmiştir. “Küme 0,0”ı oluşturan 7 genden 2’sinin biyolojik proses ve moleküler fonksiyon bilgileri bilinmemektedir. Geriye kalan 5 genin tamamının (AT5G55120, AT4G25490, AT3G55980, AT2G41010, AT4G25470) “response to stimulus, GO:0050896” adlı aynı biyolojik prosesle ilişkili olduğu, üçünün ise (AT4G25490, AT3G55980, AT4G25470) “DNA binding, GO:0003677” adlı moleküler fonksiyona sahip olduğu tespit edilmiştir.

Söz konusu 5 genden 2’sinin (AT2G41010, AT5G55120) “carboxylic acid metabolic process, GO:0019752” adlı daha spesifik bir biyolojik prosesle ilişkili olduğu tespit edilmiştir. Gen ekspresyon verisinin kullanıldığı önceki bölümde elde edilen kümeleme sonuçlarına göre bu 2 gen aynı kümede yer almamış dolayısıyla söz konusu ilişki ancak kombine verinin kullanılması ile tespit edilmiştir.

Söz konusu 5 genden bir diğer 2’si (AT4G25490, AT2G41010) daha spesifik bir biyolojik proses olan “response to water deprivation, GO:0009414” ile ilişkilidir. Ancak tespit edilen bu ilişki sadece gen ekspresyon verisinin kullanıldığı bir önceki adımda da tespit edildiği için yeni tespit edilmiş bir ilişki değildir. Yine söz konusu 5 genden 2’sinin (AT4G25490, AT4G25470) daha da spesifik bir biyolojik proses olan “cold acclimation, GO:0009631” prosesi ile ilişkili olduğu tespit edilmiştir. Bu iki genin aynı zamanda “transcription activator activity, GO:0016563” adlı aynı moleküler fonksiyona sahip olduğu tespit edilmiştir. Gen Ontolojisi veritabanına göre 36620 adet Arabidopsis gen ürününden (gene product) sadece 23’ünün “cold acclimation, GO:0009631” biyolojik prosesi ile ilişkili olduğu bilinmektedir. Buradan da anlaşılabileceği gibi “soğuğa uyum gösterme prosesi” olarak tanımlanan söz konusu proses son derece spesifik bir prosestir. Değerlendirme sırasında göz önünde bulundurulması gereken bir diğer önemli nokta ise kullanılan gen ekspresyon verisinin “soğuk şartları” altında elde edilmiş olmasıdır (bkz. Bölüm 4.2.1). Bu nedenle sadece gen ekspresyon verisinin kümeleme işlemi için kullanıldığı önceki adımda da söz konusu iki genin aynı kümede yer alacak şekilde kümeleneceğini ummak mantıklı bir beklentidir. Ancak sadece gen ekspresyon verisinin kullanıldığı önceki adımda söz konusu iki genin farklı kümelerde yer aldığı görülmektedir (bkz. Bölüm 5.1). Her iki genin de aynı kromozom üzerinde yer aldığı

(4. kromozom) ve dizilimsel konumlarının da birbirine yakın olduğu (AT4G25470 geninin baz çiftleri 13015287. ile 13016230. baz çiftleri arasında yer alırken, AT4G25490 genin baz çiftleri 13021790. ile 13022735. baz çiftleri arasında yer almaktadır) düşünülecek olursa, kombine verinin kullanılması ile birlikte gen konum verisinin söz konusu iki genin aynı kümede olacak şekilde kümelenmesinde önemli rol oynadığı söylenilebilir.

Yukarıda sıralanan tüm bu tespitlere dayanarak elde edilen bu kümenin sadece gen ekspresyon verisinin kullanılması ile tespit edilememiş ancak birbiriyle büyük oranda ilişkili genlerden oluşan, biyolojik olarak anlamlı yeni bir küme olduğu söylenilebilir.

Çizelge 5.25 Küme 0,0'a ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 0,0		
Küme 4,2	AT4G25470	involved in cold acclimation, response to cold, regulation of transcription, DNA-dependent	functions in DNA binding; has transcription factor activity, transcription activator activity
Küme 4,0	AT4G25490	involved in cold acclimation, response to cold, response to water deprivation	functions in DNA binding; has transcription factor activity, transcription activator activity
Küme 4,0	AT3G55980	involved in regulation of transcription, response to chitin	has transcription factor activity
Küme 4,0	AT4G29780	<Bilinmiyor>	<Bilinmiyor>
Küme 4,0	AT2G41010	involved in response to water deprivation, response to salt stress, regulation of salicylic acid metabolic process	functions in calmodulin binding
Küme 4,0	AT2G25250	<Bilinmiyor>	<Bilinmiyor>
Küme 3,0	AT5G55120	involved in L-ascorbic acid biosynthetic process, response to jasmonic acid stimulus, response to ozone	has galactose-1-phosphate guanylyltransferase (GDP) activity, GDP-D-glucose phosphorylase activity

5.2.2 Küme “0,1”e Ait Genler ve Analiz Sonuçları

“Küme 0,1”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.26’da gösterilmektedir. “Küme 0,1” 21 adet genden oluşmaktadır. Söz konusu 21 gen önceki adımda (gen ekspresyon verisi ile kümelemede) 4 farklı küme ile ilişkilendirilmiştir. “Küme 0,1”i oluşturan 21 genden 5’inin biyolojik proses ve moleküler fonksiyon bilgileri bilinmemektedir. Geriye kalan 16 genin 8’i (AT3G26320, AT4G25480, AT4G14540, AT4G28270, AT2G25900, AT2G45340, AT5G28770, AT3G28040) “binding, GO:0005488” adlı moleküler fonksiyona sahiptir. 16 genin 5’i ise (AT2G45340, AT4G25480, AT3G28040, AT5G28770, AT2G25900) “regulation of cellular process, GO:0050794” adlı aynı biyolojik prosesle ilgilidir. Söz konusu 5 genden 2’si ise (AT2G45340, AT3G28040) daha spesifik bir biyolojik proses olan “transmembrane receptor protein tyrosine kinase signaling pathway, GO:0007169” prosesi ile ilişkilidir ve bu iki gen aynı zamanda “protein serine/threonine kinase activity, GO:0004674” adlı aynı moleküler fonksiyona da sahiptir. Bununla beraber söz konusu iki gen, önceki bölümde de aynı küme içerisinde yer aldığı için bu ilişkinin yeni tespit edildiği söylenemez (bkz. Bölüm 5.1).

Yukarıda sıralanan ilişkiler, bu kümenin biyolojik olarak kayda değer olduğunu kesinlikle söyleyebileceğimiz derecede spesifik değildir (bkz. Çizelge 5.24, p-değerleri). Söz konusu kümenin 21 adet genden oluştuğunu düşünecek olursak, kümenin biyolojik olarak anlamlı olduğunu söyleyebilmek için küme üyeleri arasında daha fazla genin katılımı ile daha spesifik gen ilişkilerinin (ortak GO anotasyonlarının) bulunması gerektiğini söylemek yanlış olmayacaktır.

Bu kümeyle ilgili dikkat edilmesi gereken bir diğer önemli nokta ise “AT4G25480” genidir. Bu gen, “Küme 0,0” da yer alan “AT4G25470” ve “AT4G25490” genleri ile birlikte test veri setine özellikle dahil edilen, “cold acclimation, GO:0009631” adlı son derece spesifik biyolojik prosese dahil olan üç genden birisidir. Görüldüğü üzere bu gen, diğer iki genin üyesi olduğu “Küme 0,0”ın komşu kümesi olan “Küme 0,1”de yer almıştır. Daha önce de belirtildiği gibi SOM tekniğine göre kümeler arasındaki komşuluk ilişkileri, küme elamanları arasındaki benzerliklere göre belirlenmektedir. Gen ekspresyon, motif sıklık ve gen konum verilerinin bir arada kullanılması ile söz konusu üç gen, ikisi aynı kümede olmak üzere 2 komşu kümeyle üye olmuştur. Oysa sadece gen ekspresyon verisinin kullanıldığı önceki adımda her üç gen de ayrı ve birbirine komşu olmayan kümelere üye olmuştur (bkz. Bölüm 5.1.).

Çizelge 5.26 Küme 0,1'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 0,1		
Küme 3,0	AT2G45340	involved in protein amino acid phosphorylation, transmembrane receptor protein tyrosine kinase signaling pathway	functions in ATP binding; has protein serine/threonine kinase activity
Küme 3,0	AT2G46225	involved in actin nucleation, trichome morphogenesis	<Bilinmiyor>
Küme 4,4	AT3G16860	<Bilinmiyor>	<Bilinmiyor>
Küme 3,0	AT1G15260	<Bilinmiyor>	<Bilinmiyor>
Küme 4,0	AT2G25900	involved in regulation of transcription	has transcription factor activity
Küme 3,0	AT1G29530	<Bilinmiyor>	<Bilinmiyor>
Küme 3,1	AT5G28770	involved in regulation of transcription, DNA-dependent	functions in DNA binding, has transcription factor activity, protein heterodimerization activity
Küme 3,0	AT5G58770	involved in dolichol biosynthetic process	has dehydrololichyl diphosphate synthase activity
Küme 4,0	AT4G28270	<Bilinmiyor>	functions in protein binding, zinc ion binding
Küme 3,0	AT4G39770	involved in trehalose biosynthetic process	has trehalose-phosphatase activity
Küme 3,0	AT5G49730	involved in response to light stimulus	does not have ferric-chelate reductase activity; has oxidoreductase activity, ferric-chelate reductase activity
Küme 4,4	AT4G25480	involved in cold acclimation, response to cold, response to water deprivation, regulation of transcription, DNA-dependent	functions in DNA binding; has transcription factor activity, transcription activator activity
Küme 4,0	AT4G33666	<Bilinmiyor>	<Bilinmiyor>
Küme 3,0	AT1G63260	involved in aging	<Bilinmiyor>
Küme 3,0	AT3G47430	involved in peroxisome fission, peroxisome organization	<Bilinmiyor>
Küme 3,0	AT1G55360	<Bilinmiyor>	<Bilinmiyor>
Küme 3,0	AT3G28040	involved in protein amino acid phosphorylation, transmembrane receptor protein tyrosine kinase signaling pathway	functions in ATP binding; has kinase activity, protein serine/threonine kinase activity
Küme 3,0	AT4G14540	involved in regulation of transcription, DNA-dependent	has transcription factor activity
Küme 3,0	AT3G26320	<Bilinmiyor>	functions in oxygen binding
Küme 3,0	AT1G32900	involved in biosynthetic process, glucan biosynthetic process	has transferase activity, transferring glycosyl groups
Küme 3,0	AT2G01670	<Bilinmiyor>	has hydrolase activity

5.2.3 Küme “0,2”ye Ait Genler ve Analiz Sonuçları

“Küme 0,2”ye ait genler Çizelge 5.27’de gösterilmektedir. “Küme 0,2” 8 adet genden oluşmaktadır. Söz konusu 8 gen önceki adımda 3 farklı küme ile ilişkilendirilmiştir. “Küme 0,2”yi oluşturan 8 genlerden 2’sinin biyolojik proses ve moleküler fonksiyon bilgileri bilinmemektedir. Geriye kalan 6 genden 2’sinin (AT1G07350, AT1G28560) “RNA metabolic process, GO:0016070” adlı aynı biyolojik prosesle ilişkili olduğu tespit edilmiştir. Söz konusu iki gen, sadece gen ekspresyon verisinin kullanıldığı önceki adımda da aynı kümede yer almıştır. Bununla beraber, önceki adımda bu iki gen 23 farklı gen ile birlikte “Küme 3,0”da kümelenmiştir. Bu nedenle, söz konusu iki genin kombine genomik veri kullanılması durumunda daha homojen yapıda bir kümede yer aldıkları söylenebilir.

Hakkında bilgi sahibi olduğumuz 6 genden AT2G26740 ve AT4G34150 genlerinin ise “response to abiotic stimulus, GO:0009628” adlı biyolojik prosesle ilişkili olduğu tespit edilmiştir. Söz konusu iki gen önceki bölümde iki farklı kümede yer almıştır. Bu nedenle bu ilişkinin yeni tespit edilmiş bir ilişki olduğu söylenebilir.

Yukarıda belirtilen ortak anotasyonların Arabidopsis türündeki yaygınlığı (bkz. Çizelge 5.24 artalan frekansları) göz önünde bulundurulduğunda, elde edilen bu kümenin birbiriyle büyük oranda ilişkili genlerden oluştuğunu söylemek zordur.

Çizelge 5.27 Küme 0,2’ye ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 0,2		
Küme 3,1	AT2G26740	involved in response to water deprivation, response to auxin stimulus	has epoxide hydrolase activity
Küme 2,3	AT4G34150	involved in response to cold	<Bilinmiyor>
Küme 3,1	AT5G58375	<Bilinmiyor>	<Bilinmiyor>
Küme 3,0	AT3G26960	<Bilinmiyor>	<Bilinmiyor>
Küme 3,1	AT1G08390	<Bilinmiyor>	functions in nucleic acid binding
Küme 3,0	AT1G67830	involved in lipid metabolic process	has alpha-L-fucosidase activity, carboxylesterase activity
Küme 3,0	AT1G28560	involved in snRNA transcription, organ morphogenesis	functions in DNA binding
Küme 3,0	AT1G07350	involved in RNA splicing	functions in RNA binding

5.2.4 Küme “0,3”e Ait Genler ve Analiz Sonuçları

“Küme 0,3”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.28’de gösterilmektedir. “Küme 0,3” 4 adet genden oluşmaktadır. Söz konusu 4 gen önceki adımda 3 farklı küme ile ilişkilendirilmiştir. “Küme 0,3”ü oluşturan 4 genden 1’inin biyolojik proses ve moleküler fonksiyon bilgileri bilinmemektedir. Geriye kalan 3 genden ikisinin moleküler fonksiyon bilgileri, birinin ise ilişkili olduğu biyolojik proses bilgisi bilinmemektedir. Eldeki bilgiler ışığında söz konusu genler arasında biyolojik olarak kayda değer bir benzerlik bulunamamıştır.

Çizelge 5.28 Küme 0,3’e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 0,3		
Küme 2,4	AT4G02350	involved in pollen germination, pollen tube growth	<Bilinmiyor>
Küme 2,2	AT1G64460	<Bilinmiyor>	has inositol or phosphatidylinositol kinase activity
Küme 2,4	AT1G34780	involved in cell redox homeostasis	<Bilinmiyor>
Küme 2,1	AT3G54040	<Bilinmiyor>	<Bilinmiyor>

5.2.5 Küme “0,4”e Ait Genler ve Analiz Sonuçları

“Küme 0,4”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.29’da gösterilmektedir. “Küme 0,4” 14 adet genden oluşmaktadır. Söz konusu 14 gen önceki adımda 7 farklı küme ile ilişkilendirilmiştir. “Küme 0,4”ü oluşturan 14 genden 4’ünün biyolojik proses ve moleküler fonksiyon bilgileri bilinmemektedir. Geriye kalan 10 genden 4’ünün (AT2G22080, AT1G16850, AT1G29395, AT3G61890) “response to stress, GO:0006950” adlı biyolojik prosesle ilişkili olduğu tespit edilmiştir. Söz konusu 4 genden 2’sinin (AT1G29395, AT3G61890) daha spesifik bir biyolojik proses olan “response to water deprivation, GO:0009414” ile ilişkili olduğu tespit edilmiştir. Aynı iki genin “response to abscisic acid stimulus, GO:0009737” adlı bir başka biyolojik prosesle de ilişkili olduğu tespit edilmiştir. Söz konusu iki gen, gen ekspresyon verisi kullanılarak yapılan kümeleme işleminde aynı kümede yer almayıp, iki komşu kümede yer almıştır. Bu nedenle kombine genomik verinin bu ilişkiyi göstermede daha başarılı olduğu söylenilebilir.

Tespit edilen bu benzerlikler “Küme 0,4”ün kombine verinin kümeleme işleminde kullanılması ile elde edilmiş biyolojik olarak ilişkili genlere sahip yeni bir küme olduğunu göstermektedir.

Çizelge 5.29 Küme 0,4’e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 0,4		
Küme 0,0	AT1G16850	involved in response to salt stress	<Bilinmiyor>
Küme 0,1	AT5G39410	involved in metabolic process	functions in binding; has catalytic activity
Küme 1,2	AT3G05130	<Bilinmiyor>	<Bilinmiyor>
Küme 1,1	AT5G40390	involved in sucrose biosynthetic process, mannitol biosynthetic process, raffinose family oligosaccharide biosynthetic process	has hydrolase activity, hydrolyzing O-glycosyl compounds, galactinol-sucrose galactosyltransferase activity
Küme 3,1	AT3G13780	<Bilinmiyor>	<Bilinmiyor>
Küme 1,0	AT2G17280	involved in metabolic process	has catalytic activity
Küme 1,0	AT3G15357	<Bilinmiyor>	<Bilinmiyor>
Küme 1,0	AT3G61890	involved in response to water deprivation, response to osmotic stress, response to salt stress, regulation of transcription, DNA-dependent, response to abscisic acid stimulus, positive regulation of transcription	has transcription factor activity, transcription activator activity

Küme 1,0	AT3G09350	<Bilinmiyor>	functions in binding
Küme 1,1	AT2G22080	involved in response to oxidative stress	<Bilinmiyor>
Küme 3,0	AT3G21790	involved in metabolic process	has UDP-glycosyltransferase activity
Küme 0,0	AT1G29395	involved in cold acclimation, response to abscisic acid stimulus, cellular response to water deprivation	<Bilinmiyor>
Küme 3,0	AT3G14620	<Bilinmiyor>	functions in oxygen binding
Küme 1,2	AT3G07640	<Bilinmiyor>	<Bilinmiyor>

5.2.6 Küme “1,1”e Ait Genler ve Analiz Sonuçları

“Küme 1,1”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.30’da listelenmektedir. “Küme 1,1” 3 adet genden oluşmaktadır. Söz konusu 3 gen önceki adımda, yani sadece gen ekspresyon verisi kullanılarak yapılan kümeleme işleminde, 2 farklı küme ile ilişkilendirilmiştir. “Küme 1,1”i oluşturan 3 genden 1’inin biyolojik proses ve moleküler fonksiyon bilgileri bilinmemektedir. Geriye kalan 2 genin (AT1G75460, AT4G23300) “catalytic activity, GO:0003824” adlı aynı moleküler fonksiyona sahip oldukları tespit edilmiştir. Ancak bu benzerlik, söz konusu kümenin biyolojik olarak anlamlı olduğunu söyleyebilmemiz için yeterli spesifiklikte değildir.

Çizelge 5.30 Küme 1,1’e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 1,1		
Küme 3,0	AT4G23300	involved in protein amino acid phosphorylation	has kinase activity
Küme 3,1	AT1G75460	<Bilinmiyor>	has ATP-dependent peptidase activity
Küme 3,1	AT4G17670	<Bilinmiyor>	<Bilinmiyor>

5.2.7 Küme “1,2”ye Ait Genler ve Analiz Sonuçları

“Küme 1,2”ye ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.31’de gösterilmektedir. “Küme 1,2” 5 adet genden oluşmaktadır. Söz konusu 5 gen önceki adımda 2 komşu küme ile ilişkilendirilmiştir. “Küme 1,2”yi oluşturan 5 genden 1’inin biyolojik proses ve moleküler fonksiyon bilgileri bilinmemektedir. Geriye kalan 4 genden 2’sinin (AT2G41940, AT1G12520) “transition metal ion binding, GO:0046914” adlı aynı moleküler fonksiyona sahip oldukları tespit edilmiştir. Söz konusu 2 gen önceki bölümde de aynı kümede gruplanmıştır. Bununla beraber sadece gen ekspresyon verisinin kullanıldığı önceki bölümde söz konusu iki gen toplam 25 üyesi olan daha heterojen yapıdaki bir küme altında (Küme 3,0) gruplanmıştır. Bu nedenle, elde edilen bu yeni küme, söz konusu genler arasındaki ilişkiyi daha net gösterebilmektedir.

Çizelge 5.31 Küme 1,2’ye ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 1,2		
Küme 2,0	AT5G22250	involved in RNA modification	has ribonuclease activity
Küme 3,0	AT1G12520	involved in cellular copper ion homeostasis	has superoxide dismutase activity, superoxide dismutase copper chaperone activity
Küme 2,0	AT2G18050	involved in response to water deprivation, nucleosome assembly	functions in DNA binding, nucleosomal DNA binding
Küme 2,0	AT1G21050	<Bilinmiyor>	<Bilinmiyor>
Küme 3,0	AT2G41940	involved in regulation of transcription, trichome differentiation	functions in nucleic acid binding, zinc ion binding; has transcription factor activity

5.2.8 Küme “1,3”e Ait Genler ve Analiz Sonuçları

“Küme 1,3”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.32’de gösterilmektedir. “Küme 1,3” 1 adet genden oluşmaktadır. Söz konusu gen, sadece gen ekspresyon verisinin kullanıldığı önceki adımda kendi başına kümelenmeyip, bir diğer gen ile birlikte kümelenmiş ve söz konusu iki gen arasında bir benzerlik tespit edilememiştir (bkz. Bölüm 2.1). Elde edilen bu yeni küme tek bir genden oluştuğu için biyolojik doğruluğu hakkında yorum yapmak olanaklı değildir.

Çizelge 5.32 Küme 1,3’e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 1,3		
Küme 2,1	AT3G55430	involved in carbohydrate metabolic process	has hydrolase activity, hydrolyzing O-glycosyl compounds

5.2.9 Küme “1,4”e Ait Genler ve Analiz Sonuçları

“Küme 1,4”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.33’te gösterilmektedir. “Küme 1,4” 2 adet genden oluşmaktadır. Söz konusu 2 gen önceki adımda da aynı küme ile ilişkilendirilmiştir. “Küme 1,4”ü oluşturan 2 genin de “catalytic activity, GO:0003824” adlı aynı moleküler fonksiyona sahip olduğu tespit edilmiştir. Gen ekspresyon verisinin kullanıldığı önceki adımda da bu ilişki tespit edilmiştir. Üstelik bu iki gen önceki adımda aynı moleküler fonksiyona sahip 2 farklı geni de kapsayan bir kümede gruplanmıştır (bkz. Bölüm 5.1). Dolayısıyla söz konusu ilişki önceki adımda daha net bir biçimde tespit edilmiştir. Bu nedenlerle “Küme 1,4” daha önce tespit edilememiş yeni bir ilişkiyi göstermemektedir.

Çizelge 5.33 Küme 1,4'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 1,4		
Küme 1,0	AT5G66460	involved in carbohydrate metabolic process	has hydrolase activity, hydrolyzing O-glycosyl compounds
Küme 1,0	AT4G03430	involved in response to cold, response to abiotic stimulus	has RNA splicing factor activity, transesterification mechanism

5.2.10 Küme “2,2”ye Ait Genler ve Analiz Sonuçları

“Küme 2,2”ye ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.34’te gösterilmektedir. “Küme 2,2” 2 adet genden oluşmaktadır. Söz konusu 2 gen önceki adımda 2 komşu küme ile ilişkilendirilmiştir. “Küme 2,2”yi oluşturan 2 gen arasında biyolojik bir benzerlik tespit edilememiştir.

Çizelge 5.34 Küme 2,2’ye ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 2,2		
Küme 2,4	AT3G50310	involved in protein amino acid phosphorylation	has kinase activity
Küme 2,3	AT2G29670	<Bilinmiyor>	functions in binding

5.2.11 Küme “2,4”e Ait Genler ve Analiz Sonuçları

“Küme 2,4”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.35’te gösterilmektedir. “Küme 2,4” 5 adet genden oluşmaktadır. Söz konusu 5 gen önceki adımda 2 komşu küme ile ilişkilendirilmiştir. “Küme 2,4”ü oluşturan 5 genden 2’sinin (AT2G02100, AT5G62350) “enzyme inhibitor activity, GO:0004857” adlı moleküler fonksiyona sahip olduğu tespit edilmiştir. Söz konusu iki gen, sadece gen ekspresyon verisinin kümelemede kullanılması ile de aynı kümede gruplanmıştır. Ancak önceki adımda bu iki gen 10 elemanlı bir kümede yer alırken, kombine verinin kullanıldığı bu adımda 5

genden oluşan bir kümede yer almıştır. Bu nedenle kombine verinin kullanılması ile söz konusu ilişkinin daha homojen bir küme ile temsil edildiği söylenebilir.

“Küme 2,4”ü oluşturan 5 genden bir diğer 2’sinin ise (AT4G22590, AT5G27930) “phosphatase activity, GO:0016791” adlı aynı moleküler fonksiyona sahip olduğu tespit edilmiştir. Söz konusu iki gen önceki adımda iki komşu kümede yer almıştır. Bu nedenle, bu ilişkinin kombine genomik verinin kullanılması ile tespit edilebildiği ve “Küme 2,4”ün biyolojik olarak anlamlı yeni bir küme olduğu söylenebilir.

Çizelge 5.35 Küme 2,4’e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 2,4		
Küme 1,2	AT4G15420	involved in proteolysis, ubiquitin-dependent protein catabolic process	functions in zinc ion binding; has peptidase activity
Küme 1,2	AT4G22590	involved in trehalose biosynthetic process	has trehalose-phosphatase activity
Küme 1,2	AT2G02100	involved in defense response	has peptidase inhibitor activity
Küme 1,2	AT5G62350	<Bilinmiyor>	has pectinesterase inhibitor activity
Küme 0,2	AT5G27930	<Bilinmiyor>	has protein serine/threonine phosphatase activity

5.2.12 Küme “3,0”a Ait Genler ve Analiz Sonuçları

“Küme 3,0”a ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.36’da gösterilmektedir. “Küme 3,0” 1 adet genden oluşmaktadır. Söz konusu gen, sadece gen ekspresyon verisinin kullanıldığı önceki adımda kendi başına kümelenmemiştir. Elde edilen bu yeni küme tek bir genden oluştuğu için biyolojik doğruluğu hakkında yorum yapmak olanaklı değildir.

Çizelge 5.36 Küme 3,0’a ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 3,0		
Küme 1,0	AT5G15950	involved in spermidine biosynthetic process, spermine biosynthetic process; related to polyamine biosynthetic process	has adenosylmethionine decarboxylase activity

5.2.13 Küme “3,1”e Ait Genler ve Analiz Sonuçları

“Küme 3,1”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.37’de gösterilmektedir. “Küme 3,1” 1 adet genden oluşmaktadır. Söz konusu gen, sadece gen ekspresyon verisinin kullanıldığı önceki adımda kendi başına kümelenmemiştir. Elde edilen bu yeni küme tek bir genden oluştuğu için biyolojik doğruluğu hakkında yorum yapmak olanaklı değildir.

Çizelge 5.37 Küme 3,1’e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 3,1		
Küme 1,1	AT5G41580	<Bilinmiyor>	functions in zinc ion binding

5.2.14 Küme “3,3”e Ait Genler ve Analiz Sonuçları

“Küme 3,3”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.38’de gösterilmektedir. “Küme 3,3” 3 adet genden oluşmaktadır. Söz konusu 3 gen önceki adımda 2 komşu küme ile ilişkilendirilmiştir. “Küme 3,3”ü oluşturan 3 genden 1’inin biyolojik proses ve moleküler fonksiyon bilgileri bilinmemektedir. Geriye kalan 2 genin sahip oldukları moleküler fonksiyon ve ilişkili oldukları biyolojik prosesler arasında bir benzerlik bulunamamıştır.

Çizelge 5.38 Küme 3,3’e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 3,3		
Küme 1,1	AT1G24575	<Bilinmiyor>	<Bilinmiyor>
Küme 1,2	AT5G55990	involved in calcium-mediated signaling	functions in calcium ion binding
Küme 1,2	AT4G39090	involved in response to desiccation, response to water deprivation, response to osmotic stress, proteolysis, response to salt stress, defense response to bacterium	has cysteine-type endopeptidase activity, cysteine-type peptidase activity

5.2.15 Küme “3,4”e Ait Genler ve Analiz Sonuçları

“Küme 3,4”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.39’da gösterilmektedir. “Küme 3,4” 7 adet genden oluşmaktadır. Söz konusu 7 gen önceki adımda 4 farklı küme ile ilişkilendirilmiştir. “Küme 3,4”ü oluşturan 7 genden 1’inin biyolojik proses ve moleküler fonksiyon bilgileri bilinmemektedir. Geriye kalan 6 genden 4’ünün (AT1G71710, AT4G11600, AT4G26270, AT3G55450) “catalytic activity, GO:0003824” moleküler fonksiyonuna sahip olduğu tespit edilmiştir. Söz konusu 4 genden 2’sinin ise (AT4G26270, AT3G55450) “kinase activity, GO:0016301” adlı daha spesifik moleküler fonksiyona sahip olduğu tespit edilmiştir. Yine de bu ilişki, söz konusu kümenin biyolojik olarak kayda değer olduğunu söyleyebileceğimiz spesifiklikte değildir. Bununla beraber söz konusu 2 gen önceki adımda da aynı kümede yer almıştır. Önceki adımda elde edilen bu kümenin eleman sayısı 6 olup, “Küme 3,4” ile homojeniteleri birbirine çok yakındır. Bu nedenle “Küme 3,4”ün daha önce tespit edilememiş yeni ve kayda değer ilişkileri içerdiğini söylemek zordur.

Çizelge 5.39 Küme 3,4’e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 3,4		
Küme 1,3	AT4G26270	involved in glycolysis	has 6-phosphofructokinase activity
Küme 1,3	AT3G60980	involved in response to oxidative stress	<Bilinmiyor>
Küme 1,4	AT5G51050	involved in transport	functions in calcium ion binding
Küme 1,3	AT4G11600	involved in response to salt stress, response to cadmium ion	has glutathione peroxidase activity
Küme 0,3	AT1G71710	<Bilinmiyor>	has inositol-polyphosphate 5-phosphatase activity
Küme 1,3	AT3G55450	involved in protein amino acid phosphorylation	has kinase activity
Küme 1,2	AT1G34220	<Bilinmiyor>	<Bilinmiyor>

5.2.16 Küme “4,2”ye Ait Genler ve Analiz Sonuçları

“Küme 4,2”ye ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.40’ta gösterilmektedir. “Küme 4,2” 1 adet genden oluşmaktadır. Söz konusu gen, sadece gen ekspresyon verisinin kullanıldığı önceki adımda kendi başına kümelenmemiştir. Elde

edilen bu yeni küme tek bir genden oluştuğu için biyolojik doğruluğu hakkında yorum yapmak olanaklı değildir.

Çizelge 5.40 Küme 4,2'ye ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 4,2		
Küme 1,2	AT5G17050	involved in metabolic process	has UDP-glycosyltransferase activity, transferase activity, transferring glycosyl groups, anthocyanidin 3-O-glucosyltransferase activity, quercetin 3-O-glucosyltransferase activity

5.2.17 Küme “4,3”e Ait Genler ve Analiz Sonuçları

“Küme 4,3”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.41’de gösterilmektedir. “Küme 4,3” 1 adet genden oluşmaktadır. Söz konusu gen, sadece gen ekspresyon verisinin kullanıldığı önceki adımda bir başka gen ile birlikte kümelendi. Elde edilen bu yeni küme tek bir genden oluştuğu için ve söz konusu genin ilişkili olduğu biyolojik proses ve sahip olduğu moleküler fonksiyon bilinmediği için, bu kümenin biyolojik doğruluğu hakkında yorum yapmak mümkün değildir.

Çizelge 5.41 Küme 4,3’e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 4,3		
Küme 0,1	AT2G38465	<Bilinmiyor>	<Bilinmiyor>

5.2.18 Küme “4,4”e Ait Genler ve Analiz Sonuçları

“Küme 4,4”e ait genler, biyolojik proses ve moleküler fonksiyon bilgileriyle birlikte Çizelge 5.42’de gösterilmektedir. “Küme 4,4” 14 adet genden oluşmaktadır. Söz konusu 14 gen, kümeleme işlemi için sadece gen ekspresyon verisinin kullanıldığı önceki adımda 3 farklı küme ile ilişkilendirilmiştir. “Küme 4,4”ü oluşturan 14 genden 1’inin biyolojik proses ve moleküler fonksiyon bilgileri bilinmemektedir. Geriye kalan 13 genden 3’ünün (AT2G34650, AT4G36930, AT3G12980) “reproductive developmental process, GO:0003006” ve “reproductive structure development, GO:0048608” adlı biyolojik prosesler ile ilişkili olduğu tespit edilmiştir. Söz konusu 3 genden 2’sinin (AT4G36930, AT3G12980) ise daha spesifik bir biyolojik proses olan “flower development, GO:0009908” ile ilişkili olduğu tespit edilmiştir. Söz konusu 3 gen, gen ekspresyon verisi ile yapılan bir önceki kümeleme sonucunda 2 farklı kümeye yerleştirilmiştir.

“Küme 4,4” ü oluşturan ve hakkında bilgi sahibi olduğumuz 13 genden bir diğer 3’ünün (AT2G39130, AT5G02180, AT5G26340) “substrate-specific transmembrane transporter activity, GO:0022891” adlı aynı moleküler fonksiyona sahip olduğu tespit edilmiştir. Söz konusu üç gen aynı zamanda “active transmembrane transporter activity, GO:0022804” adlı daha spesifik bir moleküler fonksiyona da sahiptir. Söz konusu 3 gen bir önceki adımda 2 komşu kümeye yerleştirilmiştir (bkz. Bölüm 5.1). Bu nedenle bu ilişki ancak kombine verinin kümelemede kullanılması ile tespit edilmiştir. Bununla beraber söz konusu 3 genden 2’sinin (AT2G39130, AT5G02180) daha da spesifik bir moleküler fonksiyon olan “amino acid transmembrane transporter activity, GO:0015171” adlı fonksiyona sahip olduğu tespit edilmiştir. Ancak söz konusu iki gen, sadece gen ekspresyon verisinin kullanıldığı bir önceki adımda da aynı kümeye yerleştirilmiştir. Bu nedenle iki gen arasındaki bu ilişkinin yeni tespit edildiği söylenemez.

Yukarıda sıralanan ilişkilerden görüleceği üzere, Küme “4,4” sadece gen ekspresyon verisinin kümeleme işleminde kullanılması durumunda tespit edilememiş ancak kayda değer ilişkileri içeren, biyolojik olarak anlamlı ve yeni tespit edilmiş bir kümedir.

Çizelge 5.42 Küme 4,4'e ait genlerin biyolojik proses, moleküler fonksiyon bilgileri ve gen ekspresyon verisi kullanılarak yapılan bir önceki kümeleme işlemi sonucunda üyesi oldukları tespit edilen eski küme bilgileri.

Eski Küme	Gen	Biyolojik Proses	Moleküler Fonksiyon
	Küme 4,4		
Küme 0,4	AT2G37580	<Bilinmiyor>	functions in protein binding, zinc ion binding; has transferase activity, transferring glycosyl groups
Küme 1,4	AT2G22830	involved in sterol biosynthetic process	has oxidoreductase activity
Küme 0,4	AT5G02180	involved in amino acid transport	has amino acid transmembrane transporter activity
Küme 0,4	AT4G36930	involved in response to cold, flower development, fruit development, response to red light, negative regulation of seed germination, carpel development	functions in DNA binding, has transcription factor activity
Küme 1,4	AT5G48250	<Bilinmiyor>	functions in zinc ion binding; has transcription factor activity
Küme 1,4	AT4G23040	<Bilinmiyor>	<Bilinmiyor>
Küme 0,4	AT2G46620	<Bilinmiyor>	functions in ATP binding; has ATPase activity
Küme 1,4	AT5G26340	is downregulated by phosphorylation	has carbohydrate transmembrane transporter activity, hexose:hydrogen symporter activity, high-affinity hydrogen:glucose symporter activity, sugar:hydrogen symporter activity
Küme 1,3	AT3G12980	involved in protein amino acid acetylation, flower development	has H3/H4 histone acetyltransferase activity, histone acetyltransferase activity
Küme 0,4	AT2G34650	involved in response to auxin stimulus, auxin mediated signaling pathway, auxin polar transport, root hair elongation, root hair initiation, cotyledon development, phyllome development	has kinase activity, protein kinase activity, protein serine/threonine kinase activity
Küme 0,4	AT2G39130	<Bilinmiyor>	has amino acid transmembrane transporter activity
Küme 0,4	AT2G21380	involved in microtubule-based movement	has microtubule motor activity
Küme 0,4	AT5G44620	<Bilinmiyor>	functions in oxygen binding
Küme 1,3	AT1G20030	involved in response to other organism	<Bilinmiyor>

6. SONUÇLAR VE ÖNERİLER

Bu tez çalışması kapsamında geliştirilen yöntemin amacı, tek bir genomik veri tipinin gen kümeleme işleminde girdi olarak kullanılması ile elde edilemeyecek, ancak biyolojik olarak anlamlı olan yeni gen kümelerini, farklı genomik veri tiplerinin birarada kullanımı ile tespit etmektir. Bu amaç ile Kasturi ve diğerleri (2005) tarafından önerilen kümeleme algoritması referans olarak alınmıştır.

Kasturi ve diğerleri (2005) tarafından sunulan algoritma, SOM yapay sinir ağı modelini ve bu modele ait öğrenim algoritmasını genişleterek genomik bilgi füzyonunu mümkün kılmaktadır. Böylelikle genlere ait farklı genomik veriler bir arada kullanılarak gen kümeleme işlemi gerçekleştirilebilmektedir. Kasturi ve diğerleri (2005) tarafından sunulan bu yöntemde iki farklı genomik veri tipi bir arada kullanılmıştır. Bunlardan ilki literatürde gen kümeleme işlemi için en çok kullanılan ve bu nedenle de gen kümeleme işleminin temel veri tipi olan gen ekspresyonu verisi iken, ikinci veri tipi ise motif sıklık verisidir.

Bu tez çalışması kapsamında söz konusu metod, yukarıda sıralanan iki farklı genomik veri tipine ek olarak bir üçüncü genomik veri tipini de kullanarak kümeleme işlemini gerçekleştirebilecek şekilde genişletilmiştir. Bu amaç ile kullanılan üçüncü genomik veri tipi gen konum verisidir. Kasturi ve diğerleri (2005) tarafından sunulan algoritma ayrıca farklı uzaklık fonksiyonlarını kullanacak şekilde de genişletilmiştir. Bu amaç ile gen ekspresyonu veri tipi için vektör-açı-uzaklığı, üçüncü veri tipi olarak kullanılan gen konum verisi için ise GLDist adı verilen ve bu çalışma kapsamında geliştirilen yeni uzaklık fonksiyonu kullanılmıştır.

Arabidopsis türüne ait toplam 100 adet gen, test veri setini hazırlamak için kullanılmıştır. Soğuk etkisi altında elde edilmiş yedi farklı zaman noktasından oluşan gen ekspresyonu verisi, 614 farklı motifin her gende kaç defa tekrar ettiği bilgisinden oluşan motif sıklık verisi ve her genin kromozomal konumunu belirten gen konum verisi bir araya getirilerek test veri seti oluşturulmuştur. Böylelikle, toplam 100 adet satırdan (yani genden) ve 624 adet kolondan (yani öznelikten) oluşan test veri seti elde edilmiştir.

Geliştirilen yöntemin başarımının ve kombine genomik verinin kullanımı ile tespit edilen yeni kümelerin biyolojik kayda değerliğinin değerlendirilmesi için, Gen Ontolojisi ve TAIR veritabanları kullanılarak her genin literatüre geçmiş moleküler fonksiyon ve biyolojik proses

bilgileri tespit edilmiştir. Kümeleme sonuçlarının değerlendirildiği fazda ilk olarak, tek bir genomik veri tipinin, yani gen ekspresyon verisinin kümeleme işleminde kullanılması ile elde edilen kümeler tespit edilmiştir. Daha sonra ise, kombine genomik verinin, yani gen ekspresyon, motif sıklık ve gen konum verilerinin bir arada kullanılması ile elde edilen kümeler tespit edilmiştir. Bu iki adımdan sonra, kombine verinin kullanılması ile elde edilmiş “yeni” kümeler belirlenmiş ve elde edilen bu yeni gen kümeleri Gene Ontolojisi’nden elde edilen anotasyonlar (moleküler fonksiyon ve biyolojik proses bilgileri) baz alınarak detaylı bir şekilde analiz edilmiştir.

Elde edilen yeni kümelerin başarımı, biyolojik anlamlılıkları değerlendirilerek ölçülmüştür. Veri seti üç farklı genomik veri tipinden oluştuğu ve her veri tipinin kendisine özgü karakteristik özellikleri olduğu için, elde edilen kümelerin başarımlarının istatistiksel yöntemlerle (örneğin kompaktlık ölçümü gibi) değerlendirilmesi mümkün değildir. Üç farklı veri tipinin bir araya getirilmesi ile elde edilen söz konusu test veri seti biyolojik olarak birbirini tamamlayıcı bilgilerden oluşmuş olsa da, farklı veri tiplerini teşkil eden kolonlar arasındaki matematiksel korelasyon, tek bir veri tipinin kolonları arasındaki korelasyondan daha düşüktür. Bu nedenle, bu çalışmanın sonuçlarının istatistiksel yöntemlerle değerlendirilmesi uygun bir yaklaşım değildir.

Söz konusu kümelerin detaylı analizi ve yapılan biyolojik değerlendirmeler sonucunda, kombine verinin geliştirilen yöntem ile birlikte kümeleme işleminde kullanılması sonucunda yeni ve biyolojik olarak anlamlı gen kümelerinin tespit edildiği görülmüştür. Özellikle Küme “0,0”, Küme “0,1”, Küme “0,4”, Küme “2,4” ve Küme “4,4” bu anlamda elde edilmiş kayda değer kümelerdendir. Yapılan bu tespitlere dayanarak, biyolojik olarak anlamlı yeni gen kümeleri elde etmek için, kombine genomik verinin bu çalışma kapsamında sunulan yöntem ile birlikte kullanılabileceği gösterilmiştir.

Uzaklık fonksiyonu seçiminin elde edilecek kümeleme sonuçlarını doğrudan etkilediği gerçeğinden hareketle, bu çalışmayı takip edecek olası bir çalışma kapsamında genomik veri tipi-uzaklık fonksiyonu çiftleri üzerine yoğunlaşarak, kullanılan genomik veri tiplerine uyumlu olabilecek farklı kümeleme metriklerinin kullanılması durumunda ne gibi sonuçlar elde edileceği değerlendirilebilir. Bunun da ötesinde her zaman daha iyi uzaklık (veya yakınlık) metriklerine ihtiyaç duyulduğu gerçeğinden hareketle, kullanılan genomik veri tiplerinin karakteristiğine uygun olabilecek yeni ve daha iyi kümeleme metriklerinin bulunmasına yönelik çalışmalar da devam çalışması olarak yürütülebilir. Benzer şekilde farklı

komşuluk topolojilerinin bu tez çalışması kapsamında geliştirilen yöntem üzerindeki etkileri de bir devam çalışması ile detaylı şekilde analiz edilebilir.

SOM'ların kümeleme için kullanılması durumunda SOM tekniğinin doğası gereği her çalıştırma sonrasında farklı sayıda küme ve yeni gen-küme ilişkileri elde edilebilmektedir. Bunun temel nedeni, nöron ağırlıklarının ilk değerlerinin eğitim safhasının başlangıcında rastgele atanmasıdır. Söz konusu ilk değerler her çalıştırmada değiştiği için nihai küme merkezleri de her çalıştırmada farklı konumlara yakınsayabilmektedir. Bu nedenle bu tez çalışmasında geliştirilen algoritma kullanılarak elde edilecek küme adetleri her çalıştırmada değişebilmektedir. Sonuçların değerlendirilmesi açısından bakıldığında SOM tekniğinin bu karakteristiğinin avantaj ve dezavantajları beraberinde getirdiği söylenilebilir. Amaç biyolojik olarak anlamlı yeni gen kümeleri elde etmek olduğu için, her çalıştırmada biyolojik olarak ilişkili genlerden oluşan yeni kümeler elde edilebilecek olunması bir avantajdır. Öte yandan bu çeşitlilik, elde edilen kümelerin biyolojik anlamlılığının değerlendirilmesi fazında karmaşıklığı arttırmaktadır. Çünkü önerilen metodun yüzlerce, binlerce genden oluşan veri setleri için kullanılması durumunda, sonuç uzayı son derece büyük boyutlara ulaşabilecektir. Bu dezavantajın önüne geçebilmek için nöron ağırlık ilk değerlerinin rastgele verilmeyip, akıllı bir yöntem ile verilmesi üzerinde bir devam çalışması yürütülebilir. Böyle bir çalışma ile nöron ilk değerlerinin sabit olarak belirlenmesi gerçekleştirilebilir ise, elde edilen küme adetleri ve küme-gen ilişkileri de her çalıştırma sonucunda değişmeyecektir. Ancak belirtildiği gibi bu noktada da dikkatli olunmalıdır. Çünkü sabit ilk değerlerin atanması, rastgele değerler atanması durumunda elde edilebilecek olası farklı ve anlamlı gen kümelerinin elde edilememesine yol açabilecektir.

KAYNAKLAR

- Azuaje, F., Dopazo, J. (2005) *Data Analysis and Visualization in Genomics and Proteomics*. Wiley.
- Baldi, P., Brunak, S. (2001) *Bioinformatics: The Machine Learning Approach*. The MIT Press, England.
- Barash, Y., Frieichman, N. (2002) Context-specific Bayesian clustering for gene expression data. *Journal of Computational Biology*, Vol. 9, pages 169-191.
- Basset, D.E., Eisen, M.B., Boguski, M.S. (1999) Gene expression informatics - it's all in your mine. *Nature Genetics Supplement*, Vol. 21, pages 51-55.
- Ben-Dor, A., Shamir, R., Yakhini, Z. (1999) Clustering gene expression patterns. *Journal of Computational Biology*, Vol. 6, pages 281-297.
- Berry, M., Linoff, G. (2004) *Data Mining Techniques*. Wiley Publishing.
- Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.*, 97, pages 262-267.
- D'haeseleer, P., Liang, S., Somogyi, R. (1999) Gene expression data analysis and modeling. *Pacific Symposium on Biocomputing*, Hawaii.
- Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, Vol. 95, pages 14863-14868.
- Friedman, N., Linial, M., Nachman, I., Peer, D. (2000) Using Bayesian networks to analyze gene expression data. *J. Comp. Biol.* 7, 601-20.
- Giudici, P. (2003) *Applied Data Mining*. Wiley Publishing.
- He, J., Tan, A.H., Tan, C.L. (2003) Self-organizing Neural Networks for Efficient Clustering of Gene Expression Data. *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, USA, pages 1684-1689.

- He, J., Dai, X., Zhao, X. (2007) Associative Artificial Neural Network for Discovery of Highly Correlated Gene Groups Based on Gene Ontology and Gene Expression. *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 17-24.
- Hejlsberg, A., Wiltamuth, S., Golde, P. (2003) *The C# Programming Language*. Addison-Wesley Professional.
- Holmes, I., Bruno, W.J. (2000) Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, San Diego, USA.
- Hunter, L. (1993) *Artificial Intelligence and Molecular Biology*. AAAI Press.
- Jakt, L.M., Cao, L., Cheah, K.S.E., Smith, D.K. (2001) Assessing clusters and motifs from gene expression data. *Genome Res.* 11, 1, 112-113
- Kasturi, J., Acharya, R., and Ramanathan, M. (2003) An information theoretic approach for analyzing temporal patterns of gene expression. *Bioinformatics* 19, 4, 449-458
- Kasturi J., Acharya, R. (2005) Clustering of Diverse Genomic Data using Information Fusion. *Bioinformatics* Vol. 21, Issue 4, pages 423-429.
- Knudsen, S. (2002) *A Biologist's Guide to Analysis of DNA Microarray Data*. Wiley-Liss.
- Kohonen, T. (1995) *Self-Organizing Maps*. Berlin: Springer.
- Laurene Fausett, (1993) *Fundamentals of Neural Networks: Architectures, Algorithms And Applications*. Prentice Hall.
- Laurent, L.C., Chen, J., Ulitsky, I., Mueller, F.J., Lu, C., Shamir, R., Fan, J.B., Loring, J.F. (2008) Comprehensive MicroRNA Profiling Reveals a Unique Human Embryonic Stem Cell Signature Dominated by a Single Seed Sequence. *Stem Cells* published online, DOI: 10.1634/stemcells.2007-1081.
- Lesk, A.M. (2002) *Introduction to Bioinformatics*. Oxford University Press.

- Michaels, G.S., Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X., Somogyi, R. (1998) Cluster analysis and data visualization of large scale gene expression data. *Proceedings of the Pacific Symposium on Biocomputing*, 3:42-53.
- Mount, D.W. (2001) *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press.
- Richter, J. (2002) *Applied Microsoft .NET Framework Programming*. Microsoft Press.
- Setubal, J., Meidanis, J. (1997) *Introduction to Computational Molecular Biology*. PWS Publishing Company.
- Speer, N., Spieth, C., Zell, A. (2004) A Memetic Co-Clustering Algorithm for Gene Expression Profiles and Biological Annotation. *Proceedings of the 2004 Congress on Evolutionary Computation (CEC 2004)*, USA, Vol. 2, pages 1631-1638.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps : Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci USA* Vol. 96, pages 2907-2912.
- Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, 22, pages, 281-285.
- Tisdall, J. (2001) *Beginning Perl for Bioinformatics*. O'Reilly.
- Tompa, M., Li, N., Bailey, T.L., Church, G.M., Moor, B.D., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., Helden, J.V., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., Zhu, Z. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, Vol. 23-1, pages 137-144.
- Törönen, P., Kolehmainen, M., Wong, G., Castren, E. (1999) Analysis of gene expression data using self-organizing maps. *FEBS Lett.*, 451, pages 142-146.
- Wang, H., Azuaje, F., Bodenreider, O. (2005) An Ontology-Driven Clustering Method for Supporting Gene Expression Analysis. *Proc. 18th IEEE International Symposium on CBMS*, pages 389-394

Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., Somogyi, R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA*, Vol. 95, pages 334-339.

Woolf, P.J., Wang, Y. (2000) A fuzzy logic approach to analyzing gene expression data. *Physiol Genomics*, 3, pages 9-15.

İNTERNET KAYNAKLARI

[1] <http://en.wikipedia.org/wiki/Bioinformatics>

[2] <http://www.geneontology.org>

[3] <http://www.biltek.tubitak.gov.tr/bilgipaket/gen/index.html>

[4] <http://en.wikipedia.org/wiki/Gene>

[5] <http://www.cs.uiuc.edu/homes/sinhas/work.html>

[6] [http://en.wikipedia.org/wiki/Locus_\(genetics\)](http://en.wikipedia.org/wiki/Locus_(genetics))

[7] <http://www.obgynacademy.com/basicsciences/fetology/genetics/>

[8] <http://www.arabidopsis.org>

[9] http://www.geneontology.org/teaching_resources/presentations/2006-01_PAG_newbies_smundodi.ppt

[10] <http://search.cpan.org/dist/GO-TermFinder/>

EKLER

Ek 1 Kullanılan Genler

Ek 2 Geliştirilen Uygulama Programı

Ek 1 Kullanılan Genler

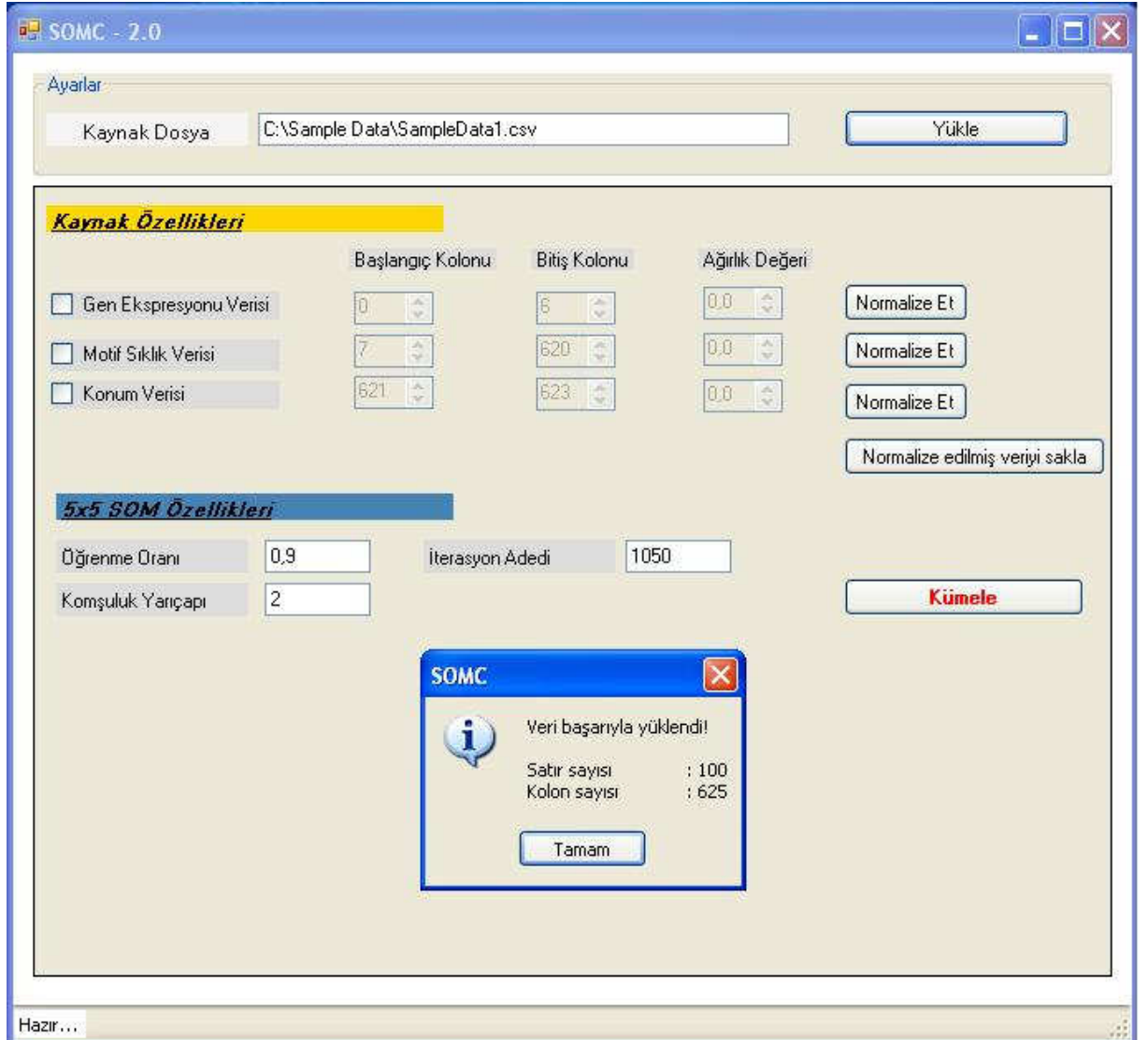
Bu çalışma kapsamında kullanılan 100 adet *Arabidopsis thaliana* genine ait tam liste aşağıdaki gibidir.

1. AT1G07350	21. AT2G01670	41. AT3G05130	61. AT4G02350	81. AT5G02180
2. AT1G08390	22. AT2G02100	42. AT3G07640	62. AT4G03430	82. AT5G15950
3. AT1G12520	23. AT2G17280	43. AT3G09350	63. AT4G11600	83. AT5G17050
4. AT1G15260	24. AT2G18050	44. AT3G12980	64. AT4G14540	84. AT5G22250
5. AT1G16850	25. AT2G21380	45. AT3G13780	65. AT4G15420	85. AT5G26340
6. AT1G20030	26. AT2G22080	46. AT3G14620	66. AT4G17670	86. AT5G27930
7. AT1G21050	27. AT2G22830	47. AT3G15357	67. AT4G22590	87. AT5G28770
8. AT1G24575	28. AT2G25250	48. AT3G16860	68. AT4G23040	88. AT5G39410
9. AT1G28560	29. AT2G25900	49. AT3G21790	69. AT4G23300	89. AT5G40390
10. AT1G29395	30. AT2G26740	50. AT3G26320	70. AT4G25470	90. AT5G41580
11. AT1G29530	31. AT2G29670	51. AT3G26960	71. AT4G25480	91. AT5G44620
12. AT1G32900	32. AT2G34650	52. AT3G28040	72. AT4G25490	92. AT5G48250
13. AT1G34220	33. AT2G37580	53. AT3G47430	73. AT4G26270	93. AT5G49730
14. AT1G34780	34. AT2G38465	54. AT3G50310	74. AT4G28270	94. AT5G51050
15. AT1G55360	35. AT2G39130	55. AT3G54040	75. AT4G29780	95. AT5G55120
16. AT1G63260	36. AT2G41010	56. AT3G55430	76. AT4G33666	96. AT5G55990
17. AT1G64460	37. AT2G41940	57. AT3G55450	77. AT4G34150	97. AT5G58375
18. AT1G67830	38. AT2G45340	58. AT3G55980	78. AT4G36930	98. AT5G58770
19. AT1G71710	39. AT2G46225	59. AT3G60980	79. AT4G39090	99. AT5G62350
20. AT1G75460	40. AT2G46620	60. AT3G61890	80. AT4G39770	100. AT5G66460

Ek 2 Geliştirilen Uygulama Programı

Bu tez çalışması kapsamında geliştirilen uygulama programının kullanıcı arayüzü ana ekranı iki temel bölümden oluşmaktadır:

- İlk bölüm “Ayarlar” adı verilen bölüm olup, analiz edilmek istenen kaynak veri setini temsil eden girdi dosyasının programa tanıtılabildiği bölümdür. Kümeleme işlemine başlamak için ilk olarak girdi veri setinin “Yükle” düğmesine basılarak program belleğine yüklenmesi gerekmektedir. Girdi dosyasının yüklenmesi tamamlandığında program yükleme başarımlı durumu ve yüklenen veri seti hakkında özet bilgileri içeren bir bilgilendirme mesajı vermektedir (bkz. Şekil 1).



Şekil 1. Uygulama arayüzü ana ekranının veri seti yükleme işlemi sonrası görünümü.

Geliştirilen programa girdi olarak tanıtılacak dosyalar için, “.csv” dosya formatını baz alan özel bir girdi dosyası formatı oluşturulmuştur (bkz. Şekil 2). Bu formata göre veri setini oluşturan kolonlar “;” karakteri ile ayrılmaktadır.

Girdi dosyalarının ilk satırı “>Gene” etiketi ile başlayan ve veri seti tanımlarını içeren özel bir satırdır. Söz konusu etiketi takip eden diğer etiketler yine “;” karakteri ile ayrılmış olup her bir etiket bir kolon adına (örneğin bir gen ekspresyon zaman noktasına, bir motif adına) karşı düşmektedir. Takip eden her bir satır belirli bir gene ait verileri içermektedir. Her satırın ilk kolonu temsil ettiği genin ID bilgisini içerir ve takip eden diğer kolonlar ise söz konusu gene ait gen ekspresyonu, motif sıklık veya gen konumu bilgilerine karşı düşer.

Veri formatı ilk satırda tanımlanır

Gen ekspresyonu verisi kolonları

Motif sıklık verisi kolonları

>Gene	0	0.5	1	3	6	12	24	A0000001	A0000002	A0000003	A0000004	A0000005	A0000006	A0000007	A0000008
AT1G07350	0,998	1,174	0,945	0,645	0,221	0,0769	0,285	5	0	1	0	1	1	0	2
AT1G08390	0,991	1,312	0,999	0,807	0,637	0,441	0,284	12	1	0	1	0	0	0	3
AT1G12520	0,943	0,438	0,567	0,374	0,213	0,125	0,183	6	3	1	0	0	0	0	1
AT1G15260	0,999	1,455	1,402	1,052	0,733	0,208	0,306	7	1	0	0	1	0	0	2
AT1G16850	1,156	1,208	1,24	19,08	284	1,780	2,9	1	0	0	0	0	0	0	2
AT1G20030	0,968	1,232	1,344	8,653	95,04	267,4	136,9	12	1	1	0	1	0	0	2
AT1G21050	0,972	1,088	1,997	3,779	1,899	1,348	1,995	11	0	0	0	0	0	0	2
AT1G24575	0,993	1,316	1,208	1,158	1,582	3,494	4,934	11	1	1	0	0	0	0	2
AT1G28560	0,967	1,511	1,174	0,906	0,613	0,436	0,295	3	3	0	0	1	0	0	2
AT1G29395	0,999	1,676	1,239	5,538	47,57	154,3	312,9	10	0	1	0	1	0	0	2
AT1G29530	0,994	1,027	1,324	0,936	0,576	0,277	0,123	8	1	0	1	1	0	0	2
AT1G32900	0,994	1,329	1,182	0,995	0,817	0,478	0,146	4	2	0	0	1	1	0	2
AT1G34220	1,096	1,19	1,097	1,728	4,067	3,655	11	2	1	0	0	0	0	0	2
AT1G34780	0,997	1,304	0,916	1,354	3,01	2,821	0,688	9	2	0	0	0	0	0	2
AT1G55360	0,997	1,266	0,969	1,035	0,927	0,242	0,0622	13	2	0	0	1	0	0	2
AT1G63260	1,081	0,808	0,733	0,601	0,259	0,092	10	4	0	0	0	0	0	0	2
AT1G64460	0,998	1,182	1,735	2,396	3,883	2,969	2,675	8	1	0	0	0	0	0	2
AT1G67830	0,999	0,929	0,952	0,751	0,489	0,313	0,281	10	1	1	0	0	0	0	2
AT1G71710	1,093	0,905	2,3	4,617	8,04	6,002	15	2	0	0	1	0	0	0	2
AT1G75460	0,99	1,75	1,47	1,088	0,915	0,541	0,27	9	2	1	0	2	0	0	2
AT2G01670	0,968	0,858	0,707	1,616	1,494	0,295	0,383	11	2	0	0	0	0	0	2
AT2G02100	0,99	1,125	1,335	8,608	18,57	34,76	37,41	13	1	1	0	0	0	0	2

Şekil 2. Program Girdi dosyası formatı.

- Uygulama arayüzü ana ekranının ikinci bölümü, girdi dosyasının ve kullanılacak olan SOM özelliklerinin tanımlanabilmesi için kullanılır:

“Kaynak Özellikleri” başlığı altında tanımlı olan özellikler ile kullanıcı girdi olarak programa tanıttığı veri seti içinde tanımlı olan genomik veri tiplerini tanımlayabilmektedir. Bu bağlamda üç farklı genomik veri tipi kullanıcı tarafından tanımlanabilir. Bunlar; gen ekspresyonu, motif sıklık ve gen konum verileridir. Veri tiplerini tanımlamak için öncelikle söz konusu veri tipine ait bilgileri içeren kolonların başlangıç indeksi ve bitiş indeksi bilgileri girilmelidir. Bu işlemten sonra söz konusu veri tipi için ağırlık değeri tanımlanabilir. Ağırlık değerlerinin arttırım hassasiyeti “0,1”dir. Kümeleme işlemini başlatabilmek için ağırlık değerlerinin toplamı “1” olmalıdır. Aksi halde program bir uyarı mesajı ile kullanıcıyı uyarır ve kümeleme işlemi başlatılmaz. Her bir veri tipi tanımının yanında yer alan “Normalize et” adlı düğmeye basılarak her veri tipi için ayrı ayrı maksimum-minimum normalizasyonu gerçekleştirilebilmektedir (uygulanan yöntemeye göre her veri tipi ayrı ayrı normalize edilerek kullanılmalıdır, detaylı bilgi için bkz. Bölüm 3.2.1). Kullanıcı ayrıca normalize ettiği veri setini “Normalize edilmiş veriyi sakla” düğmesine basarak saklayabilmektedir.

“SOM Özellikleri” başlığı altında tanımlı olan seçenekler kullanılarak kullanılacak olan 5x5 SOM’a ait parametreler tanımlanabilmektedir. Bu bağlamda öğrenme oranı (α), komşuluk yarı çapı (R) ve iterasyon adedi özellikleri kullanıcı tarafından tanımlanabilmektedir. Komşuluk topolojisi ön tanımlı olarak dikdörtgensel topolojide olup komşuluk yarıçapı R , otomatik olarak toplam iterasyon sayısının $(R+1)$ ’de birlik kısmına her ulaşıldığında bir azaltılır. Öğrenme oranı ise yine otomatik olarak geometrik biçimde azaltılmaktadır (daha fazla detay için bkz. Bölüm 4.5).

Girdi dosyası yüklenip girdi veri seti ve SOM özellikleri tanımlandıktan sonra “Kümele” düğmesine basılarak kümeleme işlemi başlatılabilir. Kümeleme işlemi sona erdiğinde program ana ekranın alt kısmında kümeleme sonuçlarına ait özet bilgiler verilir (bkz. Şekil 3). Özet bilgi ekranına bakarak kullanıcı elde edilen küme sayısını ve her bir kümenin 2 boyutlu harita üzerindeki konumunu görebilir.

SOMC - 2.0

Ayarlar

Kaynak Dosya: C:\Sample Data\SampleData1.csv Yükle

Kaynak Özellikleri

	Başlangıç Kolonu	Bitiş Kolonu	Ağırlık Değeri	
☒ Gen Ekspresyonu Verisi	0	6	0,4	Normalize Et
☒ Motif Sıklık Verisi	7	620	0,3	Normalize Et
☒ Konum Verisi	621	623	0,3	Normalize Et

Normalize edilmiş veriyi sakla

5x5 SOM Özellikleri

Öğrenme Oranı: 0,9 İterasyon Adedi: 1050

Komşuluk Yarıçapı: 2 **Kümele**

<< Özet Sonuçlar >>

Elde edilen küme adedi: 19

Elde edilen kümeler ve komşuluklar:

Küme-[0,0]	Küme-[0,1]	Küme-[0,2]	Küme-[0,3]	Küme-[0,4]
Küme-[1,0]		Küme-[1,2]	Küme-[1,3]	Küme-[1,4]
				Küme-[2,4]
Küme-[3,0]	Küme-[3,1]	Küme-[3,2]	Küme-[3,3]	Küme-[3,4]
	Küme-[4,1]	Küme-[4,2]	Küme-[4,3]	Küme-[4,4]

Detaylı Sonuçları Göster

Sonuçları Kaydet

Hazır...

Şekil 3. Kümeleme işlemi sonrası uygulama arayüzü ana ekranının görünümü.

Elde edilen kümeleri, küme-gen üyelik ilişkilerini ve küme komşuluk ilişkilerini içeren detaylı kümeleme sonuçlarını görebilmek için “Detaylı sonuçları göster” düğmesine basılarak, “Kümeleme Sonuçları” adlı ekran açılabilir (bkz. Şekil 4).



Şekil 4. Kümeleme sonuçlarının detaylı biçimde gösterildiği “Kümeleme Sonuçları” ekranının görünümü.

Elde edilen kümeleme sonuçları “Sonuçları Kaydet” düğmesine basılarak kayıt edilebilir. Sonuçların kayıt edildiği çıktı dosyaları üç ana bilgilendirme bloğundan oluşmaktadır (bkz. Şekil 5). SOM yapay sinir ağı modelinde elde edilen kümeler arası komşuluk ilişkileri önem taşıdığı için, çıktı dosyalarında ilk olarak elde edilen kümelerin 2 boyutlu SOM haritası üzerindeki konumları gösterilmektedir. Çıktı dosyalarında ikinci olarak elde edilen küme adet ve isimlerine ait özet bir bilgilendirme bloğu yer almaktadır. Ve son olarak her bir kümenin nihai ağırlık değerlerinin ve kendisine üye olan genlerin listelendiği blok çıktı dosyalarında yer almaktadır.

SOMC

KUMELEME

SONUÇLARI

Kume-0,0

Kume-1,0

Kume-3,0

Kume-0,1

Kume-3,1

Kume-4,1

Kume-0,2

Kume-1,2

Kume-3,2

Kume-4,2

Kume-0,3

Kume-1,3

Kume-3,3

Kume-4,3

Kume-0,4

Kume-1,4

Kume-2,4

Kume-3,4

Kume-4,4

Elde edilen kümelerin 5x5
SOM üzerindeki görsel
gösterimi

>> Elde edilen kume sayisi "19". Soz konusu kumeler asagidaki gibidir :

Kume-0,0

Kume-0,1

Kume-0,2

Kume-0,3

Kume-0,4

Kume-1,0

Kume-1,2

Özet sonuçlar

>> Kume 0,0

Koordinatlar ; 0,000893035415737185; 0,00093615501187412; 0,00174768595726812; 0,01
AT2G22080
AT3G54040
AT2G25250
AT1G20030
AT1G08390

>> Kume 0,1

Koordinatlar ; 0,000893228625217407; 0,000924762135565245; 0,00152978572506089; 0,0
AT5G15950

>> Kume 0,2

Koordinatlar ; 0,000894266599813749; 0,000926799601299011; 0,00143281843591929; 0,0
AT5G41580

Detaylı sonuçlar:

Kümeler, final koordinatları
ve kendisine üye genler

Şekil 5. Program çıktı dosyası formatı.

ÖZGEÇMİŞ

Doğum Tarihi	01.01.1984	
Doğum Yeri	İstanbul	
Lise	1997-2001	Habire Yahşi Süper Lisesi
Lisans	2001-2005	Kocaeli Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü
Yüksek Lisans	2006-	Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı
Çalıştığı Kurumlar		
	2005-2007	IMS Health Türkiye Yazılım Mühendisi
	2008-	Verdi Sistem Yazılım Mühendisi, Analist