T.C.
SELÇUK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİYOLOJİK ÇİZGE MADENCİLİĞİ: ALT ÇİZGE ÖRÜNTÜLERİNİN BULUNMASI VE ETKİLEŞİM TAHMİNİNDE KULLANILMASI

MEHMET EMİN TURANALP

DOKTORA TEZİ

ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ
ANABİLİM DALI

Konya, 2008

T.C.
SELÇUK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİYOLOJİK ÇİZGE MADENCİLİĞİ: ALT ÇİZGE ÖRÜNTÜLERİNİN BULUNMASI VE ETKİLEŞİM TAHMİNİNDE KULLANILMASI

Mehmet Emin TURANALP

DOKTORA TEZİ

ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

Bu tez 15.12.2008 tarihinde aşağıdaki jüri tarafından oy birliği ile kabul edilmiştir.

Prof. Dr. Ahmet ARSLAN	**Prof. Dr. Saadetdin HERDEM**
Üye	Danışman
Yrd. Doç. Dr. Halis ALTUN	**Yrd. Doç. Dr. Yüksel ÖZBAY**
Üye	Üye

Yrd. Doç. Dr. Salih GÜNEŞ
Üye

ÖZET

Doktora Tezi

BİYOLOJİK ÇİZGE MADENCİLİĞİ: ALT ÇİZGE ÖRÜNTÜLERİNİN BULUNMASI VE ETKİLEŞİM TAHMİNİNDE KULLANILMASI

Mehmet Emin TURANALP

Selçuk Üniversitesi
Fen Bilimleri Enstitüsü
Elektrik-Elektronik Mühendisliği Anabilim Dalı
Danışman: Prof. Dr. Saadetdin HERDEM
İkinci Danışman: Yrd. Doç. Dr. Tolga CAN
2008, 93 sayfa
Jüri: Prof. Dr. Saadetdin HERDEM
Prof. Dr. Ahmet ARSLAN
Yrd. Doç. Dr. Salih GÜNEŞ
Yrd. Doç. Dr. Yüksel ÖZBAY
Yrd. Doç. Dr. Halis ALTUN

Giderek artan sayıda organizmaya ait genom-ölçekli gen ve/veya protein ağlarının ortaya çıkmasına bağlı olarak, son yıllarda önemli ölçüde araştırma çabası biyolojik ağların analizine yöneltilmiştir. Bir protein-protein etkileşim (PPE) ağı, bir organizmanın etkileşen protein çiftlerinin yönsüz olarak bağlandığı bir çeşit biyolojik ağdır. Yönsüz bağlantı ağlarında örüntü bulmanın zorluğundan ötürü, PPE ağlarının analizi için geliştirilen tekniklerin çoğu aralarındaki etkileşimin yoğun olduğu proteinlerin oluşturduğu kümeleri bulmayı veya lineer etkileşim patikalarını bulmayı

hedeflemişlerdir. Bu çalışmada, PPE ağındaki proteinler tekil olmayan Gen Ontolojisi açıklamalarıyla etiketlenmiş ve bu ağda sık görülen etkileşim örüntüleri ortaya çıkarılmıştır. Bulduğumuz örüntüler küme bulma algoritmaları tarafından bulunamayan ve herhangi bir topolojiye sahip olabilecek çizgelerdir.

Bu çalışmada, PPISpan adını verdiğimiz ve gSpan olarak bilinen bir alt-çizge bulma metodundan PPE ağlarına özgü olarak uyarladığımız yeni bir sık geçen örüntü bulma tekniğini öneriyoruz. Bu teknik *S. cerevisiae* (Hamur Mayası) organizmasının PPE ağına uygulanmış ve sık görülen etkileşim örüntüleri belirlenmiştir. Bulgularımız PPE ağlarındaki modüler organizasyonun biyolojik içyüzünü kavramak için faydalı olabilecektir.

Genom ölçekli ve çok miktarda çıktı veren deneyler yardımıyla bol miktarda protein-protein etkileşim verisi ortaya çıkarılmış olmasına rağmen, bu veri setleri hala tamam değildir ve çelişkilerle doludur. Bu çalışmada, yanlış eksi etkileşimleri tahmin ederek bu veri setlerinin kalitesini artırmak amaçlanmıştır. Bunun için PPE ağında sık bulunan örüntüler kullanılmıştır. PPE Ağında olmayan ama aday bir etkileşimin eklenmesiyle birlikte, var olan bir örüntünün yeni bir gömüntüsü ortaya çıkıyor ise, bu aday etkileşim için bir oy olarak kabul edimiştir. Bütün ağ taranarak, ağdaki bütün adayların hangi örüntülerden oy aldıkları belirlenmiştir. Adayların aldıkları oylar ve örüntülerin belirgin özelliklerine bağlı olan bir güven ölçütü belirlenmiştir. Bu ölçüt kullanılarak, girdi olarak verilmiş herhangi bir etiketli ağda etkileşimleri kestiren bir algoritma geliştirilmiştir. GO terimleri ile etiketlenmiş Hamur Mayası organizması ağı üzerinde yapılan deneylerle, geliştirilen algoritmanın pratikte etkili ve iyi bir kestirim performansına sahip olduğu gösterilmiştir.

Anahtar kelimeler: Protein-protein etkileşim ağları; çizge madenciliği; ağ motifleri, altçizge eşitliği, etkileşim örüntüleri, protein etkileşim kestirimi, protein fonksiyon kestirimi.

ABSTRACT

Ph. D. Thesis

BIOLOGICAL GRAPH MINING: DISCOVERY OF SUBGRAPH PATERNS AND THEIR UTILIZATION IN INTERACTION PREDICTION

By

Mehmet Emin TURANALP

Selçuk University
Graduate School of Natural and Applied Sciences
Department of Electrical-Electronics Engineering
Supervisor: Prof. Dr. Saadetdin HERDEM
Co-supervisor: Assistant Prof. Dr. Tolga CAN
2008, 95 pages
Jury: Prof. Dr. Saadetdin HERDEM
Prof. Dr. Ahmet ARSLAN
Assist. Prof. Dr. Salih GÜNEŞ
Assist. Prof. Dr. Yüksel ÖZBAY
Assist. Prof. Dr. Halis ALTUN

In recent years, a considerable amount of research effort has been directed to the analysis of biological networks with the availability of genome-scale networks of genes and/or proteins of an increasing number of organisms. A protein-protein interaction (PPI) network is a particular type of a biological network, in which physically interacting pairs of proteins of an organism are connected with undirected links. Due to the difficulty of identifying patterns in an undirected network, most of the techniques developed for analysis of PPI networks have targeted finding clusters of highly interacting proteins or linear paths of interaction. In this work, we utilize the

Gene Ontology annotations to assign non-unique labels to proteins of a PPI network, and identify frequently occurring interaction patterns. We map known functional annotations of proteins onto a PPI network in order to identify frequently occurring interaction patterns in the functional space. The patterns we identify are not necessarily clique-like highly interacting protein clusters; therefore, are ignored by cluster detection techniques.

We propose a new frequent pattern identification technique, PPISpan, adapted specifically for PPI networks from a well-known frequent subgraph identification method, gSpan. We have applied PPISpan on protein-protein interaction networks of *S. cerevisiae* (Baker's Yeast) and identified frequently occurring interaction patterns, providing further biological insights into the modular organization of protein-protein interaction networks.

Although large amounts of protein-protein interaction (PPI) data have been identified using various genome-scale and high-throughput methods, these data sets are often incomplete and contradictory. We describe a method for improving the quality of these datasets by predicting false negative interactions, using common network motifs (interaction patterns) found in PPI networks. A motif votes for a missing interaction if an embedding of this motif can be realized with the addition of the missing interaction to the PPI network. We make a search in a given PPI network for interaction candidates that complete found motifs and rank them using a confidence measure based on motif votes and distinctive features of motifs such as z-scores and number of edges. We formulate a generic algorithm to apply this method to any labeled large-scale network. Our experiments on yeast PPI network labeled with GO terms indicate that the algorithm is efficient in practice and has good predictive performance.

Keywords: Protein-protein interaction networks; graph mining; network motifs; subgraph isomorphism, protein interaction prediction, protein function prediction.

TEŞEKKÜR

Bu tezimi, beni ben yapan Sevgili Annem Mukaddes'e armağan, varlığımı kendisinden miras aldığım Babam Ahmet Turanalp'in anısına ithaf ediyorum.

Doktora tez çalışmamın başlangıcında beni teşvik ederek cesaret veren, bu süreçte her türlü desteğini ve üstün anlayışını üzerimden eksik etmeyen danışmanım sayın Prof. Dr. Saadetdin Herdem'e ve beraber çalışmaktan büyük heyecan duyduğum, yönlendirmesiyle tezimin şekillenmesine önemli katkıda bulunan ikinci danışmanım sayın Yrd. Doç. Dr. Tolga Can'a içten teşekkürlerimi sunuyorum. Ayrıca, tezim süresince eleştiri ve önerileriyle bana destek veren sayın Prof. Dr. Ahmet Arslan ile sayın Yrd. Doç. Dr. Salih Güneş'e teşekkürü bir borç bilirim.

İÇİNDEKİLER

1. GİRİŞ 1
2. KAYNAK ARAŞTIRMASI 7
2.1. Protein Etkileşimleri 9
2.2. Çizge Madenciliği ve Protein-Protein Etkileşim Ağları 10
2.3. Protein Etkileşim Ağlarında Etkileşim Kestirimi 16
2.4. Protein Fonksiyon Tahmini 20
3. MATERYAL VE METOT 22
3.1. Materyal 22
3.2. Metot 30
3.2.1. PPISpan örüntü bulma algoritması 30
3.2.2. Bir örüntünün istatistiksel önemi 34
3.2.3. Örüntü oylarıyla etkileşim kestirimi 39
3.2.3.1. Etkileşim kestirimi algoritması 43
3.2.3.2. Örüntülerin benzeşmezlik ve örtüşme ölçütleri 44
3.2.4. Altçizge eşitliği 45
4. ARAŞTIRMA SONUÇLARI 53
4.1. Bilinen Moleküler Bileşikler ve Patikalar ile Karşılaştırma 56
4.2. Bazı İlginç Etkileşim Örüntüleri 69
4.3. Etkileşim Kestirimi Sonuçları 74
4.4. Etkileşim Tahmini – Biyolojik Örnekler 79
5. TARTIŞMA 82
6. SONUÇ VE ÖNERİLER 85
7. KAYNAKLAR 87

ŞEKİLLER LİSTESİ

Şekil 1.1. İleri besleme döngüsü örüntüsü 2
Şekil 1.2. Örnek örüntüler ve onların çizge içerisinde bulunan gömüntüleri 3
Şekil 3.1. GO seviyeli yapısından örnek bir kısım. 23
Şekil 3.2. Bir GO terimine düşen PPE ağındaki ortalama düğüm sayısı. 25
Şekil 3.3. Hamur Mayası organizmasının proteinlerinin tüm GO seviyelerine dağılımı (Her bir GO seviyesine düşen protein sayısı) 26
Şekil 3.4. Vasküler iletim'de SNARE etkileşimleri üzerine KEGG patikası. 28
Şekil 3.6. Normal dağılımın yoğunluk fonksiyonunun grafiği üzerinde Z ile α'nın ilişkisi. 37
Şekil 3.7. Örnek G ağının örnek bir parçası. 41
Şekil 3.8. (a)Bağlantı matrisleri verilmiş iki çizge (H ve G), (b) H ve G çizgelerinin temsil ettiği çizgenin bir çizimi. 47
Şekil 4.1. Farklı GO seviyelerinde PPISpan algoritmasının performansı 55
Şekil 4.2. Farklı destek eşik değerleri için PPISpan algoritmasının performansı. 56
Şekil 4.3. STRING ağında bulunuş örüntülerden iki örnek. 60
Şekil 4.4. DIP ve WI-PHI ağlarında bulunmuş örüntülerden dört örnek 61
Şekil 4.5. DIP, STRING ve WI-PHI ağlarından seçilmiş örüntüler. 63
Şekil 4.6. Seçilmiş örüntülerin MIPS bileşikleriyle karşılaştırılarak elde edilen *cpörtüşme* değerleri. 65
Şekil 4.7. Seçilmiş örüntülerin MIPS bileşikleriyle karşılaştırılarak elde edilen *cpsayısı* değerleri. 66
Şekil 4.8. Seçilmiş örüntülerin nakliye ve sinyal iletme patikalarıyla karşılaştırılarak elde edilen *cpörtüşme* değerleri. 67
Şekil 4.9. Seçilmiş örüntülerin nakliye ve sinyal iletme patikaları ile karşılaştırılarak elde edilen *cpsayısı* değerleri. 68
Şekil 4.10. DIP ağında bulunan bir fonksiyonel etkileşim örüntüsü. 69
Şekil 4.11. WI-PHI ağında bulunan bir fonksiyonel etkileşim örüntüsü. 71
Şekil 4.12. MAPK sinyal iletme patikası ile ilgili bir fonksiyonel etkileşim örüntüsü. 72

Şekil 4.13. Vasküler iletim patikasındaki SNARE etkileşimleri ile ilişkili bir fonksiyonel etkileşim örüntüsü. 73
Şekil 4.14. PPE ağında bulunan bazı önemli örüntüler..... 77
Şekil 4.15. Yu'nun yöntemi ile bizim örüntü oyuna dayalı yöntemimizin ROC eğrileri. 77
Şekil 4.16. Farklı örüntü kümeleri kullanıldığında algoritmamız tarafından elde edilen farklı Olasılık Oranları. 79
Şekil 4.17. PRP6 ve BRR1 arasındaki etkileşime oy veren Y3 ve Y4 örüntülerinin durumunu gösteren PPE parçası. 80
Şekil 4.18. APC9 ve DOC1 arasındaki etkileşime oy veren Y1 ve Y2 örüntülerinin durumunu gösteren PPE parçası. 81

ÇİZELGELER LİSTESİ

Çizelge 1. “Moleküler fonksiyon bilinmemektedir” teriminin kökü teşkil ettiği GO kütüphanesinin “moleküler fonksiyon” kategorisinin bir alt kümesi olan GO Slim’in içinde yer alan 22 adet terim (*S. Cerevisiae* organizması için). 24

Çizelge 2. Bulunan örüntülerin sayısı: PPE ağlarında en az 15 gömüntüsü bulunan örüntülerin sayısı. 53

Çizelge 5. Bazı eşik destek değerleri için PPISpan ve gSpan algoritmalarının farklı GO seviyelerinde ve GO Slim terimleri kümesiyle verdikleri çıktı örüntü sayısı ve çalışma süreleri (saniye). 54

Çizelge 3. Tüm örüntülerin rastgele örüntüler ile MIPS bileşikleri kullanılarak elde edilmiş cpörtüşme değerleri açısından karşılaştırılması. 58

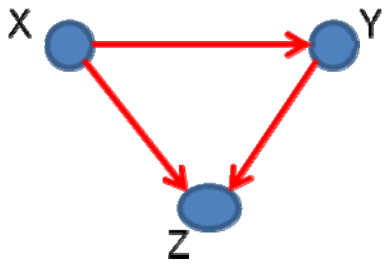
Çizelge 4. Tüm örüntülerin rastgele örüntüler ile nakliye ve sinyal iletme patikaları kullanılarak elde edilmiş cpörtüşme değerleri açısından karşılaştırılması. 59

1. GİRİŞ

Son yıllarda, çok miktarda veri sağlayan deney biçimlerinde sağlanan ilerleme ile birlikte, önemli sayıda organizmanın protein etkileşim bilgilerinin tamamına yakın kısmı elde edilmiştir. Bu deneyler arasında en fazla veri sağlayanlar olarak maya iki-melez ve kütle spektrometreli birleşme eğilimi arıtması zikredilebilir. Benzeri deneylerden elde edilen sonuçlar olasılıksal teknikler yardımıyla birleştirilerek hakkında birden fazla kanıt bulunan etkileşimlerin ön plana çıkarıldığı, daha güvenilir protein-protein etkileşim (PPE) ağları ortaya konulmuştur. Bu ağların ortaya çıkması ile birlikte bunlardan çeşitli veri madenciliği teknikleri ile bilgi çıkartımı ve keşfi yapılmaya başlanmıştır. Örneğin bu ağlarda patika ve moleküler öbeklerin bulunması için, proteinlere fonksiyon atanması için ve öbeklere mensubiyet testi için çeşitli yöntemler geliştirilmiştir. Protein-protein etkileşim ağları, bilgimiz dâhilinde, sık geçen örüntüler için henüz geniş çaplı olarak madenciliğe tâbi tutulmamışlardır. Ağlarda proteinlerin tekil belirteçler ile ifade edilmelerinin ve etkileşimlerin yönsüz ifade edilmelerinin bunda etkili olduğu düşünülmektedir.

Mevcut sık örüntü bulma algoritmaları iki tür ağda başarıyla uygulanmışlardır: Düğümleri tekil belirteçli olmayan kimyasal bileşimler ağında ve etkileşimleri yönlü olarak belirtilen genlerin biyolojik olarak ilginç örüntüler oluşturduğu gen düzenleyici ağında. İlk olarak Uri Alon'un çalışma takımı tarafından keşfedilmiş olan bazı örüntülerin biyolojik olarak anlamlı olabileceği gösterilmiştir (Yeger 2004). Örneğin onların *E. Coli* organizmasının gen düzenleyici ağında buldukları bazı küçük (3-4 düğümlü) fakat önemli örüntülerin (onlar bunu motif olarak adlandırıyorlar) biyolojik olarak anlamlı açıklamaları yapılmıştır. Bunların gen ifadesinin miktarını belirlemede etkili oldukları, zamansal ifade kesitini ürettikleri ve değişen dış kaynaklı sinyallere verilecek tepkileri yönettikleri ifade edilmiştir. Alon ve arkadaşları daha sonra algoritmalarını birden fazla tipteki etkileşimlerden oluşabilen örüntüler arayabilecek şekilde geliştirerek protein-protein etkileşim ve gen düzenleyici ağının birleşimi olan, *S. cerevisiae* organizmasının bir

ağına uygulamışlardır. Örüntü bulmaya çalıştıkları ağ yönlü bir gen düzenleyici ağı olmuştur. Onların buldukları örüntülerden birisi olan ve ileri besleme döngüsü (ing. "feed-forward loop") adını verdikleri örüntü Şekil 1.1'de verilmiştir. Bu örüntünün hedef gen Z'nin cevabını (kodladıkları proteinlerin ifadesini) hızlandırıcı yönde çalışan bir mekanizma sağladığı ifade edilmiştir.

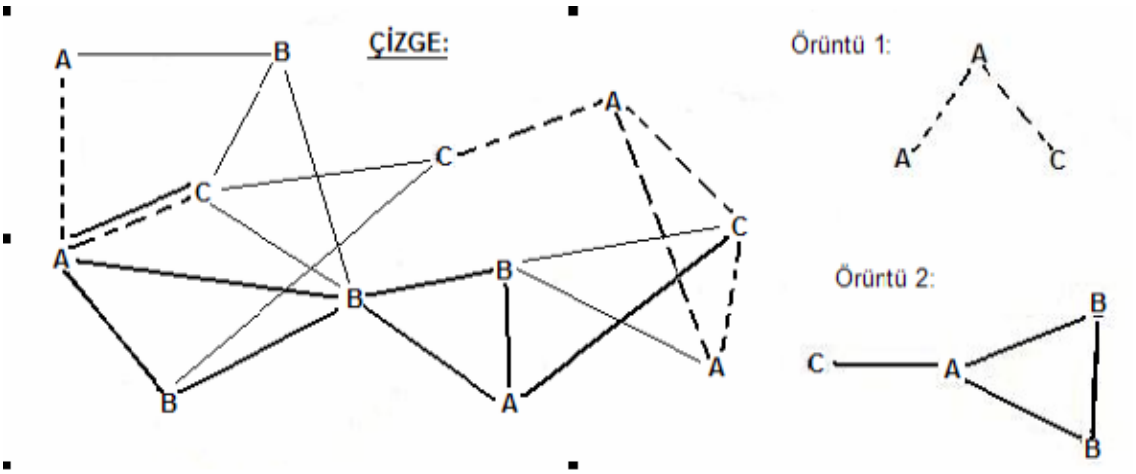


Şekil 1.1. İleri besleme döngüsü örüntüsü

Biyolojik sonuçlar çıkartılabilecek şekilde örüntülerin bulunabilmesi için proteinlere (ve/veya etkileşimlere) tekil olmayan özelliklerin atfedilmesi gerekmektedir. Bu amaçla proteinlere Gen Ontolojisi kütüphanesinden elde edilen fonksiyonel kategoriler atfedilerek bir protein-protein etkileşim ağı tanımlanabilir. Örneğin, bir proteinin dâhil edilebileceği 20 tane genel fonksiyonel kategori belirlenip, ağda bulunan (mümkün olduğunca) bütün proteinler bu kategorilerle etiketlenebilirse, böyle bir ağda değişik fonksiyonel kategoriler arasındaki etkileşimleri ihtiva eden örüntüler keşfedilebilir ve dolayısıyla PPE ağlarının yapısal organizasyonuna ışık tutacak bilgiler ortaya konulabilir.

Diğer pek çok "ağ" gibi PPE ağları da çizge olarak ifade edilirler ve gösterilirler. Çizgelerde sık görülen örüntüleri bulan çeşitli algoritmalar geliştirilmiştir. Bu algoritmaların tamamına yakını iki adımda bunu gerçekleştirirler. Birinci adımda mümkün olan tüm adaylar üretilir/belirlenir. Hiçbir aday örüntünün atlanmadan bunun yapılması önem arz eder. İkinci adımda ise frekans sayımı gerçekleştirilir. Frekans sayımı, bir örüntü adayının çizgede geçen gömüntü sayısını bulma işlemidir. Eğer örüntünün frekansı – ki bu destek olarak da adlandırılır - belli bir eşik değerin- ki bu en düşük destek olarak da adlandırılır- üstünde ise sık geçen

örüntü olarak kabul edilir. Üretilen her bir aday örüntüden eşik değerinin üstünde frekansa sahip olanlar algoritmanın çıktısını oluştururlar. Şekil 1.2.'de örnek örüntüler ve bunların etiketli bir çizge içindeki gömüntüleri verilmiştir. Çizge ve örüntülerdeki düğümler A, B ve C etiketleri ile etiketlenmiştir. Bu şekilde, Örüntü 1'in 2 adet örtüşmeyen ve 4 adet örtüşen örüntüsü bulunmakta olup bunlar kesik çizgilerle ifade edilmiştir. Örüntü 2'ni 2 adet olan örtüşmeyen gömüntüleri kalın çizgilerle çizge içinde belirtilmiştir.



Şekil 1.2. Örnek örüntüler ve onların çizge içerisinde bulunan gömüntüleri

Bir aday örüntünün bir büyük çizgedeki (örneğin genom-ölçekli PPE ağındaki) frekansını bulmak, alt-çizge eşitlik testi adı verilen bir işlemin tekrar tekrar yapılmasını gerektirir. Bu işlemin NP-complete olduğu bilinmektedir (Yan 2006, Washio ve Motoda 2003, Valiente 2002). Dolayısı ile çoğu örüntü bulma algoritmaları aday üretimi aşamasında gereksiz ve fazlalık aday üretimini azaltmayı hedeflemişlerdir. Bunlardan Yan ve Han tarafından geliştirilen gSpan zamansal performansı ile öne çıkmıştır. Onlar, derinlik öncelikli aramaya dayalı bir standart etiketleme sırasına göre adayları üretmişler ve aynı standart etikete sahip adayları eleyerek arama uzayını küçültmüşlerdir. Bir alt-çizgenin sık olup olmadığını belirleyebilmek için, Kuramochi ve Karypis yaklaşımsal en büyük bağımsız küme algoritmasını kullanmışlar ve örüntülerin gömüntülerinin örtüşme çizgesi belli bir eşik değerinden daha büyük ebatta olan bağımsız bir küme olup olmamasına göre

karar vermişlerdir. Aralarında 20,000 düğümlü PPE ağı da olan çeşitli bilimsel alanlardan elde ettikleri gerçek verilerle yöntemlerini test etmişler ve 8 düğüme kadar büyüklükte olabilen sık örüntüler elde etmişlerdir. Amaçları tekil etiketli ağda denedikleri algoritmaların performansını göstermek olduğundan biyolojik bulgular vermemişlerdir.

Bu tez çalışmamızda yeni bir sık geçen örüntüleri bulma tekniği önerilmiştir. Üzerine geliştirmede bulunulan gSpan algoritmasından esinlenerek PPISpan adının verildiği bu algoritma ile Gen Ontolojisiyle (GO) açıklanmış PPE ağlarında sık örüntülerin bulunması amaçlanmaktadır. PPISpan, gSpan algoritmasının her bir GO Terimi etiketine düşen düğüm sayısının az olduğu GO terimleri ile etiketlenmiş PPE ağlarına uygun biçimde geliştirilmişidir. gSpan, sadece azami örüntüleri bulacak şekilde değiştirilmiştir. Bir azami örüntü, başka bir sık geçen çizgenin alt çizgesi olmayan bir sık geçen çizgedir. PPISpan ve gSpan arasındaki bir başka fark da gSpan 'in literatürde "işlem veritabanı" denilen bir çizge kümesi içinde örüntü araması, PPISpan'in ise tek bir büyük çizge içinde örüntü aramak üzere uyarlanmış olmasıdır. PPISpan algoritması Maya (*S. cerevisiae*) canlısının 3 farklı PPE ağında örüntü bulmak amacıyla uygulanmış ve bazı ilginç örüntüler tespit edilmiştir.

PPISpan DIP ağında 15 adetten daha fazla geçen toplam 205 değişik örüntü bulmuştur. Bunların 199 adedi Z-sayısı 2,3'den büyük olan ve dolayısıyla istatistiksel olarak önemli kabul edilebilecek örüntüler olmuştur. Bu örüntülerin gömüntüleri PPE ağının yaklaşık %37'sini kapsamaktadır. STRING PPE ağı için 287 sık geçen etkileşim örüntüsü bulunmuş olup, bunlardan 17'sinin Z-sayısı 2,3'ün üstündedir. Sık geçen örüntüler STRING ağının yaklaşık %40'ının kapsamaktadır. WI-PHI ağında 321'i istatistiksel olarak önemli olan toplam 378 adet örüntü belirlenmiştir. Sık geçen örüntüler WI-PHI ağının yaklaşık %37'sini kapsamaktadır. Örüntülerin gömüntüleri örtüşmemesine rağmen örüntüler örtüşebilirler ancak algoritmamız azami örüntü bulduğundan bir örüntü bir başkasının altçizgesi olamaz. Bulduğumuz örüntülerin büyük bir çoğunluğu ağaç yapısındadır. Yıldız yapısı en fazla görülen topoloji olmuştur. Bu bulgumuz, yaklaşık ağaç yapılarını bulan ama daha hızlı (anında etkileşimli) örüntü bulma algoritmalarının kullanılabileceğini göstermektedir.

PPISpan algoritmamız ile bulunan örüntüler, gürültülü, eksik ve çelişkili olduğu bilinen PPE ağlarında yanlış eksi etkileşimleri belirlemek amacıyla kullanılmıştır. Yukarıda bahsedilen protein çiftleri arasındaki etkileşimleri belirleme yöntemlerinin çok sayıda eksik, gürültülü ve çelişkili veriler ürettikleri bilinmektedir. Maya canlısı için tahmin edilen etkileşim sayısı şu an için tespit edilenden çok daha yüksektir. PPE ağlarını ortaya çıkarmak amacıyla, zaman alıcı ve iş yoğunluklu pek çok biyolojik deney biçiminin yanında, protein-protein etkileşimlerini doğru biçimde belirleyebilen çok sayıda hesaplamaya dayalı metot geliştirilmiştir. Protein-protein etkileşimleri protein özelliklerinden, etraflarındaki etkileşimlerden ve topolojik özelliklerden yararlanarak kestirilebilir. Bu amaçla geçmişte, destek vektör makineleri, Bayes ağları gibi makine öğrenmesi yöntemleri ile topolojik özelliklerden yararlan yöntemler kullanılmıştır.

Bu çalışmada, PPE etkileşim ağlarında bulunan örüntüleri kullanarak protein etkileşimlerini kestiren bir yöntem önerilmektedir. Bir örüntünün sadece bir kenarı eksik kalmış gömüntüsünün bulunması halinde, bunun eksik olan kenarın aslında tespit edilmemiş bir etkileşimin işareti olduğu varsayımından hareketle, bulunan örüntülerin tam olmayan gömüntülerinin eksik kalan kenarının bir etkileşim adayı olabileceğini ve bunun şartları gösterilmiştir.

Bir PPE ağı, protein etkileşim verilerinin çizge biçiminde belirtilmiş gösterimidir. Bu gösterimde proteinler düğümlere, kenarlar da etkileşimlere karşılık gelir. Yukarıda ifade edilen çalışmamızda bazı örüntülerin PPE ağlarında, yapay olarak oluşturulduğu ancak aynı matematiksel özelliklere sahip ağlarda olduğundan önemli ölçüde daha sık geçtikleri tespit edilmişti. Bu tez çalışmamızda, bulunan örüntülerin etkileşimleri kestirmede kullanılabilecekleri gösterilmektedir. Bu amaçla, maya canlısına ait bir PPE ağında sık örüntülerin olmasının yanında, bu örüntülerin bir etkileşimi eksik olan gömüntülerinin kayıp etkileşimlerini belirlemede kullanılabilecekleri gösterilmiştir. Bir örüntünün ağda rastlanan bir görüntüsü (kopyası), gömüntü olarak adlandır. Ağdaki bazı gömüntüler sadece tek bir kenarları örüntüden eksik olduğundan geçerlenememişlerdir. Bir başka deyişle, bu gömüntülerin eksik olan kenarı da olsaydı örüntünün fazladan bir kopyasını teşkil ediyor olacaklardı. Bu kenarlar aday etkileşimler olarak görülmüş ve örüntülerin oylamasına dayalı bir kestirim yöntemi geliştirilmiştir.

DIP veritabanından elde edilen ve en derin GO etiketleriyle zenginleştirilen maya canlısına ait PPE ağında 38 adet örüntü belirlenmiştir. Bu örüntülerin aday etkileşimler için oy vermesini sağlayarak, 964 adet iyi derecede güvenilirliği olan etkileşim tahmininde bulunulmuştur. Tahminlerimiz 2 farklı altın standardı veri kümesiyle test edilmiş ve 2648'e varan olasılık oranlarına ulaşılmıştır.

Bu tez, sırasıyla aşağıdaki bölümlerden oluşmaktadır. Kaynak araştırması bölümünde çizge madenciliği yöntemleri ile protein etkileşim ağları üzerinde gerçekleştirilmiş araştırmalara değinilmektedir. Bunlar arasından çeşitli alanlarda uygulamaları yapılmış olan alt-çizge bulma yöntemleri üzerinde durulmaktadır. Ardından PPE ağlarındaki topolojik bilgiler kullanılarak protein etkileşim kestirimi ve protein fonksiyon kestirimi yapan yöntemler ele alınmaktadır. Ayrıca bulunan örüntülerin ne derece özel olduklarına dair değerlendirme yapmamızı sağlayan rastgele ağlar ile ilgili çalışmalar açıklanmıştır. Materyal bölümünde önce tezde kullanılan materyaller, kullanılan veritabanları, GO protein sınıflandırma kütüphanesi ve bunların özellikleri açıklanmıştır. Metot bölümünde önce, protein etkileşim ağlarında sık geçen alt-çizgeleri (örüntüleri) bulmak için geliştirdiğimiz özgün PPISpan algoritması, ardından protein etkileşim tahmini için geliştirdiğimiz özgün örüntü oylaması algoritması açıklanmıştır. Araştırma Sonuçları bölümünde geliştirdiğimiz algoritmaların performansları ve çıktıları okuyucu ile paylaşılmıştır. Örüntü bulma algoritmamız tarafından *S. cerevisiae* organizmasının protein etkileşim ağında bulunan ve biyolojik açıdan önem arz edebilecek örüntülerden örnekler verilmiş ve bunların biyolojik anlamları konusunda fikir yürütülmüştür. Yine bu bölümde bulunan örüntüler bilinen moleküler bileşikler ve metabolik patikalarla karşılaştırılmıştır. Ayrıca etkileşim tahmini yapan yöntemimizin etkinliği gösterilmiş ve tahminlerimizden örnekler verilmiştir. Tartışma bölümünde proteinlerin fonksiyonlarını tahmin etme amacıyla örüntülerin nasıl kullanılabilecekleri tartışılmıştır. Sonuç ve öneriler kısmında tez çalışmamızın sonuçları özetlenmiş, muhtemel genişletme yönleri konusunda fikirler verilmiştir.

2. KAYNAK ARAŞTIRMASI

Son yıllarda proteinlerin kendi aralarında yaptıkları etkileşimlerin belirlenebilmesi için yeni yöntemler başarıyla uygulanmış, sayıları gittikçe artan pek çok organizmanın protein etkileşim ağları büyük oranda ortaya konmuştur (Mering 2002, Bork 2004). Bu deneysel yöntemler arasında en fazla veri sağlayanlar maya iki-melez (Uetz 2000, İto 2001), kütle spektrometreli birleşme eğilimi arıtması (Gavin 2002, Ho 2002), DNA Mikrodizisi/Gen ortak ifadesi (Eisen 1998), protein mikrodizileri (MacBeath ve Schreiber 2000, Zhu 2001, Jones 2006), yapay ölümcüllük (Ye 2005, Berman 2000) ve bakterideki virüs parazit görüntüsü (Smith 1985) olmuştur. Bu metotlardan bazıları proteinlerin fiziksel etkileşimlerini tespit ederken (ör. maya iki-melez), bazıları da proteinlerin fonksiyonel olarak ilişkili oluşlarını tespit etmektedir (ör. Yapay ölümcüllük). Fiziksel etkileşimleri belirleyen yöntemlerden bazıları ikili ilişkileri bulabilmekte (ör. maya iki-melez), veya proteinin bir bileşiğe aidiyetini tespit edebilmektedir (ör. kütle spektrometreli birleşme eğilimi arıtması). Aynı bileşikte bulunan proteinlerin büyük olasılıkla etkileştikleri varsayılmaktadır. Bu yöntemlerle *Drosophila melanogaster*-Meyve sineği (Giot 2003), *Caenorhabditis elegans*-Küçük solucan (Li 2004), *Saccharomyces cerevisiae*-Hamur Mayası (Krogan 2006), *Escherichia coli* (Bartel 1996), *Helicobacter pylori* (Lin 2005), *Homo sapiens*-İnsan (Stelzl 2005) ve *Mus musculus*-Ev faresi (Lourdes 2008) gibi pek çok organizmanın protein etkileşim ağları önemli miktarda ortaya çıkarılmıştır.

Protein etkileşimleri ile ilgili bilgileri saklamak ve düzenli olarak yayınlamak amacıyla pek çok veritabanı ortaya çıkmıştır. Bunlar arasında ihtiva ettikleri protein bilgileri ve verilerinin güvenilirliği açısından şu veritabanları öne çıkmıştır: Yüksek bir kalite değerlendirmesinden geçmiş etkileşimleri barındıran DIP (Salwinski 2004, Duan 2002); proteinlerin yapısal verilerini de dâhil etmiş olan BIND (Alfarano 2005); literatürden tek tek toplanan etkileşim bilgileriyle oluşturulmuş Mpact/MIPS (Guldener 2006); fonksiyonel kestirimler de dahil olmak üzere, 730 000 adet etkileşim ile en geniş PPE ağına sahip olan STRING (Mering 2005); proteinlerin alt alan etkileşimlerini saklayan iPfam (Finn 2005), ve güvenilirlik derecesi en yüksek

sayıda insan PPE verisi barındıran HPRD (Peri 2003). Bu veritabanlarından DIP (Database of Interacting Proteins), deneysel olarak belirlenmiş protein etkileşimlerinden sadece belli kalite değerlendirmesini geçmiş olanları ihtiva etmesiyle ön plana çıkmıştır. Etkileşim verileri bilimsel literatürden, maya iki-melez deneylerinden, DNA ve protein mikrodizilerinden, kütle spektrometreli birleşme eğilimi arıtması deneylerinden elde edilen verilerden oluşturulmuştur. Etkileşimlerin kalitelerini değerlendirmek için çeşitli değerlendirme metotları geliştirilmiş ve bunlara göre sorgulama yaparak seçme imkanı sağlanmıştır. Bu tezin yazıldığı tarihlerde DIP veritabanında 204 organizmaya ait toplam 19935 proteinin arasındaki 56638 etkileşim bulunmaktadır. Bu organizmalardan meyve sineğinin 21022, Hamur Mayasının 18343, insanın ise 1923 adet protein-protein etkileşimi listelenmiştir. MIPS (Munich Information Center for Protein Sequences) Hamur Mayası canlısı için literatürden tek tek toplanmış etkileşimler kütüphanesidir. Çok fazla miktarda veri üreten deney sonuçları da veritabanında bulundurulmakta ancak ayrı tutulmaktadır. MIPS genellikle literatürde, altın standardı verisi olarak, etkileşim kestirim metotlarının kalitesini ve doğruluğunu değerlendirmede kullanılmaktadır.

Biyolojik protein-protein etkileşim belirleme deneylerine ek olarak, değişik kaynaklardan gelen verileri dolaylı gensel delillerden yararlanarak olasılıksal tekniklerle birleştiren yöntemler sayesinde genom kapsamı büyütülmüştür. Örneğin Jansen ve ark. (2003) Bayes ağları yaklaşımı ile bütün gensel delilleri hesaplanan ağırlıklarına göre birleştirmiştir. Hamur Mayası organizmasında genom ölçekli protein ilişkilerini tahmin etmek için kullandıkları Bayes Ağı yaklaşımı, değişik kaynaklardan gelen, değişik güvenilirlik seviyeleri olan etkileşim bilgilerini ihtimal tabanlı birleştirmeye tabi tutarak, her muhtemel etkileşim için bir güven değeri belirlemeye dayanır. Kendi başına zayıf olan her bir kaynağa güvenirliğine göre ağırlık vermek suretiyle oluşturulan Bayes ağı sayesinde doğruluk oranı yüksek tahminlerde bulunmuşlardır. Benzer bir şekilde, Lee ve ark. (2004) ise genlerin birbirleriyle olan fonksiyonel bağlarının sayısal ihtimallerini değişik gen özelliklerini kullanarak hesaplamışlar ve buna dayalı olasılıksal bir ağ ortaya koymuşlardır. Hesaplamaya dayalı yöntemlerle etkileşimleri belirleyen yöntemler değişik gen özelliklerinden yararlanmışlardır. Bunlardan bazıları şunlardır: genlerin DNA dizini içerisinde aynı gen öbekleri veya yakın protein zincirleri içerisinde yer almaları;

etkileşen proteinlerdeki eş-evrimsel örüntüler; genlerin eş-ifadeleri; protein alanlarının birlikte görülmelerindeki örüntüler; gen fenotipleri; etkileşen proteinlere has dizimsel/yapısal motifler (Shoemaker ve Panchenko 2007).

Farklı kaynaklarda verilerin bir araya getirildiği bir başka PPE ağı çalışmasında, Lee ve ark. (2004), nümerik olasılıkları genler arasındaki fonksiyonel eşleşme bilgilerini göz önüne alarak belirlemişlerdir. Deneylerin özelliklerini ağa etkileşimlerin ağırlıkları yoluyla yansıtmışlardır. Bu şekilde Hamur Mayası organizmasının proteinlerinin %80'inin içinde yer aldığı yüksek doğruluklu bir ağ ortaya koymuşlardır.

2.1. Protein Etkileşimleri

Protein etkileşimleri hücre süreçlerinin sonuçlarını belirleyen en önemli faktördür. Proteinler birbirleriyle oldukça özel şartlarda ve seçici olarak etkileşmektedirler. Öyle ki, etkileşim arayüzlerindeki en küçük bir bozulma hastalıklara neden olmaktadır. Dolayısıyla, protein-protein etkileşimlerini belirleme ve nitelendirme çabası moleküler seviyede biyolojik süreçleri anlamak için büyük önem taşımaktadır. Moleküler seviyede protein tanımanın mekanizmalarını açığa çıkarmak ve hücredeki protein etkileşimlerinin bütünsel bir resmini ortaya koyabilmek amacıyla deneysel teknikler geliştirilmiştir. Bu teknikler yüksek miktarda çıktı verip ve vermemeleri, yaşayan hücrede yapılıp yapılmadıkları, etkileşimlerin ikili veya bileşik düzeyinde olmaları, ilişkilerin fonksiyonel ilişkilendirme mi yoksa fiziksel etkileşimden mi kaynaklandığı ve etkileşimleri nitelendirebilme gibi özellikleri açısından farklılık arz etmektedirler. Bu tekniklerden en önemlilerinin temel özelliklerini burada özetleyeceğiz.

Hamur Mayası iki-melez (Y2H) yöntemi, hücre dışında yapılan etkileşim taramalarını önemli miktarda hızlandırmıştır. Y2H, ökaryotik canlılardaki kopyalama etkinleştiricilerinin iki temel alana sahip oldukları gerçeğinden faydalanılarak geliştirilmiştir. Bu alanlardan biri kurucu DNA dizisine bağlanırken(BD), diğeri kopyalamayı etkinleştirmektedir(AD). BD ve AD'nin ayrılmalarının kopyalamayı imkân dışı bıraktığı ancak bu ikisinin fiziksel olarak ilintili olmalarının kopyalamayı mümkün kıldığı belirlenmiştir. Y2H yöntemine göre, incelenen bir protein BD'ye,

diğeri ise AD'ye bağlanır. Bu yapay proteinler çoğaltılarak Hamur Mayası hücresinin içine sokulurlar. Şayet bu iki incelenen protein birbirleriyle etkileşiyorlar ise, işaretçi gen etkinleşir. Daha önceden bilinen işaretçi genin etkinleştiğinin tespiti ile bu iki proteinin etkileştiği anlaşılmış olur (Walhout 2000, Finley 1994, Bartel 1996). Bu yöntem ile Hamur Mayası haricinde Solucan, Sinek ve İnsan'daki protein etkileşimlerinden yüzlercesi belirlenmiştir.

Kütle spektrometresi (KS), makromoleküler etkileşimleri canlı hücrede belirleyebilmek için geliştirilmiş güçlü bir yöntemdir. KS'nin temel yöntemi, kütle/yük oranlarından yükleri tespit edilebilen iyonlar üretmeye dayalıdır. Bu şekilde polipeptit dizileri belirlenebilir.

Gen birlikte ifadesi yöntemi, değişik şartlar altında iki genin ilişkili oranlarda ifade edilmesinin tespitine dayanır. Bir protein bileşiğinin fonksiyonu alt parçalarının fonksiyonlarına bağlı olduğu için, bileşiğin alt parçalarının ifade seviyeleri ilişkili olmalıdır. Değişik şartlar altında bu ilişkiler korunduğundan, gen ifade biçimlerine bakarak tespit edilebilir. İfade biçimi benzerliği, iki genin göreceli ifade seviyelerin ilişki katsayısı olarak veya mutlak ifade seviyelerinin normalleştirilmiş farkı şeklinde hesaplanabilir. Daha sonra hedef proteinler için bu değerlerin dağılımları rastgele ikililer için olanlarla karşılaştırılarak istatistiksel olarak önemli oranda farklı olanlar belirlenir (Jansen 2002).

2.2. Çizge Madenciliği ve Protein-Protein Etkileşim Ağları

Gen haritası ölçekli protein ağlarını analiz etmek için farklı çalışmalar yapılmıştır. Bu ağların topolojik düzenleniş biçimini belirlemek (Przulj 2004, Valente 2006, Luo 2007, Girvan ve Newman 2002), farklı türler arasıda değişmeden kalmış küçük ağ örüntülerini bulmak (Sharan 2005, Hirsh 2007), etkileşim öbeklerinin keşfetmek (Bader ve Hogue 2003, Spirin ve Mirny 2003, Asthana 2004, Scott 2005, Brohee 2006, Cakmak ve Ozsoyoglu 2007), ayırt edilmemiş proteinlerin fonksiyonlarını kestirmek(Letovsky ve Kasif 2003, Lanckriet 2004, Hu 2005) ve mevcut ağların doğruluğunu artırmak (Mering 2002, Patil ve Nakamura 2005, Chan 2006, Suthram 2006, Collins 2007, Mahdavi ve Lin 2007) amacıyla pek çok yöntem geliştirilmiştir.

Scott ve ark. (2005) protein etkileşim ağlarında önemli işaretleşme patikalarını bulan algoritma geliştirmişlerdir. Bunu yaparken renk kodlaması algoritmasını, ağırlık değerli kenarlar üzerinde uyarlamışlardır. Protein ağları ile ilgili sınırlamalardan yararlanarak algoritmanın verimliliğini artırmışlardır. Örneğin proteinlerin hücrede içe dönük sıralamada etkinleşmeleri kullanılmıştır. Algoritmalarını Hamur Mayası organizmasının protein ağında denemişlerdir. Benzer şekilde, organizmalar arasındaki bilinmeyen ve önemli patikaları ortaya çıkarmak için, Cakmak ve Ozsoyoglu (2007) danışmanlı öğrenme yöntemini kullanmışlardır. KEGG (Kanehisa 2000) patika veritabanında, patikaları temsil eden fonksiyonel şablonları öğrenmişlerdir. Bu şablonları kullanarak başka organizmaların metabolik ağında yeni patikalar keşfetmişlerdir. Ancak bu danışmanlı öğrenme yöntemi alıntı yapılan patikalarla sınırlıdır ve dolayısıyla tamamen özgün patikalar bulunmasına yaramaz. Pandey (2007,2008)'de Cakmak ve Ozsoyoglu gibi düzenleyici ve işaret aktarma ağlarını GO terimleriyle (Gene Ontology Consortium 2000) etiketlemiş, ancak onlardan farklı olarak moleküler etkileşim ağlarında önemli derecede tekrar eden patikaları bulmuşlardır. Ürettikleri NARADA isimli yazılım, her türden etiketleme için PPE veya gen düzenleyici ağlarında önemli derecede sık geçen etkileşim örüntülerini bulmaya imkân tanımaktadır. Ancak, önerilen yöntem sadece 2 ila 5 düğümlü lineer patikaları bulmaya yaramakta, değişik topolojilerdeki örüntüleri bulamamaktadır.

Chen ve Yuan (2006), proteinlerin fonksiyonel ve yapısal verileriyle birlikte mikrodizi veri kümelerini kullanarak kenar ağırlıklı protein çizgesi elde etmişlerdir. Ara ölçeği tabanlı bir algoritma ile bu çizgeyi 266 fonksiyonel öbeğe parçalamışlardır. Bu fonksiyonel öbeklerin yoğun alt-çizgeler olduklarını ispatlamışlardır. İhtimalli protein etkileşim ağında, öbekleri yani topluluk yapılarını bulan algoritmaları yardımıyla gen/proteinlerin öbek ortaklığının fenotip ortaklığına da işaret ettiğini göstermişlerdir.

Patikalara ve öbeklere ek olarak, herhangi bir topolojiye sahip ve tekrarlayan alt-çizgeleri çeşitli büyük ağlarda bulmak için çalışmalar yapılmıştır. Kashtan (2002) ve Yeger (2004) gen uyarlama- düzenleme ağlarında sık geçen alt-çizgeleri bulmuşlardır. Inokuchi (2004), Nijssen ve Kok (2005) ve Yan ve Han (2002) kimyasal bileşen veritabanlarında sık geçen alt-çizgeleri bulmuşlardır. Gen

düzenleme ağlarındaki sık geçen örüntülerin biyolojik olarak ilginç olabileceği fikri ilk olarak Uri Alon'un gurubu tarafından ortaya atılmıştır (Kashtan 2002, Yeger 2004). Onlar, *E. Coli* gen düzenleme ağında motif adını verdikleri küçük (3-4 düğümlü) fakat önemli örüntüler bulmuşlar ve bunlardan bazıları için biyolojik olarak anlamlı açıklamalar vermişlerdir. Sundukları örüntüler gen ifadesini belirlemede özel fonksiyon icra etmektedirler. Örneğin, bazıları zamansal ifade kesitlerini meydana getirmede, bazıları da dışarıdan gelen değişken sinyallere verilecek tepkileri yönetmede etkili olmaktadırlar. Alon grubu daha sonra algoritmalarını geliştirerek birden fazla etkileşim biçimini barındıran ağlara uygulamışlar ancak buldukları örüntüler sadece gen düzenlemeye yönelik olmuştur.

PPE ağlarındaki etkileşim örüntülerini bulmak amaçlı çalışmalar yapılmıştır (Sharan 2005, Hirsh ve Sharan 2007, Oyama 2002, Besemann 2004, Koyuturk 2006). Sharan (2005), Koyuturk (2006) ve Hirsh ve Sharan (2007) çeşitli canlılara ait PPE ağlarını analize tabi tutarak, bu canlılarda ortak olan korunmuş etkileşim örüntülerini bulmuşlardır. Rapor ettikleri örüntüler bu organizmalarda çok görülen özel biyolojik süreçlere karşılık gelmektedirler. Oyama (2002) ve Besemann (2004) birliktelik kuralları madenciliği tekniklerini kullanarak protein çiftleri arasındaki etkileşim kurallarını ortaya çıkarmaya çalışmışlardır. Bilgimiz dâhilinde, tek bir organizmaya ait PPE ağındaki tekrar eden rastgele topolojiye sahip olabilen etkileşim örüntülerini çizge madenciliği yöntemleri ile bulma çalışması bu teze kadar yapılmamıştır.

Çizgelerdeki bütün sık geçen örüntüleri(alt-çizgeleri) bulan değişik algoritmalar önerilmiştir (Hu 2005, Yan ve Han 2002, Koyuturk 2004 2006, Kuramochi ve Kaypis 2005, Wernicke 2005, You 2006). Bu algoritmaların tümü iki aşamalıdır: aday üretimi ve frekans sayımı. İlk adım olan "aday üretimi" aşamasında muhtemel bütün örüntüler birer birer üretilirler. Sonrasında "frekans sayımı" aşamasında her adayın desteği hesaplanarak geçerli olanlar yani örüntü olarak kabul edilenler belirlenir. Desteği belli bir eşik seviyenin üstünde olan adaylar sık geçen olarak kabul edilir. Bir aday alt-çizgenin büyük bir çizgedeki (örneğin gen ölçekli bir PPE ağındaki) frekansını hesaplamak altçizge eşitlik sınamasını defaatle kullanmayı gerektirir. Bu sınama işleminin NP-complete olduğu bilinmektedir(Ullman 1976, McKay 1981, Cordella 2004). Dolayısıyla yukarıda

zikredilen algoritmaların çoğu birinci adım olan aday üretim aşamasında, fazlalık olan adayların üretimini engellemeyi hedeflemişlerdir. Yan ve Han (2002) bunu aday örüntülerin derinlik öncelikli arama temelli kadim etiketlerini hesaplayarak ve eşit etiketli olanları arama uzayından çıkarmak suretiyle sağlamaktadır.

Kuramochi ve Karypis (2005) literatürdeki önceki çalışmalardaki çok çizgede destek aramadan farklı olarak, tek ve büyük bir çizgede sık çizge arama üzerine yatay ve düşey iki algoritmayı geliştirmiştir. Bir alt-çizgenin sık geçip geçmediğine karar vermek için, aday örüntülerin gömüntülerinin örtüşme çizgelerindeki bağımsız kümelerin ebadını hesaplamışlardır. Bunun için en büyük bağımsız küme algoritmasını kullanmışlardır. Her ikisi içinde yakınsayan ve tam sonuç bulan çeşitlerini üretmişlerdir. Bu algoritmaların değişik özelliklerdeki çeşitli çizgelerde ölçeklenebilir bir şekilde çalışabildiğini göstermiş ve kendi aralarında karşılaştırmıştır. Aralarında 20000 düğümlü PPE ağı da bulunan farklı alanlardan temin edilmiş gerçek veri kümeleri ile deneyler yapmışlardır. Algoritmalarının tam sonuç bulan çeşidi protein etkileşim ağında sonlanamamış ancak yakınsayan çeşidi 8'e kadar düğümlü sık altçizgeleri bulabilmiştir. Amaçları algoritmalarının performansını göstermek olduğundan biyolojik olarak ilginç örüntü bulmamışlardır. Temelde farklılıkları olmakla birlikte mümkün olan alanlarda SUBDUE (You 2006) ve SeUS (Ghazizadeh ve Chawathe 2002) algoritmaları ile karşılaştırma yapılmıştır. gSpan ile karşılaştırılmamıştır. Algoritmaları oldukça geniş yelpazeye hitap etmekle birlikte karmaşıktır.

Hu ve ark. (2005) sık bütünleşik alt-çizgeleri girdi ebadından ve sayısından bağımsız olarak çıkarabilen bir algoritma geliştirmişler, bunu 30'un üzerindeki organizmaya ait birlikte-ifade ağında denemişler, çok sayıda homojen kümeyi ve 169 sınıflanmamış geni belirlemişlerdir. Sık bütünleşik alt çizgeler düğümlerinin derecesi belli bir eşiğin üzerinde olan çizgelerdir. Dolayısıyla bu algoritma tam-çizge benzeri modülleri bulmak için uygundur. Nitekim bu yöntemle parçalardaki baskın fonksiyonları bilinmeyen genlere iliştirerek fonksiyon tahmininde bulunmuşlardır. Diğer yöntemlerle karşılaştırma yapılmamıştır.

You ve ark. (2006) SUBDUE ismini verdikleri çizge temeli bir veri madenciliği aracını KEGG metabolik patikalarını ve bu patikalardaki biyolojik olarak anlamlı olabilecek örüntüleri çalışmak için geliştirmişlerdir. Örüntüler

patikaları birbirinden ayırt etmek veya ortak noktaları belirtmek için kullanılmıştır. Koyuturk ve ark. (2004) sık kenarküme madenciliği yaparak, çok çizge içinde belli desteği olan alt çizgeleri bulan derinlik öncelik aramalı bir algoritma geliştirmiştir. Bu algoritma tüm olası alt çizgeleri üretmekte ve altçizge eşitlik testi yapmaktadır. Sık parka-küme madenciliği kullandıkları algoritmalarını organizmada ortak olarak bulunan KEGG metabolik patikalarını bulmakta kullanmışlardır. Hesaplama maliyetlerini düşürmek için, bu patikaların dağınık ve az yoğunluklu olma özelliklerinden faydalanmışlardır. Bir organizmada birbiriyle ortolog proteinler tek bir düğüm şeklinde ifade edilerek alt-çizge eşitlik sınamaları basitleştirilmiş ve sayıları azaltılmıştır. Algoritmanın tamlığı belirtilmemiştir. Diğer algoritmalarla karşılaştırılmamış sadece sonuçları buluş süreleri belirtilmiştir. Algoritmanın yeni kenarları çizgelere nasıl eklediği açık değildir. Koyuturk ve ark. (2006) daha sonra aynı etiketli düğümlerin altçizge eşitlik testinde darboğaz oluşturduğunun bilincine vararak, biyolojik ağlar için ortolog büzüşmesine dayalı verimli bir örüntü bulma algoritması önermişlerdir. Bu algoritmanın büzüşme olmaksızın bulunabilen tüm örüntüleri veri kaybı olmaksızın bulabildiğini ispatlamışlardır. Öte yandan algoritmaları veri kaybına müsaade etmemesine rağmen veri fazlalığı nedeniyle performans düşmelerine açık kapı bırakmaktadır. Büzüşmeyi diğer algoritmaların ön işlemi olarak kullanmayı önermişlerdir. Bununla birlikte tüm düğümler için homoloji/ortoloji bilgisinin mevcut olamayabileceği bir dezavantajdır.

Vanetik ve ark. (2002) destek hesaplamalarını en büyük bağımsız küme kullanarak hızlandırmıştır. Bu makale Apriori algoritmalarının temel özelliklerini ve zayıf yanlarını açıklaması açısından dikkate değerdir. Aday üretimi için k-1 kenarlı iki yolu birleştirerek k ve $k+1$ uzunluklu yeni yollar elde etmiş ve bunun tam olduğunu göstermiştir. Vanetik (2004) ve Gudes (2006) daha sonra kısmi etiketli örüntüleri bulabilen bir algoritma geliştirmişler, bunun tamlığını ispatlamışlardır. Bunu örüntünün düğümlerinin bir alt kümesinin değişken etiketli olabilmesine imkân vererek sağlamışlardır. Algoritmaları tekrarlı zayıflatmaya dayalıdır. Yapay ve gerçek veriler ile uygulama yapmışlardır. Vardıkları en önemli sonuçlar; örüntü arayışında yapının etiketlerden daha önemli olduğu ve kısmi etiketli örüntü bulmanın verimliliğinin en çok kullanılacak destek hesabına bağımlı olduğudur. Verimli aday üretimine yönelik algoritma önermişlerdir. Algoritma önce tüm sık patikaları üretir

ardından bunları belli kurallara göre birleştirerek yeni adaylar üretir. Ortak çekirdeği olan n-patikalarını birleştirerek *n+1* patikaları elde etmeye dayalı bu yöntemle, eleman küme madenciliğinin çalışma birimini ortak kenarı olmayan patikalar cinsinden ifade etmişler, böylece daha az döngü çalışması ve daha az aday testi amaçlamışlardır. Destek fonksiyonuna değinilmemiştir. Üretimin tamlığını göstermişlerdir. Belli çizgelerde FSG (Kuramochi ve Karypis 2005) algoritmasından daha az sayıda aday ürettiğini göstermişlerdir.

Kashtan ve ark. (2004) alt-çizgelerin rastgele örneklenmesi üzerine kurulmuş bir ağ örüntü bulma algoritması geliştirmişlerdir. Bu algoritma aynı zamanda yoğunlaşmaları belirleyebilme özelliğine de sahiptir. Algoritmanın hızlı olması daha büyük örüntülerin keşfedilmesine olanak tanımıştır. Ancak bütün sık örüntüleri bulmayı garanti edememektedir. Bulunan örüntülerin istatistiksel değeri Z-değeri ile hesaplanmıştır. Algoritma Z değerini sınır olarak kullanarak çalışma zamanını, dolayısıyla örüntü bulma derinlik seviyesini ayarlamamıza imkân vermektedir.

Berg ve Lassig (2004) benzer ama tamamen aynı olmayabilen örüntülerin bulunması için istatistiksel model geliştirmişlerdir. Bu model yardımı ile olasılıksal örüntülerin istatistiksel önemini veren değer fonksiyonu çıkarmışlardır. Bu örüntüler yeterince benzer alt çizgelerin yeterli sayıda bulunmasıyla ortaya çıkartılırlar. İhtimali örüntülerin iki belirgin özelliği kenar/düğüm oranının yüksek olması ve altçizgelerin ortalama örtüşmezliklerinin az olmasıdır. Bu özelliklere dayalı olarak, bir uzlaşma örüntü tarafından betimlenen altçizge kümesini büyük bir ağdaki rastgele altçizge ortamından seçtirecek değer fonksiyonunu tasarlamışlardır. En yüksek değeri verecek altçizge örtüştürmelerini ve parametre optimizasyonunu fizikteki fırıl modele çevirerek , benzetimli tavlama yöntemiyle çözmüşlerdir. Bu şekilde daha önce U. Alon grubu tarafından kullanılan gen düzenleyici ağında 4 ve 5 düğümlü örüntüler bulmuşlardır.

Rastgele çizgeler

Bir örüntünün istatistiksel olarak önemli olup olmadığının belirlenebilmesi amacıyla boş hipotez sınamasının yapılması gerekir. Boş hipotezin kurulması için gerekli veri ise belli bir rastgele ağ modeli kullanarak çok sayıda rastgele ağ üretilip, bu ağlar ile ilgili istatistiksel bilgilerin elde edilmesini gerekli kılar. Literatürde

biyolojik ağlara uygun rastgele ağ modelleri vardır. Bu tezde kullanılan modele en yakın olmaları açısından iki farklı rastgele ağ modeline burada değinilecektir.

Organizmaların sekanslarının belirlenmesinin ardından, gen haritalarının anlaşılması yolundaki ikinci adım protein etkileşimlerinin analizi olmuştur. Bu süreçte, başka yapay ve tabii pek çok ağa benzer biçimde, PPE ağlarının ilginç özellikler gösterdiği belirlenmiştir (Strogatz 2001). Örneğin, bu ağların güç yasasına uydukları, rastgele çizgelerin tipik özelliği olan küçük dünya etkisini gösterdikleri, ağ geçirgenliği ve topluluk yapısı özelliklerini gösterdikleri (Girvan 2002), rastgele bir düğüm çıkarımına karşı dayanıklı oldukları (Albert 2000) gösterilmiştir.

Sole ve ark.(2002) , PPE ağlarının yukarıda ifa edilen tipik özelliklerinin ve topolojilerinin kaynağını anlamak için protein etkileşimleri evrimi modeli geliştirmişlerdir. Gen kopyalanması ve değişimine dayalı olan bu modelin yukarıda belirtilen istatistiksel düzenleri sağladığını göstermişlerdir. Bu model, basit ve küçük bir ağ ile başlayıp rastgele seçilmiş düğümlerin kopyalanmasını sonra da belli ihtimaller dâhilinde bu yeni düğümlere etkileşimlerin eklenmesi veya silinmesini içerir (değişim). Sonuçta elde edilen ağların özellikleri ile gerçek PPE ağınınkinin istatistiksel olarak önemli yakınlıkta olduğunu göstererek modellerinin geçerliğini göstermişlerdir.

Vazquez ve ark. (2003), PPE ağlarının evrimini ifade eden Sole ve ark.'ninkine benzer fakat daha sade bir model geliştirmişler ve bu modelin rastgele veya hedeflenmiş hasarlara karşı hata toleranslarının gerçek ağdakine benzer olduğunu göstermişlerdir. Farklı kopyalama ve değişme oranları kullanarak elde ettikleri sonuçları karşılaştırmışlardır. Sole ve Vazquez'in çalışmaları PPE ağlarının topolojik özelliklerini daha anlaşılır kılmalarının yanında, bu tezde kullanıldığı gibi, rastgele PPE ağlarının üretimine dair prensipleri de ortaya koymuşlardır.

2.3. Protein Etkileşim Ağlarında Etkileşim Kestirimi

Canlı hücrelerinde gerçekleşen olayları anlamak için protein-protein etkileşimlerini anlamak büyük önem taşımaktadır. Bu etkileşimleri ortaya çıkarmak amaçlı deneysel yöntemlerden önceki bölümde bahsedildi (Bkz. Bölüm 2.1). Deneysel yöntemlerin ürettikleri verilerin -çoğunlukla deneylerin tabiatından

kaynaklanan nedenlerden ötürü- gürültülü, eksik ve çelişkili oldukları bilinmektedir (Mering 2002, Deane 2002, Edwards 2002). Hamur Mayası canlısında olduğu tahmin edilen protein etkileşim sayısı şu anda belirlenmiş olandan oldukça daha fazladır (Mering 2002, Bader 2002). PPE ağlarının zahmetli ve zaman alıcı canlının içinde deneylerle belirlenmesinin yanında, büyük doğruluk oranları ile hesaba dayalı kestirimler yapabilen yöntemler de geliştirilmiştir (Dohkan 2004, Jansen 2003, Martin 2005, Qi 2006, Wu 2006).

Deneysel yöntemler büyük miktarda -muhtemelen önemli miktarda gürültü barındıran - veri ürettiklerinden bu verilerin analiz edilmeleri, doğrulanmaları ve tamamlanmaları gerekir. Bu amaçla hesaplamsal yöntemler başarıyla kullanılmıştır ve bazı durumlarda deneysel yöntemlerin belirleyemediği etkileşim detaylarını da ortaya çıkarabilmişlerdir (Shoemaker 2007). Bu yöntemlerin bazıları deneysel ve hesaplamsal teknikleri birlikte kullanmış ve doğrudan fiziksel etkileşimleri belirlemek yerine potansiyel çiftler arasındaki fonksiyonel ilişkileri çıkartmaya çalışmışlardır.

Hesaplamsal yöntemlerin en çok ilgi görenleri gen haritası/protein bağlamından çıkartım yapan yöntemler olmuştur. Gen komşuluğu ve gen öbeği yöntemlerinde, yakın ve ilgili fonksiyonlara sahip genlerden potansiyel olarak etkileşen proteinleri kodlayan kümelerinin tek bir birim (operon) halinde kopyalandıkları bilgisinden yararlanılmıştır. Genler arasındaki dizilim mesafesinden yola çıkılarak operonların belirlenmesi için pek çok metod geliştirilmiştir (Ermolaeva 2001, Moreno 2002, Salgado 2000, Strong 2003).

Filogenetik yanay yöntemleri, fonksiyonel olarak ilişkili ve muhtemelen etkileşiyor olan proteinlerin birlikte evrimleştiği ve DNA dizisi tümüyle ortaya çıkarılmış canlılar arasında aynı alt kümede ortologlarının olduğu tezine dayanır. Zaten, işlevlerini gerçekleştirebilmek için bileşiklerin ve patikaların parçalarının aynı anda bulunmaları gerekmektedir. Bir gen için filogenetik yanay N elemanlı bir vektör olup, N adet canlının gen haritasında bu proteinin bulunup bulunmaması 0 ve 1'lerle ifade edilir. Yanaylarına bakılarak proteinler (bit uzaklığı ölçütü ile) kümelenirler ve aynı kümedeki genler fonksiyonel olarak ilişkili kabul edilir. Bu yöntem etkili olmakla birlikte yüksek hesaplamsal maliyeti vardır ve yüksek yanay bilgilerine gereksinim duyarlar (Eisenberg 2000, Pellegrini 1999).

Rosetta Anıt Yöntemi, değişik canlıların gen haritalarındaki protein dizilerinden yararlanarak protein etkileşimlerini çıkarmaya çalışır. Bir canlıda etkileşen bazı proteinlerin başka bir canlıda genlerinin homologlarının birbirine yapışık veya iç içe geçmiş halde bulundukları gözlemlenmiştir. Genlerin birleşikliğinin, birbiriyle etkileşen proteinlerin birlikte ifadelerinin optimize edilmesine yönelik olduğu düşünülmüştür. Bu yöntemle sadece Escherichia coli canlısında 6000'in üzerinde yüksek potansiyelli etkileşim aday çiftleri belirlenmiştir. Bu çiftlerin yüksek oranda başka bir canlıdaki tek bir proteinle benzeştikleri gözlemlenmiştir. DIP veritabanından alınan deneysel verilerle yapılan karşılaştırmada, deneysel etkileşimlerin % 6,4'ünün Rosetta Anıtı proteinleriyle bağlantılı olduğu görülmüştür (Marcotte 1999, Enright 1999, Yanai 2001, Marcotte 2002).

Gen dizisi temelli birlikte evrimleşme yöntemleri, etkileşen proteinlerin birlikte evrimleştiği dolayısıyla bir proteindeki fonksiyon kaybettirici bir değişimin diğer(ler)indeki ilgili değişim(ler)le telafi edildiği varsayımından yola çıkılarak geliştirilmişlerdir. Benzeş olmayan ancak etkileşen protein ailelerinin filogenetik ağaçlarının benzerliği ile birlikte evrimleşme ifade edilmektedir (Jothi 2006, Pazos 2001, Goh 2000, Ramani 2003, Gertz 2003). Bu yöntemler genellikle ağaçlar için uzaklık matrisleri oluşturup karşılaştırırlar. Bir ailedeki proteinlerin paralog olma ihtimali bu matrislerin katsayıları şeklinde düşünülmüştür. Ancak bu bilgi her zaman elde edilemeyebilmektedir. Bunun üstesinden gelebilmek için de paralogların farklı bağlanma özelliklerinden yararlanılmıştır (Gertz 2003). İki protein ailesinin uzaklık matrislerinin hizalanması işlemi yapılarak en yakın benzerlik seviyesi tespit edilir. Ancak bu yöntemin 30 proteinden fazla üyeli büyük aileler için pratikte kullanılamayacak kadar zaman alıcı (hesaplamsal olarak masraflı) olduğu gözlemlenmiştir.

Yukarıda sayılan gen analizine dayalı yöntemlerin dışında sınıflandırma yöntemleri de protein ve protein alanı etkileşimlerini tespit etmede başarıyla kullanılmıştır (Lee 2004, Jansen 2003, Albert 2004, Qi 2005, Ben-Hur 2005, Chen 2005, Bader 2004, Gilchrist 2004, Yamanishi 2004, Zhang 2004, Fariselli 2002, Koike 2004, Bradford 2005). Bu yöntemler değişik kaynaklardan alınan verilerle bir eğitme işlemi gerçekleştirildikten sonra protein çiftlerinin sınıflandırmasını

yapabilmektedir (pozitif olarak sınıflandırılanlar etkileştikleri kestirilenler ve negatif olarak sınıflandırılanlar ise etkileşmedikleri şeklinde tahmin edilenler). Destek Vektör Makineleri ve Rastgele Orman Kararı yöntemleri bu alanda başarıyla kullanılmıştır (Scholkopf 2004, Qi 2006). Proteinlerin özellikleriyle ve deneysel metotlardan gelen veriler ile her bir protein bir vektör olarak kodlanır ve vektör çiftlerinin uzaklıkları kullanılarak sınıflandırma gerçekleştirilir. Farklı alanlardan gelen delil niteliğindeki verilerin kullanılmasının sınıflandırma doğruluğunu artırdığı gözlemlenmiştir.

Etkileşimler protein özelliklerinden, çevredeki etkileşimlerden ve topolojik özelliklerden çıkarılabilir. Dohkan ve ark. (2004) protein alanları, aminoasit bileşimi ve hücre içi yerleşimi gibi protein özelliklerinden, destek vektör makineleri yardımıyla protein etkileşimlerini kestirmişlerdir. Jansen ve ark. (2003) etkileşimle zayıf bir ilişkisi olan gen ile alakalı özellikleri, Bayes ağı yaklaşımı ile birleştirerek kaliteli etkileşim tahminleri yapmışlardır.

Bazı çalışmalarda sadece ağda tespit edilen topolojik bilgilerden faydalanarak etkileşim kestiriminde bulunulmuştur. Özel olarak ağ motifleri kullanılmıştır. El ile tespit edilen ağ motiflerini, çokça hata barındıran PPE ağlarında, Saito ve ark. (2002) yanlış artıları tespit etmek için kullanırken, Albert ve ark. (2004) yanlış eksileri tespit etmek için kullanmışlardır. Albert ve ark. üç tane etiketsiz motif kullanarak proteinlerin ait oldukları etkileşim komşuluklarını açığa çıkarmaktadır. Daha sonra bir tavsiye motoru, proteinlerin etkileşim imzalarının benzerliğine dayalı olarak en uygun etkileşim adaylarını sıralayarak belirlemektedir. Her iki yöntem de sadece sınırlı ve önceden elle belirlenmiş bir motif kümesi kullanarak ümit vaat eden sonuçlar elde etmişlerdir. Yu ve ark. (2006), PPE ağını özürlü takımları tespit etmek için aramışlar ve özürleri kapatacak etkileşimleri kestirmişlerdir. Büyük ağlara bunu uygulamışlar ve yüksek doğruluk oranında tahmin yapabildiklerini protein bileşikleri örnekleri eşliğinde göstermişlerdir. Yöntemleri çabuk ve etkili kestirimlerde bulunmaktadır ancak takım biçiminde olmayan oluşumlardaki etkileşimleri kestirememektedir.

2.4. Protein Fonksiyon Tahmini

Önceki bölümde, saklanmış etkileşim örüntülerinin, etkileşim tahmininde (Albert 2004) ve PPE ağının altında yatan parçaları keşfetmede (Itzkovitz 2005) kullanıldıklarını ifade edilmişti. Benzer şekilde, örüntüler karmaşık ağların yapısal tasarım prensiplerini açıklamada faydalı bir kavram olarak ilgi görmektedir (Wernicke 2006, Milo 2004).

Bir proteinin hücre içindeki fonksiyonu onun başka proteinlerle olan etkileşimine bağlıdır. Bu etkileşimler pek çok organizma için önemli oranda ortaya çıkarılmış olmalarına rağmen, proteinlerin fonksiyonlarına olan etkileri açısından henüz yeterince araştırılmamışlardır. Ortaya çıkan büyük miktardaki PPE bilgisi, proteinleri sadece komşuları ile birlikte değil tüm ağ ile birlikte değerlendirmemizi sağlamaktadır. Proteinin yapısı, fonksiyonu ve hücre içindeki yeri gibi protein özelliklerinin tümü PPE ağını etkilemekte ve ondan etkilenmektedir (Chou 2000, Spirin 2003, Aloy 2003). Örneğin Spirin ve ark. ağda proteinlerin fonksiyonel olarak öbeklendiklerini göstermiştir. Bu durum olağandır çünkü belli bir biyokimyasal işlevi icra edebilmeleri için proteinlerin birlikte hareket etmeleri gerekir, dolayısıyla tamamen aynı olmasa da bazı fonksiyonları paylaşacaklardır. Başka bir örnek de proteinlerin yapısı ile ilgilidir. Proteinler etkileşirken belli yapıda olanlar belli başka yapıda olanlarla seçici olarak etkileşeceklerdir (Aloy 2003).

Yapısal ve fonksiyonel etkileşim motifleri proteinlerin PPE içindeki konumlarından faydalanarak protein fonksiyonu kestirimi yapan yeni yöntemlerin bulunmasına yol açmıştır (Schwikowski 2003, Nabieva 2005). Bunlardan en meşhur olan yöntem, fonksiyonu tahmin edilecek proteinin etrafındaki (komşularındaki) en çok geçen fonksiyonları seçen ve "çoğunluk oylaması" adı verilen yöntemdir (Schwikowski 2003). Buna göre belli bir proteinle etkileştiği bilinen proteinler onun fonksiyonunu tahmin etmede kullanılırlar. Bu proteinlerin (ki bunlara komşu proteinler denmektedir) fonksiyonları içinden en fazla olan fonksiyon tahminde bulunmak istediğimiz proteine fonksiyon olarak atanır. Bu atamada ihtimaller de göz önüne alınarak fonksiyonların sayılarıyla orantılı bir seçim de yapılabilir. Son derece basit olmakla birlikte bu yöntem en başarılı yöntemlerden biridir. Hishigaki ve ark. (2001) fonksiyonu belirlenecek proteinin belli bir yarıçapı yakınlığındaki

komşularının tümünü sayarak tahminde bulunmuşlardır. Ancak bu yöntem komşuluk alanı içindeki topoloji bilgisini kullanmamaktadır. "Çoğunluk oylaması"ndan esinlenerek ve üçüz protein etkileşim yapılarını kullanarak frekans sayımı yapan Chen ve ark. (2007) daha yüksek başarı oranları elde etmişlerdir. Proteinlerin içinde bulundukları üçüzlerin kategorilerini sayarak en çok geçeni seçmişlerdir. Anlaşılacağı üzere bu yöntemle sadece topoloji bilgisini kullanarak fonksiyon tahmini yapmışlardır.

Kirac ve ark. (2006) Gen Ontolojisi terimlerinin aralarındaki ilişkileri kullanarak protein fonksiyon tahmininde bulunmuşlardır. Bu amaçla ağda rastgele yürüme yöntemiyle elde ettikleri önemli GO terim dizilerinin sonlarına gelecek en muhtemel GO terimlerini belirlemişler ve bunları uyuşan dizilerin sonundaki etiketlenmemiş proteinlere fonksiyon olarak atamışlardır.

Vazquez ve ark. (2003) ile Karaoz ve ark. (2004) protein etkileşim ağının genel topolojik özelliklerinden faydalanarak, komşu proteinlerin farklı etiketlenmelerini asgariye çekecek şekilde fonksiyon tahmininde bulunmuşlardır. Karaoz ve ark. (2004) ek olarak gen ifadesi verileriyle etkileşimleri ağırlıklandırmışlar ve hesaplamalarına bu bilgileri dâhil etmişlerdir. Çok yönlü k-kesimi probleminin genelleştirilmiş halini yöntem olarak kullanmışlardır.

Nabieva ve ark (2005), asgari-çizge-kesi problemine başka bir yaklaşım olan ağ akış fikrini uyarlamışlardır. Fonksiyonu bilinen proteinler, çizgedeki kaynak düğümleri temsil edecek şekilde fonksiyonları bilinmeyen proteinlere doğru akış sağlanır. Yerel akış kuralları gereğince ve proteinler arasındaki uzaklığa da bağlı olarak her bir protein için her bir fonksiyonun aldığı akışlar hesaplanır. Böylelikle topoloji bilgisi fonksiyon tahmininde kullanılmış olur.

3. MATERYAL VE METOT

3.1. Materyal

Bu tezde Hamur Mayası canlısına ait üç farklı PPE ağını kullandık. Bunlar DIP, STRING ve WI-PHI'dır. Çok miktarda veri sağlayan deneylerden alınan etkileşim verilerini barındıran Database of Interacting Proteins (DIP) veritabanının 11 Nisan 2007 tarihli nüshasını kullandık(Salwinski 2004). DIP ağı 17491 etkileşim ve 4932 protein barındırmaktadır. DIP PPE ağı yönsüz ve ağırlıksız bir çizge olarak temsil edilmiş, proteinlerin kendileriyle yaptıkları etkileşimler ihmal edilmiştir.

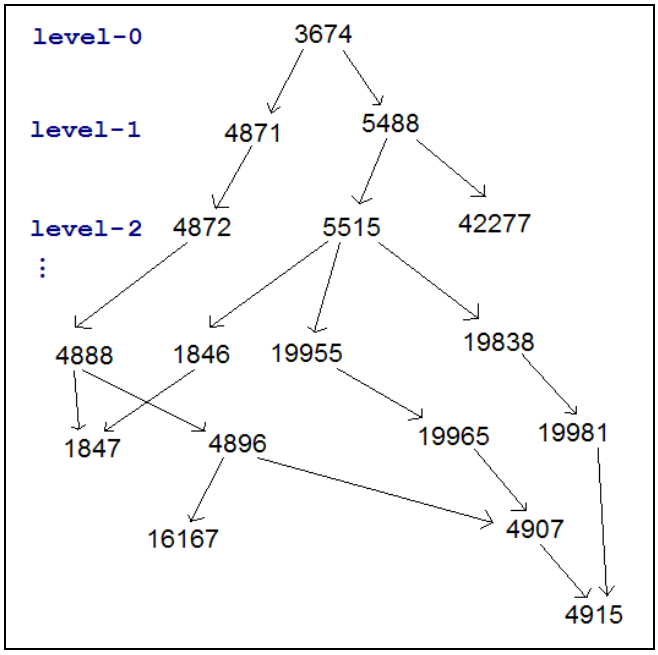
STRING veritabanı bazı organizmalar için güven ağırlıklı etkileşim bilgileri içermektedir (Mering 2007). Bu çalışmada güven eşiği 0.95'in üzerinde güven notuna sahip 20050 protein etkileşimi kullanılmıştır. Bu küme Hamur Mayası canlısının 2952 proteinini kapsamaktadır. Kullandıkları veri kaynaklarından ötürü (örneğin gen ifadesi verisi) etkileşimler fiziksel olanların yanında dolaylı etkileşimleri de kapsamaktadır.

WI-PHI veritabanı doğrudan fiziksel etkileşimlerin daha ön plana çıkarıldığı, bunun yanında doğrudan olmayanların öneminin baskılandığı ağırlıklı bir Hamur Mayası PPE ağıdır (Kiemer 2007). Etkileşimlerin tümünün sayısı 50000 civarındadır ancak biz güven derecesi 9,4183 eşiğinden yüksek olan 20097 adet etkileşimi deneylerimizde kullandık. Böylece DIP ve STRING veritabanlarıyla karşılaştırma yapabileceğimiz bir ağ büyüklüğü sağlamış olduk.

PPE ağındaki proteinlere fonksiyonel sınıf etiketleri atamak için Gen Ontolojisi (GO) etiketlerini kullandık. GO projesi, gen ürünlerinin farklı veritabanlarında tutarlı betimlemelerinin tutulabilmesi ihtiyacını karşılamayı hedefleyen bir ortak çalışma projesidir. GO'daki üç ana sınıf biyolojik süreçler, hücre içi parçalar, ve moleküler fonksiyonlar açısından türden bağımsız betimlemeler içermektedir. GO'nun aşamalı yapısı, elde bulunan verilere bağlı olarak, gen ürünlerine değişik seviyelerden özellik atanmasına imkân vermektedir. GO moleküler fonksiyon sınıfının örnek bir kısmı Şekil 3.1'de verilmiştir. Bu sınıflandırma sistemi, yönlü çembersel olmayan çizge biçiminde ifade edilir. Bir

terimin birden çok ebeveyni ve bu ebeveynlerle birden çok tipte ilişkisi olabilir. Örnekteki moleküler fonksiyon sınıfının kök terimi 3674'dür. Terimler arasındaki oklar *dir* (ing. *is a*) ilişkisini ifade etmektedir. Bu seviyeli sistemdeki bir terimin derinliği onun kök ile arasındaki en kısa mesafe olarak tanımlanır. Kökün derinliği sıfırdır. Ocak 2007 itibariyle moleküler fonksiyon ağacındaki bir terimin en yüksek derinliği 13'tür. *S. cerevisiae* canlısının DIP veritabanından alınmış PPE ağındaki proteinlere atanmış terimler için derinlik en fazla 13 ve ortalama 5,33'dür.

Hamur Mayası canlısının DIP PPE ağını GO moleküler fonksiyon terimleriyle etiketledikten sonra, bu ağda 3840 adet proteinin etiketlenmiş olduğu ve bunlar arasında 11449 etkileşim olduğu görülmüştür. Bu proteinlerden 697 tanesinin ağdan kopuk olduğu yani diğer proteinlerle etkileşimlerinin olmadığı görülmüştür. Bu ağda üçüncü GO moleküler fonksiyon seviyesinde 266 tane farklı terimin olduğu gözlemlenmiştir.



Şekil 3.1. GO seviyeli yapısından örnek bir kısım.

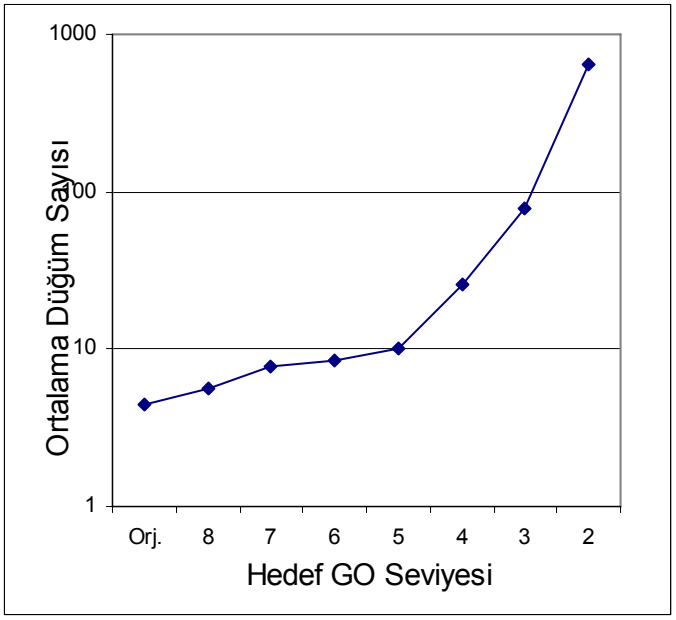
Bu çalışmada Gen Ontolojisi'nin moleküler fonksiyon sınıfının terimlerinin yanı sıra GO Slim terimleri kullanılmıştır. GO Slim, GO terim kütüphanesinin 22

terimden oluşan küçük bir alt kümesidir. GO kütüphanesini özetlemek amacıyla biyologlar tarafından derlenmişlerdir. Bunlar Çizelge 1'de listelenmiştir. Örneğin, Şekil 3.1'de görülebilen 5515 terimi bir GO Slim terimidir. Bu terimlerle proteinler etiketlenerek transkripsiyon faktörleri ve kinazlar gibi genel özelliklerin atanması sağlanmıştır. Amacımız farklı protein fonksiyonlarını barındıran ve PPE ağının değişik yerlerinde bulunan etkileşim örüntülerini belirlemektir. Bir proteinin birden fazla etiketinin olmasına müsaade edilmiştir ve örüntüler proteinlerle eşleştirilirken bütün olasılıklar (kombinasyonlar) göz önüne alınmıştır. Bu çalışmada GO web sitesinden 5 Kasım 2007 tarihinde indirilen etiketler kullanılmıştır. Etiketlerin GO Slim eşleri GO aşamalarında yukarı doğru ebeveyn bağlantısı takip edilerek bulunmuş, bu şekilde ilk bulunan GO Slim etiketi geçerli etiket olarak kabul edilmiştir.

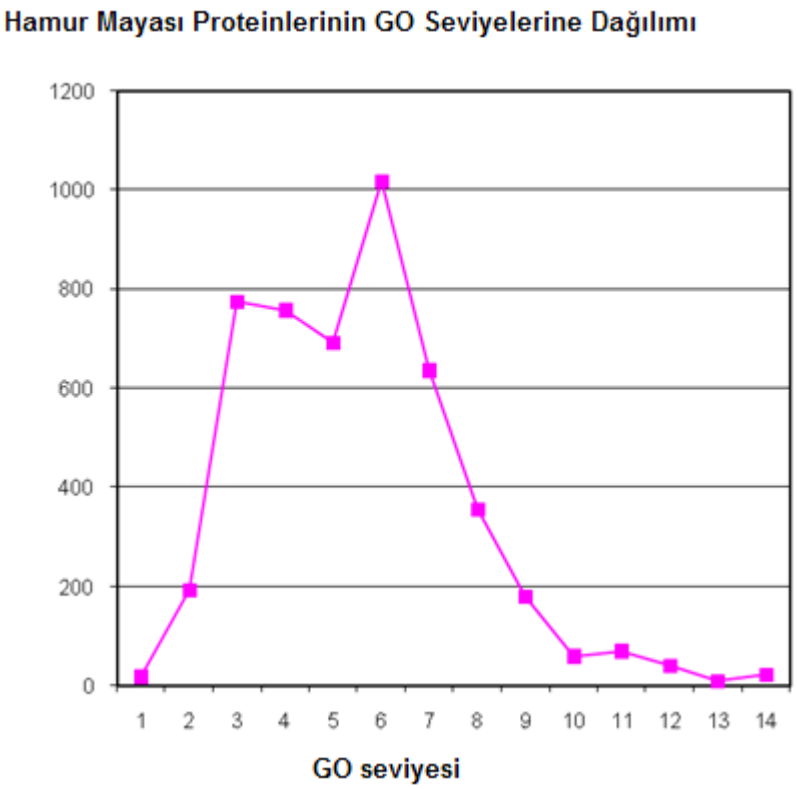
Çizelge 1. "Moleküler fonksiyon bilinmemektedir" teriminin kökü teşkil ettiği GO kütüphanesinin "moleküler fonksiyon" kategorisinin bir alt kümesi olan GO Slim'in içinde yer alan 22 adet terim (*S. Cerevisiae* organizması için).

Terim No	Tanım
GO:3674	molecular function unknown
GO:16787	hydrolase activity
GO:16740	transferase activity
GO:5515	protein binding
GO:3723	RNA binding
GO:5215	transporter activity
GO:5198	structural molecule activity
GO:30528	transcription regulator activity
GO:16491	oxidoreductase activity
GO:3677	DNA binding
GO:30234	enzyme regulator activity
GO:8233	peptidase activity
GO:16874	ligase activity
GO:4672	protein kinase activity
GO:16779	nucleotidyltransferase activity
GO:16829	lyase activity
GO:4386	helicase activity
GO:16853	isomerase activity
GO:45182	translation regulator activity
GO:4871	signal transducer activity
GO:4721	phosphoprotein phosphatase activity
GO:3774	motor activity

Hamur Mayası canlısının proteinlerine GO seviyeli yapısının çok farklı seviyelerindeki terimleriyle özellik atfı yapılmıştır. GO seviyeli yapısında 13 seviye bulunmakla birlikte proteinlerin %92'si 2-7 seviyeleri arasındaki terimlerle iliştirilmiştir. Bu durum Şekil 3.3'deki dağılım grafiğinde görülmektedir. Hamur Mayası canlısının PPE ağında GO terimlerinin dağılımını anlamak amacıyla her seviyede yaklaşık ne kadar düğümün aynı terimle etiketlendiğini araştırdık. PPE ağı sadece belli bir sevideki GO terimleriyle etiketlendiğinde ortaya çıkan ağda her bir terime düşen düğüm sayısı grafiği Şekil 3.2'de verilmiştir. Bu grafikte skala logaritmiktir. PPE ağındaki bütün proteinler sadece 3. ve daha yukarıdaki seviyeden terimlerle etiketlendiğinde, aynı GO terimiyle etiketlenmiş ortalama protein sayısının 80 olduğu görülebilmektedir. Orijinal terimin değiştirilmemesi durumu "Orj" ile ifade edilmiştir.



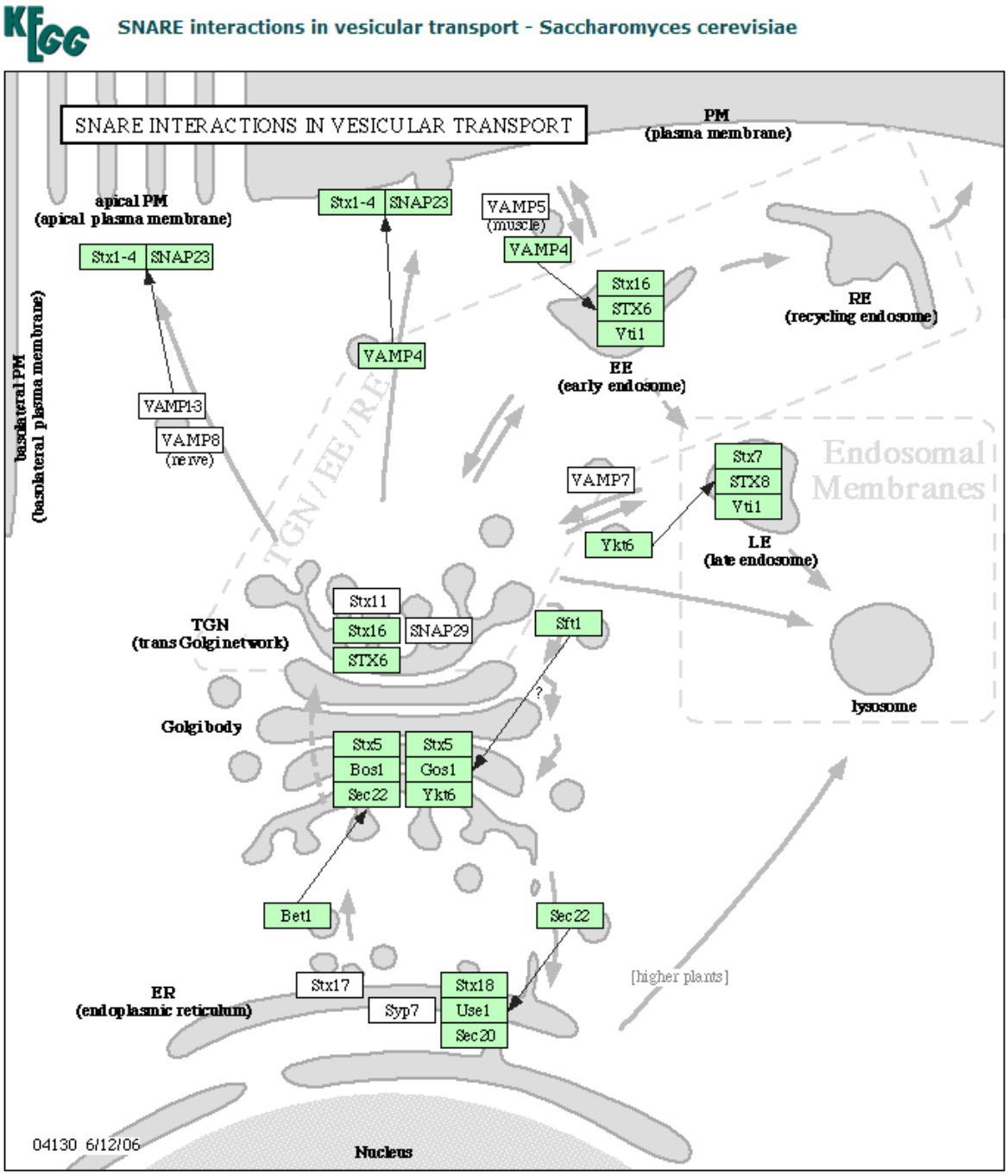
Şekil 3.2. Bir GO terimine düşen PPE ağındaki ortalama düğüm sayısı.



Şekil 3.3. Hamur Mayası organizmasının proteinlerinin tüm GO seviyelerine dağılımı (Her bir GO seviyesine düşen protein sayısı)

KEGG (Kyoto Gen ve Genom Ansiklopedisi) enzimatik patikalar ve biyolojik kimyasallar ile ilgili bir çevrimiçi veritabanları topluluğudur. Japonya İnsan Genomu Programı çerçevesinde biyolojik sistemlerin bilgisayarda temsili amacıyla genler, proteinler, biyokimyasal bileşikler ve etkileşimler/reaksiyonlar üzerine elle derlenmiş bilgileri içermektedir. Sistemler biyolojisi yaklaşımı çerçevesinde, KEGG veritabanı modelleme, benzeşme ve veri geri alımı için kullanılabilir. KEGG'deki patikalar literatürden ve uzman bilgisine başvurularak elle derlenmiştir. Genler ile ilgili bilgiler NCBI'dan, gen dizisi bilgileri RefSeq'den tedarik edilmektedir. KEGG patikalarının temsilinde kullanılan parçalar özetle şunlardır: Bir dikdörtgen gen ürününü (çoğunlula bir proteini bazen bir RNA bileşiğini) temsil eder. Küçük bir çember bir bileşiği, büyük bir çember ise başka bir patika haritasını temsil eder. Gen ürünleri arasındaki ilişki moleküler etkileşimin cinsine göre çeşitli oklarla ifade edilir. Örneğin fosforlama, fosforlamanın geri çevrilmesi, glikozlama ve metilleme

olayları için +p, -p, +g ve +m kullanılır. Gen ürünlerinin bileşikleri birbirine bitişik kutucuk kümeleri şeklinde ifade edilirler. Örnek olarak S. Cerevisiae organizmasında vasküler iletimle ilgili bir patika Şekil 3.4'de verilmiştir. Bu çalışmada, KEGG veritabanından XML biçiminde alınan patika verileri örüntülerin doğrulanmasında kullanılmıştır.



Şekil 3.4. Vasküler iletim'de SNARE etkileşimleri üzerine KEGG patikası.

Son yıllarda, kalitesi ve kapsamlılığı nedeniyle, Münih Protein Dizileri Bilgi Merkezi (MIPS)'nin Hamur Mayası'nın protein etkileşimleri ve bileşikleri pek çok çalışmada altın standardı olarak kabul edilerek kullanılmıştır (Guldener 2006, Yu ve ark. 2004). Yüksek miktarda çıktı veren deneylerden elde edilen verilerin yanında 1500'ü aşan sayıda protein ile ilgili deneysel etkileşim bilgisi uzmanlar tarafından süzgeçten geçirilerek derlenmiştir. Bu veriler web tabanlı bir araç ile protein etkileşimlerinin analizi ve görselleştirilmesi amacıyla kullanılabilmektedir. Pek çok bilimsel dergi yazarların buldukları yeni gen ve protein dizilerini umuma açık olan veritabanlarına eklemelerini zorunlu tutuyor olsalar da protein etkileşimleri, yönetimi, sinyal iletimi, hücresel yer ve fonksiyon çoğunlukla veritabanlarına iletilmemektedir. Bu bilgiler iletilseler bile uygun biçimde kaydedilmediklerinden sistematik yaklaşım ile kullanılmaları zor olmaktadır. MIPS ekibi metin madenciliği araçlarını PubMed üzerindeki seçilmiş yayınlar üzerinde kullanarak ve uzman süzgecinden geçirerek protein etkileşim işaretlemesini yapmaktadır. Bu işlemde kilit nokta etkileşimlerin uzman tarafından metinden çıkartılarak belirlenebilmesi ve kaynağın kayıt altına alınmasıdır. Bu verilere dayanarak etkileşimlerin tipe göre süzülmesi veya analizin belli bir kaynağa veya belli bir güven seviyesine göre yapılması mümkün olmaktadır. Büyük ve çok miktarda veri üreten deneylerin kendilerine has zayıf ve kuvvetli noktaları olmakla birlikte yüksek oranda yanlış pozitif üretebildiklerinden, bu verileri sıklıkla literatürden elle düzenlenerek alınmış verilerden ayırmak gerekmektedir. MIPS bu ihtiyacı gözeterek verilerini biçimlendirmiştir. MIPS veritabanındaki etkileşimleri güvenilirliği literatürdeki işaret sayısıyla orantılı olarak artmaktadır. Bir proteine ortalama 2,6 etkileşim, bir etkileşime ortalama 1,2 işaret, referans/yayın başına 2,5 etkileşim düşmektedir. Toplamda Hamur Mayası organizması için (Kasım 2007 itibariyle) 156 adet bileşikte 9518 adet etkileşim bulunmaktadır.

3.2. Metot

3.2.1. PPISpan örüntü bulma algoritması

Çizgelerdeki bütün sık geçen örüntüleri (alt-çizgeleri) bulan değişik algoritmaların hemen hepsi *apriori* adı verilen bir yaklaşımı sergiler ve tümü iki aşamalıdır: aday üretimi ve frekans sayımı. İlk adım olan "aday üretimi" aşamasında muhtemel bütün örüntüler birer birer üretilirler. Sonrasında "frekans sayımı" aşamasında her adayın desteği hesaplanarak geçerli olanlar yani örüntü olarak kabul edilenler belirlenir. Desteği belli bir eşik seviyenin üstünde olan adaylar sık geçen olarak kabul edilir. Bir aday alt-çizgenin büyük bir çizgedeki (örneğin gen ölçekli bir PPE ağındaki) frekansını hesaplamak altçizge eşitlik sınamasını defaatle kullanmayı gerektirir. Bu sınama işleminin NP-complete olduğu bilinmektedir (Ullman 1976, McKay 1981, Cordella 2004). Dolayısıyla yukarıda zikredilen algoritmaların çoğu birinci adım olan aday üretim aşamasında, fazlalık olan adayların üretimini engellemeyi hedeflemişlerdir. Yan ve Han (2002), gSpan algoritmalarıyla, bunu aday örüntülerin derinlik öncelikli arama temelli kadim etiketlerini hesaplayarak ve eşit etiketli olanları arama uzayından çıkarmak suretiyle sağlamaktadır.

Bu tezde ortaya koyduğumuz algoritma, gSpan algoritmasından uyarlanarak, GO etiketleri ile etiketlenmiş PPE ağlarında daha iyi performansla çalışmak üzere geliştirilmiştir. gSpan derinlik öncelikli arama (DÖA) kod ağacını kullanarak aday üretimini gerçekleştirir. Bu ağaçtaki her bir düğüm bir adayı temsil eder. Adaylar bu ağacın dikte ettiği sırayla ve şekilde üretilirler ve asgari destek için test edilirler. Bu asgari destek bizim problemimiz için algoritmaya girdi olarak verilen alt eşik frekans değeridir. Adayın PPE ağındaki kesişmeyen gömüntülerinin sayısı asgari destekten yani alt eşikten büyük veya eşitse, bu adayın desteğe sahip olduğunu ifade eder. Eğer bir adayın yeterli desteği yoksa bu aday ve onun DÖA kod ağacındaki çocukları göz ardı edilir. Benzer şekilde, eğer bir adayın DÖA kodu asgari ise yani daha önce DÖA kod ağacı tarafından üretilmiş ise, bu adayın kendisi ve çocukları da ilerideki aramalarda göz ardı edilir. Bu iki budama tekniği gSpan algoritmasını rakiplerine

göre daha verimli ve PPE ağları gibi değişik ağlara uyarlanmaya daha uygun kılmaktadır (Worlein 2005).

gSpan algoritması üstü kapalı olarak, bir adayın asgari DÖA kod hesaplamasının kendisi ve nesli için yapılacak toplam destek hesaplamasından daha az masraflı olacağını varsaymaktadır. Bu durum PPE ağları için özellikle etiket başına düşen düğüm sayısı düşük olduğundan ve sadece belirgin şekilde yüksek frekanslı örüntülerle ilgilendiğimizden genellikle geçerli değildir (bkz. Bölüm 4. Araştırma Sonuçları). Algoritma DÖA kod ağacının derinliklerine doğru ilerlerken, asgari DÖA kodu hesaplamasının maliyeti giderek artmakta iken destek hesaplamasının maliyeti pratikte sabit kalmaktadır. Çok büyük olasılıkla bu destek hesaplaması başarısız olacağı için (yanlış artının budanması: örüntünün desteğinin asgari destekten az çıkması) yanlış pozitifin budanmasının maliyeti asgari destek hesaplamasının maliyetinden az olacaktır. Dolayısıyla bu çalışmadaki algoritmamızda bir örüntünün destek hesaplamasının maliyetinin asgari DÖA kodu hesaplamasından az olup olmayacağını tahmin eden ve bu tahmine göre asgari DÖA kodu hesaplamasını atlayan bir fizibilite fonksiyonu kullandık.

Bu çalışmada PPE ağlarında sık geçen örüntüleri bulmayı hızlandırmak amacıyla kenar ve düğümler için özgün bir sözlüksel sıralama önerilmektedir. Düğümler için düğüm etiketlerinin frekansına bağlı olan (yüksek frekanslı etiket düşük frekanslının önünde gelmektedir) bir sıralama tanımlanmaktadır. Benzer şekilde kenarlar için de frekans tabanlı bir sıralama önermekteyiz. Bir kenar bir çift etiket ile temsil edilmekte, sıralamada düşük frekanslı çift yüksek frekanslının önünde gelmektedir. gSpan algoritması, bir kenarın kökü olduğu DÖA ağacında araması bittiğinde, o kenarı aramadan çıkarmaktadır, yani diğer kenarların kök olduğu DÖA ağacında kullanmamaktadır. Bunun yanında aynı kenar, içinde geçtiği bütün adaylar denendiği için çizgeden de çıkarılarak hız artırılmaktadır. Dolayısıyla az frekanslı kenarları aramanın erken evrelerinde devreden çıkarmak (PPE ağından silmek) daha sonraki evrelerde daha sık geçen kenarlar için yapılacak olan yanlış pozitifleri budama işlemleri için harcanacak zamanı azaltacaktır.

PINSpan algoritmamız gSpan'den farklı olarak en büyük örüntüleri bulabilmektedir. En büyük örüntü başka bir sık geçen örüntünün alt çizgesi olmayan örüntüdür. PPISpan algoritmamız *PPISpan (G, L, minSup)* ve *SubGraphs(s, minSup,*

ext) imzalı başlıkları altında aşağıda iki parça halinde verilmiştir. Algoritmanın ana parçası olan PPISpan her bir kenar için çalışan döngüyü içermektedir. Algoritmanın girdileri PPE ağını temsil eden *G* çizgesi, etiketleme yapmak istediğimiz hedef GO seviyesi olan *L* ve en düşük destek eşiği olan *minSup*'dır.

Algoritma: PPISpan (G, L, minSup)

1: G ağındaki düğümlerin etiketleri olarak GO kütüphanesinin L seviyesindeki terimleri ata.
2: S ← G'deki bir kenarlı tüm sık geçen çizgeler (frekansa dayalı sözlüksel sıralama ile)
3: Her bir e ∈ S kenarı için (artan frekans sırasına göre) yap
    3.1: SubGraphs(e, minSup, e)
    3.2: e'yi G'den çıkar

Algoritma: SubGraphs(s, minSup, ext)

1: Eğer (feasible(s, ext)) ise
2: Eğer s'in DÖA kodu onun asgari DÖA koduna eşit değilse
3: dön
4: C ← s'in tüm çocuklarını üret (ext isimli kenarı büyüterek)
5: azami ← doğru
6: Her bir c ∈ C için (DÖA sözlüksel sıralamaya göre) yap
7: Eğer destek(c) ≥ minSup ise
8: SubGraphs(c, minSup, c.ext)
9: azami ← yanlış
10: Eğer (azami) ise
11: s'i çıktı olarak ver

Derinlik öncelikli arama kod ağacının büyümesi kadim etiketleme kurallarına göre yapıldığından, bu ağaçtaki bir düğümün çocuğu kendisinden bir fazla kenara sahip olmak zorundadır. Dolayısıyla bir adayın gömüntüleri, bu adayın çocuklarının gömüntülerinin aranmasında kullanılabilir. PPISpan algoritması ana örüntünün

gömüntü bilgilerini saklayarak gerektiğinde bunları çocukların desteklerinin hesaplanmasında kullanmakta, böylelikle çok sayıda alt-çizge eşleştirme işlemlerinden kurtulunmuş olmaktadır. Her bir örüntü ile birlikte bu örüntünün gömüntülerinin çizge içindeki yerleri ve her bir gömüntü düğümünün eşleştirildiği örüntü düğümü saklanmaktadır. Saklanan gömüntü bilgileri çizge eşleştirme işlemlerini oldukça kolaylaştırmakta ve hızlandırmaktadırlar çünkü ana örüntü için başarılı olmuş bir çizge eşleştirme işlemi çocuk örüntü için tekrar edilmek zorunda kalmamaktadır. Algoritmada *s* ve *ext* parametrelerini alan bir ikili fizibilite fonksiyonu tanımlanmaktadır. Bu fonksiyon, *s*'deki kenarların ortalama frekansları ile bu kenarların frekanslarının standart sapmasının toplamı *ext*'nin frekansından küçükse doğru döndürmektedir. Bir başka deyişle *ext*'nin frekansı *s*'deki kenarların ortalama frekanslarından bir standart sapma miktarı kadar veya daha büyükse, *s*'in olabilirliğine karar verilip asgari DÖA kodu hesaplanır. Aksi takdirde bu hesaplama atlanır.

Algoritmamızı dilenilen GO seviyesinde çalışacak şekilde esnek yaptık. GO verilerinin indirildiği tarih itibariyle, maya canlısının proteinleri 14 farklı GO seviyesinden terimler ile eşlenebilmektedir(bkz. Bölüm 3.1. Materyal). Dolayısıyla sadece orijinal GO terimleri üzerinden yapılacak bir arama daha yukarı seviyelerdeki terimlerden oluşmuş olabilecek muhtemel örüntüleri bulmamıza yaramayacaktır. Bu nedenle algoritmanın birinci adımında hedeflenen GO seviyesinin tayini yapılır. Ağdaki bütün düğümlerin etiketleri olarak bu hedef seviyeden terimler atanır. Bu şekilde algoritmamız hedef olarak belirtilen GO seviyesinde örüntü araması yapmaya hazır hale gelir. Bir orijinal *t* teriminin yerine düğüme atanacak yeni hedef terimler şu şekilde saptanır: Şayet $t<=L$ ise orijinal terim korunur. Aksi takdirde ($t>L$ ise), kök terim ile *t* arasındaki patikalarda *L* seviyesinde bulunan tüm terimler atanır. Nitekim materyal bölümündeki Şekil 3.1'den de anlaşılacağı üzere GO seviyeli yapısında bir terim ile kök terim arasında birden fazla patika olabilir. Örneğin, $L=3$ ise, 4907 terimli bir düğüm (protein)'e üçüncü seviyedeki 4888 ve 19955 terimleri atanacaktır. Aynı durumda etiketi 5515 olan bir proteinin etiketi değişmeyecektir. Dikkat edilmelidir ki, bu durum bir proteinin birden fazla etiketinin olmasına imkân tanımaktadır ve algoritmamız da buna uygun olarak örüntü ararken bir proteinin bütün etiketleriyle örüntüyü eşleştirmeye çalışmaktadır.

3.2.2. Bir örüntünün istatistiksel önemi

Bir örüntünün başka örüntülere nazaran yüksek desteğe sahip olması tek başına bu örüntünün önemli olduğunu ispat etmeye yetmemektedir. Bunun yanında farklı desteklere ve farklı büyüklüklere sahip örüntüleri karşılaştırabilmek için bir örüntünün istatistiksel önemini ifade eden bir ölçüte ihtiyaç vardır. Biz bu amaçla örüntünün Bonferroni doğrulaması yapılmış Z-puanını hesapladık. Bunu 100 farklı rastgele ağda benzer örüntüleri sayarak gerçekleştirdik. Rastgele ağlar orijinal PPE ağı ile aynı derece dağılımına ve protein etiket frekans dağılıma sahip olacak şekilde üretildiler. Z-puanını (standart sapmalar cinsinden), bir örüntünün PPE ağındaki desteği ile aynı örüntünün benzerlerinin rastgele ağlardaki ortalama destekleri arasındaki uzaklık olarak hesapladık.

Rastgele ağlar, gerçek PPE ağındaki düğüm dereceleri ile aynı düğüm derecelerine sahip olacak şekilde markov zinciri algoritmaları (Kannan 1999, Maslov 2002) kullanılarak üretildiler. Bu yöntemde önce gerçek ağın bir kopyası oluşturulur ve bu kopya yeterince farklı olduğuna kanaat getirinceye dek defaatle değişikliğe uğratılır. Her bir değişiklik rastgele seçilen iki düğüm çiftinin aralarındaki etkileşimlerin karşılıklı alış-verişi şeklinde gerçekleştirilir. Aralarında etkileşim olan X1-Y1 çifti ile X2-Y2 çifti değişiklikten sonra X1-Y2 ve X2-Y1 etkileşimlerine sahip hale gelirler. Bir başka deyişle, X1-Y1 ve X2-Y2 etkileşim çiftleri, X1-Y2 ve X2-Y1 etkileşim çiftleriyle değiştirilmiş olurlar. X1-Y2 ve X2-Y1 etkileşimlerinden biri hâlihazırda varsa değişim yapılmaz. Bu değişiklikler gerçek ağın kopyası yeterince rastgele hale gelinceye kadar tekrarlanır. Biz uygulamamızda, Hamur Mayası organizmasının PPE ağları için 40000 değişiklik yaparak rastgele ağları ürettik.

Bir sonucun şans eseri olarak ortaya çıkması ihtimali çok küçük ise, bu sonuç "istatistiksel olarak önemli" olarak nitelendirilir. Şans kelimesi düzen, amaç, sebep ve belirliğin olmadığı süreçleri ifade etmek için kullanılır. Bir sonucun istatistiksel olarak farklı bir sonuç olması, farklı olduğunu gösteren istatistikî delillerin olduğu anlamına gelmektedir, farkın büyük olduğu veya biyolojik açıdan önem arz ettiği anlamına gelmemektedir. Bir denemenin önem seviyesi, sıfır varsayımı (H0) doğru olduğu halde onu reddetmeye karar verme ihtimali olarak tanımlanır. Bu karar tip-1

hata ya da yanlış pozitif belirleme olarak adlandırılır ve p-değeri'ne bakılarak verilir. p-değeri, eğer sıfır varsayımı doğru ise, bunun yanlışlıkla reddedilme ihtimalidir. P-değeri önem seviyesinden az ise sıfır varsayımı reddedilir. P-değeri ne kadar küçük olursa sonucun istatistiksel önemi o kadar çok kabul edilir.

Önem seviyesi α ile simgelenir. En çok kullanılan önem seviyeleri sağlık bilimlerinde genellikle α=0,1 ,mühendislik bilimlerinde genellikle α=0,05 ve sosyal bilimlerde genellikle α=0,01'dir. Eğer bir önem sınaması α'dan daha küçük bir p-değeri verir ise, bu durumda sıfır varsayımı reddedilir ve deney sonucu "istatistiksel olarak önemli" olarak adlandırılır. Örneğin bir kişi "bu sonucun tesadüfi olarak ortaya çıkmasının binde bir ihtimalle olacağını" belirtir ise bununla %0,1 istatistiksel önem seviyesi ima edilmiş olur. Bu seviye azaldıkça delilin gücü de artar.

Sıfır varsayımı, ortalamalar ve dağılımlar arasında bir farkın (veya ilişkinin) olmadığını savunur. O nedenle fark-yok varsayımı olarak da adlandırılır. Varsayım sınaması, varsayımın örnekten elde edilen bilgilere bağlı olarak belirli bir hata payı ile doğrulanmasına denir. Anlamlılık testi (manidarlık testi), istatiksel varsayımın (sıfır varsayımı) yanlışlanıp yanlışlanmadığını anlamak için uygulanır. Sıfır varsayımı, ortalamalar ve dağılımlar arasında bir farkın (veya ilişkinin) olmadığını savunur. Karşı hipotez, araştırma varsayımı (H_a), ortalamalarda ve dağılımda bir fark olduğunu savunur. Araştırma varsayımı, bizim evrendeki ortalamalarda ve dağılımda bir fark olduğu şüphe ve gözlemimize dayanır. Varsayım sınamasında, evrenden alınacak sonsuz tesadüfî örneklemin bir normal dağılım eğrisinin altında (ana ortalamanın etrafında, 1.96 standart sapma uzaklıkta) toplanacağı kabul edilir. Seçtiğiniz örneklemin, normal dağılım eğrisinin sağ ve solundaki red bölgelerinde olup olmadığına bakılır. O bölgelerin dışında (kenarlarda) yer alıyorsa, farklılıklar tesadüfî olabilir. Yoksa fark artık tesadüfî değildir (rastgele örneklem hatası değildir). Dolayısıyla, sıfır varsayımının reddedilme ölçütü α, H_0 gerçekte doğru olduğunda, onun yanlışlıkla reddedilme ihtimalidir. Araştırmalarda yanlış bir iddiada bulunmamak ve sıfır varsayımını korumak için α seviyesi küçük tutulur.

Bu çalışmamızda kullandığımız sıfır varsayımı bulduğumuz bir örüntünün gözlemlenen frekansının şans eseri ortaya çıkmış olduğu üzerinedir. Benzer özelliklere sahip başka rastgele herhangi bir örüntünün de aynı veya daha fazla sayıda gömüntüye sahip olabileceğini, dolayısıyla bu örüntünün belirtilen frekansla

ortaya çıkmasının şansla izah edilebileceğini ifade eder. Bu durumda karşı varsayımımız, elimizdeki örüntünün kaydedilen frekansının rastgele ağlardaki ortalamasından farklı olduğu biçiminde olacaktır. Bu da tabii olarak bizim araştırma varsayımımızdır.

Fizik ve mühendislik gibi bazı bilimlerde yaygın olan istatistikî önemi ifadeye yarayan bir başka ölçüt de Z-puanıdır. Z-puanı Gaus (Normal) dağılımının standart sapmasını ölçü birimi olarak kullanılarak ifade edilir ve α'ya hata fonksiyonu (erf) aracılığıyla dönüştürülebilir:

$$\alpha = 1 - \mathrm{erf}\left(Z/\sqrt{2}\right)$$

Z-puanı ile önem seviyesi arasındaki ilişki Şekil 3.6'deki standart normal dağılımın ihtimal dağılım fonksiyonu üzerinde ifade edilmiştir. Bu fonksiyon eğrisinin altında ve Z-puanı değerinin sağında kalan alan α olup, toplam dağılım fonksiyonu (cdf) kullanılarak hesaplanır. Grafikte, kırmızı/taralı kesimin alanı olarak ifade edilen α=1-cdf(Z) olduğundan, Z-puanı değeri Z=2,33' e karşılık gelen önem seviyesi α=0,01 olmaktadır. cdf ve ters cdf fonksiyonlarının kapalı formülleri olmayıp seri açılımı ile yakınsanabilen erf ve erf^{-1} fonksiyonları cinsinden ifade edilirler.

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

$$\mathrm{erf}^{-1}(z) = \sum_{k=0}^{\infty} \frac{c_k}{2k+1} \left(\frac{\sqrt{\pi}}{2} z \right)^{2k+1},$$

Burada $c_0 = 1$ ve

$$c_k = \sum_{m=0}^{k-1} \frac{c_m c_{k-1-m}}{(m+1)(2m+1)} = \left\{ 1, 1, \frac{7}{6}, \frac{127}{90}, \ldots \right\}.$$

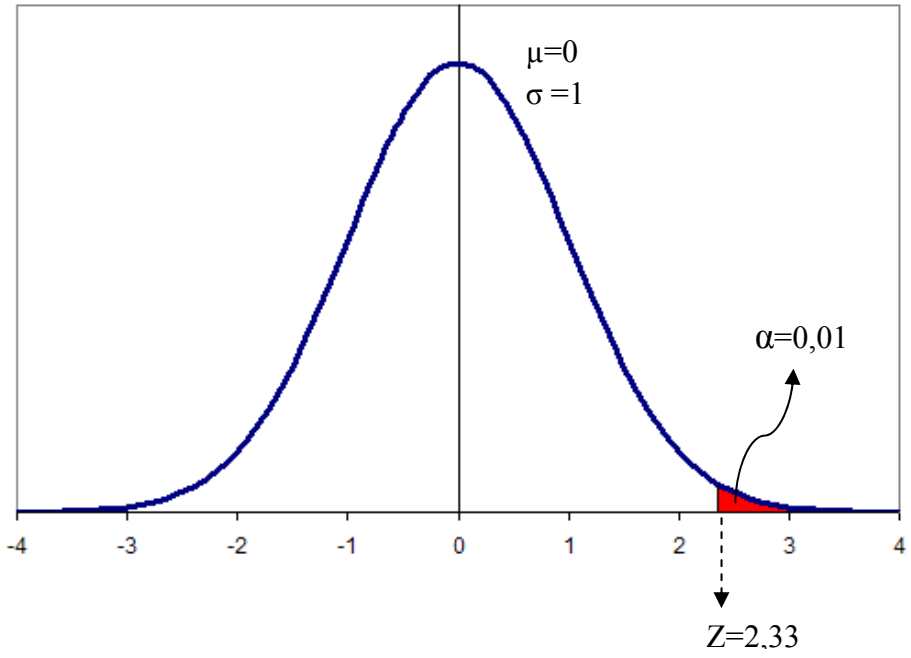
Normal dağılımın toplam dağılım fonksiyonu cdf'nin erf cinsinden bağıntısı

$$\mathrm{cdf}(x, \mu, \sigma) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

olup standart normal dağılım için μ=0 ve σ=1'dir. Cdf'nin tersi ise;

$$\mathrm{cdf}^{-1}(p) = \sqrt{2}\, \mathrm{erf}^{-1}(2p - 1) \quad .$$

Çalışmamızda, cdf^{-1} fonksiyonu, Z-puanını Bonferrroni düzeltmesine tabi tutmak için kullanılmıştır (Montgomery 2003, Mendenhall 1994).



Şekil 3.6. Normal dağılımın yoğunluk fonksiyonunun grafiği üzerinde Z ile α'nın ilişkisi.

Yapılan örüntü bulma deneylerinin sonucunda istatistiksel önemi ölçülmesi gereken çok sayıda (genellikle yüzlerce) örüntü ortaya çıkmaktadır. Bundan dolayı yüzlerce istatistiksel varsayımın sınanması gerekmektedir. Bu tür durumlarda örneğin α=0.05 kullanılması Tip-1 hatalara karşı yeterince koruma sağlamamaktadır. Bu çalışmada yapılabilecek Tip-1 hata, önemli olmadığı halde bir örüntüyü önemli kabul etmektir (Gardiner 1997). Bu durum şöyle gerçekleşmektedir: Bir deneyin sonucunda ortaya çıkan ayrı varsayımların sayısı arttıkça bütün bir deney için geçerli olacak gerçek α değeri (bizim kararlaştırdığımız değer ne olursa olsun)

büyümektedir. Sınadığımız varsayım sayısına bağlı olarak gerçek α seviyesi yaklaşık olarak şu şekilde hesaplanabilir:

Gerçek α-seviyesi=1-(1-α)^k

Burada *k* deney sonucunda sınanması gereken ayrı varsayım sayısı, yani bizim çalışmamızda, bulunan örüntü sayısıdır. Örneğin α'yı 0,05 seviyesine kurduğumuzu ve 20 tane örüntü için istatistiksel varsayımı sınayacağımızı düşünürsek, doğru olmadığı halde bunlardan birinin önemli olduğunu iddia etme ihtimalimiz 1-(1-0,05)^{20}=0,64'dür. Bir başka deyişle Tip-1 hatası yapma (boş varsayımı uygun olmadığı halde reddetme) ihtimalimiz %64'dür. Bu oldukça yüksek bir ihtimaldir ve daha ötesi hangilerinin hatalı hangilerinin doğru olduğunu bilme imkanımız yoktur. Öyleyse 20 tane sınamanın arasında bir tane önemli sonuç gördüğümüzde bunun gerçekten önemli olduğuna güvenimiz ne derecede olacaktır? Bir tanesinin önemli çıkma ihtimali, yani deneyin sonuçları bütüncül olarak değerlendirildiğinde Tip-1 hatası oranı %64 gibi arzuladığımız %5'den çok yüksek bir rakam olarak karşımıza çıkmaktadır.

Bu sorunu çözmek için birkaç farklı yol vardır. Bunlardan bir tanesi çok değişkenli istatistiksel yöntemler kullanarak α büyümesi sorununa yaklaşmaktır. Bu bütün bağımlı değişkenler üzerinde eş zamanlı sınamalar yapmak şeklinde olabileceği gibi iştigal edilen değişken sayısını istatistiksel yöntemlerle düşürme şeklinde de yapılabilir. Bu yöntemler genellikle karmaşık istatistiksel problem analizi gerektirir.

Bir başka yaklaşım Bonferroni'ye atfedilmiştir. Bonferroni α-büyümesi sorununu dizginlemek için basit bir yol önermiştir. Bir çalışmanın önceden belirlenen bir α seviyesi(örneğin 0,05) ile yürütülmesi için, bu basitçe (k tane) önem sınaması sayısına eşit bir şekilde bölünür. Böylece her bir önem sınaması için yeni α seviyesi

$$yeni\ \alpha\ seviyesi = \frac{Bütün\ bir\ çalışma\ için\ arzu\ edilen\ \alpha\ seviyesi}{k}$$

şeklinde olur. Buna göre önceki 20 önem sınamalı örnek çalışma için yeni α seviyesi 0,05/20=0,0025 olacaktır. Böylece çoklu önem sınamasından kaynaklanan α seviyesi büyümesine karşı daha iyi korunmuş olunur. Yeni α seviyesini kullanarak yapacağımız önem sınamalarının sonucunda karşılaşacağımız deneyin bütününe ait

büyümüş α seviyesi $1-(1-0{,}0025)^{20}=0{,}0488$ olacak, dolayısıyla bütün bir çalışma için arzuladığımız genel α seviyesine çok yakın olacaktır. Bu şekilde α seviyesini düzeltme yöntemine Bonferroni düzeltmesi yöntemi adı verilir. Bu çalışmamızda α=0,01 seviyesini arzulandığından, Bonferroni düzeltmesi yöntemi kullanılmamış olsaydı, örneğin 100 örüntülü bir deney için gerçek α seviyesi $=1-(1-0{,}01)^{100}=0{,}64$ olarak gerçekleşecekti. Bonferroni düzeltme yöntemiyle bu engellenmiş oldu. Ancak bu durumda, önemli fark olduğu halde bunu bulamama sorunu ile yani Tip-2 hatası yapma tehlikesi ile karşılaşılır. Bu örnek için, p-değeri 0,0025'den küçük örüntü bulmak zor olabilir, dolayısıyla önemli olabilecek bir bulgu atlanabilir. Bonferroni yönteminin en çok eleştirilen yanı tekil sonuçlar için yüksek standart kullanmak suretiyle çok muhafazakâr bir yaklaşım sergilemesidir. Bu durum yukarıda belirttiğimiz gibi Tip-2 hatası ihtimalini artırmaktadır. Dolayısıyla bu çalışmada ortaya çıkarılan önemli örüntülerin bütün önemli örüntüleri kapsamayabileceği, bazı örüntülerin önemli oldukları halde sonuç listesinde yer almayabilecekleri ifade edilmelidir.

3.2.3. Örüntü oylarıyla etkileşim kestirimi

Bu bölümde bulunan örüntülerin etkileşimleri kestirmede kullanılabileceklerini göstereceğiz. Bu amaçla geliştirdiğimiz, örüntülerin aday kenarları oylamasına dayalı yöntemimizi açıklayacağız.

Bir $G=(V,E)$ çizgesi bir V düğüm kümesi ve bir $E \subseteq V \times V$ kenar kümesinden oluşur. Bir kenar bir düğüm çiftinden oluşur. Bir PPE ağı proteinleri V kümesine, etkileşimleri ise E kümesine karşılık gelecek şekilde ifade edilir. Her bir düğüm GO kütüphanesindeki çeşitli seviyelerde olabilen GO terimleriyle etiketlenir. Bir örüntü G içinde sıkça geçen bir alt-çizgedir. Bir örüntünün G içinde bulunan eşi olduğu alt-çizge gömüntü olarak adlandırılır. Bir başka deyişle bir örüntünün ağda rastlanan görüntüleri (kopyaları), onun gömüntü kümesidir.

$Oy_verir(G,g,e_{i,j})$ fonksiyonu, g örüntüsünün i ve j düğümleri arasındaki kenar için oy verip vermediğini belirten bir ikili fonksiyondur. Muntazam olarak;

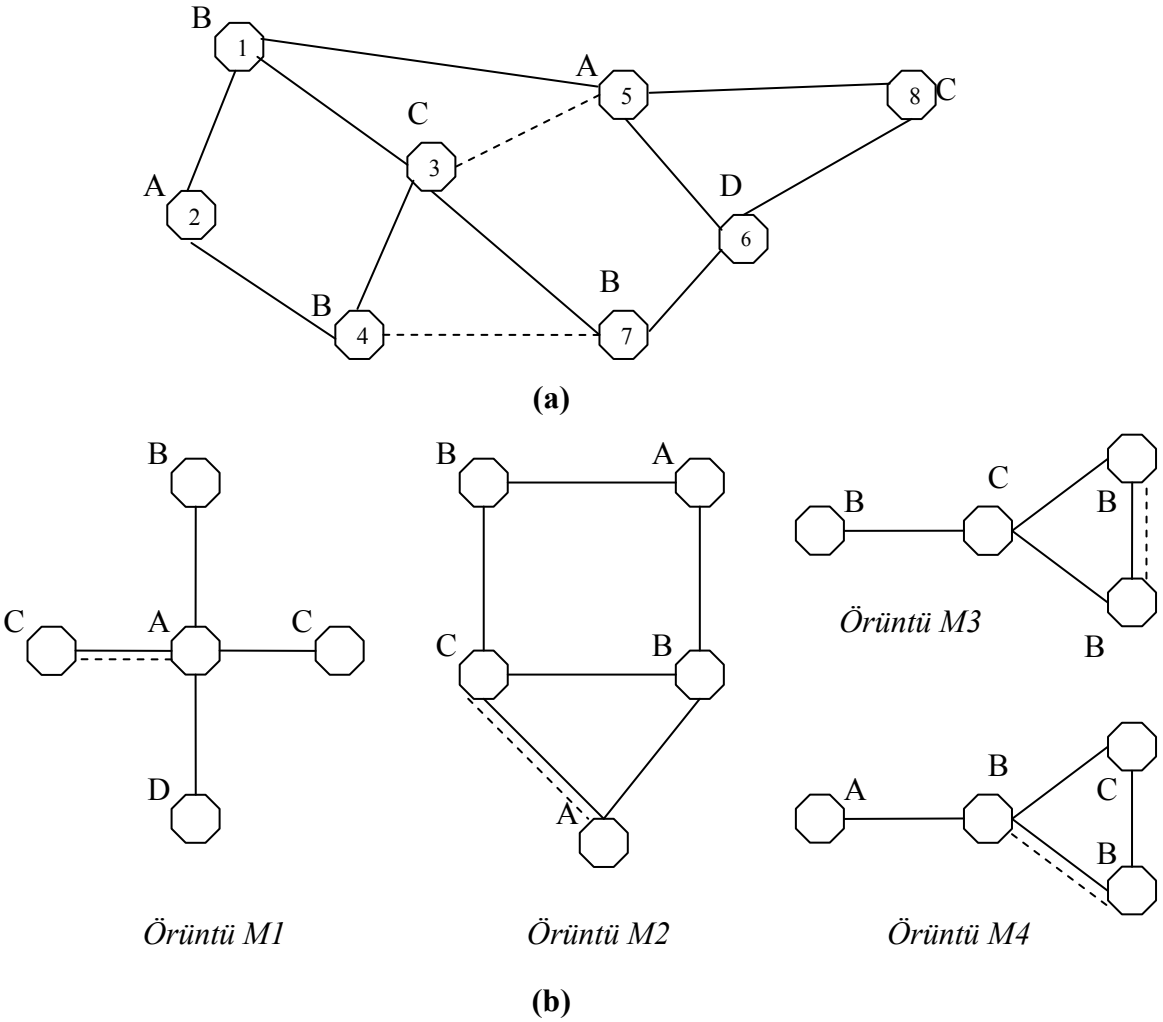
$$Oy_verir(G,g,e_{i,j}) = \begin{cases} 1, & \exists s \mid \ s \subseteq G \cup e_{i,j} \wedge \ s \cong g \ \wedge \ e_{i,j} \in s \\ 0, & Aksi\ durumda \end{cases}$$

Oy_verir(G,g,$e_{i,j}$)=1 ise g örüntüsünün G∪$e_{i,j}$, nin içinde en az bir gömüntüsü $e_{i,j}$'yi kapsamaktadır. $s \cong g$, *s* ve *s*'nin eş çizgeler olduğunu ifade eder.

Yapacağımız kestirimler için iki güven ölçütü tanımlıyoruz. Birinci ölçüt sade oy vermeye dayalı iken, ikincisi ağırlıklı oy verme kuralına bağlıdır. Birinci ölçütte sadece bir aday kenarın farklı örüntülerden toplam aldığı oyu sayıyoruz. Buna göre, M örüntü kümesinin oylarıyla, *i* ile *j* düğümleri arasındaki aday kenarın doğru bir kestirim olduğuna olan güvenimizi ifade eden $C_{i,j}$ değeri aşağıdaki şekilde hesaplanır:

$$C_{i,j}^{sade}(M) = \sum_{\forall g \in M} Oy_verir(G,g,e_{i,j})$$

Aynı aday etkileşime oy veren iki örüntü Şekil 3.7.b 'de verilmiştir. Burada M1 ve M2 örüntüleri kesik çizgi ile belirtilen $e_{3,5}$ kenarı için oy vermektedirler. Kesikli çizgi Şekil 3.7.a'da verilmiş olan asıl PPE ağında bulunmamaktadır, ama örüntü tarafından ima edilmektedir. $C_{3,5}$(M) = 2 'dir çünkü M = {M1, M2, M3, M4} iken *Oy_verir(G,M1,$e_{3,5}$)* = 1, Oy_verir (G,M2, $e_{3,5}$) = 1, *Oy_verir(G,M3, $e_{3,5}$)* = 0 ve *Oy_verir(G,M4, $e_{3,5}$)* = 0 'dir.



Şekil 3.7. Örnek G ağının örnek bir parçası.

Ağılıklı oy verme mekanizmasında oyların iliştirilmiş ağırlıkları vardır. Bu ağırlıklar gömüntülerin örtüşmeleri ve örüntülerin benzeşmezlikleri kullanılarak hesaplanır.

$$C_{i,j}^{ag}(M) = \frac{1}{N_{i,j}} \sum_{\forall p,q \in M} D_{p,q}.(1 - O_{p,q,e_{i,j}}).Oy_verir(G,p,e_{i,j}).Oy_verir(G,q,e_{i,j})$$

$D_{p,q}$, p ve q örüntüleri arasındaki benzeşmezlik seviyesidir. Bu değer [0,1] aralığında olup, p ve q'nun benzerlikleri artarken azalmaktadır. $p=q$ ise $D_{p,q}=0$'dır. $O_{p,q,e_{i,j}}$, p ve q örüntülerinin $e_{i,j}$ kenarını kapsayan gömüntülerinin örtüşme oranıdır. Bu örtüşme oranı, her ikisi tarafından kapsanan kenar sayısının, ikisinin kapsadığı tüm kenarların sayısına bölümü olarak hesaplanır. $N_{i,j}$, $e_{i,j}$ kenarına oy veren toplam örüntü sayısıdır. $1/N_{i,j}$ ifadesi benzer ve büyük oranda birbiriyle örtüşen örüntüler tarafından verilen oyları amorti etmeye yaramaktadır. $D_{p,q}$ ve $O_{p,q,e_{i,j}}$ değerleri ile ilgili daha ayrıntılı bilgi 3.2.3.2. Örüntülerin benzeşmezlik ve örtüşme ölçütleri başlıklı gelecek bölümde bulunmaktadır.

Ağırlıklı oy verme mekanizması, bir aday kenarın adaylığını kuvvetlendirmesi açısından, iki benzer olmayan örüntünün aday kenara oy vermesinin, iki benzer örüntünün aynı aday kenara oy vermesinden genel anlamda daha önemli bir işaret olduğu öngörüsüne dayanmaktadır. Aynı şekilde, aday kenara oy veren iki örüntünün gömüntüsünün örtüşme oranının yüksekliği, bu ölçütlere göre yapacağımız kestirimlere olan güvenimizi artırmış olacaktır. Bu durum Şekil 3.7'de görülebilmektedir. Burada, M1 ve M2, $e_{3,5}$ için, M3ve M4 ise $e_{4,7}$ için oy vermektedir. Sade oylama yöntemiyle $C_{3,5}^{sade}(M)=2$ ve $C_{4,7}^{sade}(M)=2$ olduğundan bu adaylar($e_{3,5}$ ve $e_{4,7}$) güven sıralamasında aynı derecede olmaktadırlar. Ağırlıklı oy verme ölçütü uygulandığında ise $C_{3,5}^{ag}(M)=0.15$ ve $C_{4,7}^{ag}(M)=0.05$ olduğundan $e_{3,5}$ sıralamada $e_{4,7}$ ye göre daha önde olmaktadır.

Şekil 3.7'deki her bir düğüm (A, B, C veya D ile) etiketlidir. 3 numaralı ve 5 numaralı düğümler arasındaki kenar ($e_{3,5}$), M1 ve M2 örüntüleri tarafından kestirilmektedirler. 4 numaralı ve 7 numaralı düğümler arasındaki kenar ($e_{4,7}$), M3 ve M4 örüntüleri tarafından kestirilmektedirler. Dikkat edilmeli ki $e_{4,7}$ ve $e_{3,5}$ G'nin

parçası olmayıp, örüntüler tarafından ima edilmektedirler, kastedilmektedirler, kestirilmektedirler. Şekil (b)'de bazıları G'deki bazı aday kenarlara oy veren 4 adet örüntü bulunmaktadır. Bunların kesikli kenarları Şekil (a)'daki G'de oy verdikleri kenarlara karşılık gelen kenarlardır.

3.2.3.1. Etkileşim kestirimi algoritması

Yukarıda açıklanan oy vermeye dayalı mekanizmayı kullanarak etkileşim tahmini yapan özgün bir algoritma öneriyoruz. Bu algoritma verilen bir PPE ağını önemli örüntüler için taramaktadır. Örüntüler arasından belli örüntü özelliklerine göre bir alt kümesi seçilir. Bazı özelliklere sahip örüntüler diğerlerine nazaran daha iyi kestirim performansı göstermektedirler. Dolayısıyla algoritmamız, tüm örüntüleri kullanmak yerine daha uygun bit alt kümesiyle etkileşim adaylarını aramaktadır. Seçilen örüntülerin verdikleri oylarla adaylar belirlenir. Son olarak algoritmamız, her bir aday etkileşim için önceki bölümde açıklanan biçimde güven puanlarını hesaplayarak adayları sıralamaya sokmaktadır.

Algoritmamızda kullanılan terimler aşağıda açıklanmıştır.

Terim	**Tanım**	**Karşılık gelen Eşik**
m.frekans	m örüntüsünün ağdaki örtüşmeyen gömüntüleri sayısı	F
m.z-puan	m örüntüsünün Z-puanı	Z
m.kenar	m örüntüsündeki kenar sayısı	N
m.oy	m örüntüsünün oy verdiği farklı aday kenar sayısı	T
K	Aday listesinin üstünden seçilecek yüzde	K

Aşağıda Etkileşim Kestirimi Algoritmamızın adımları verilmiştir.

Etkileşim Kestirimi Algoritması:

1. PPE ağındaki mümkün olan tüm proteinlere (düğümlere) GO kütüphanesinden terimler ata.
2. PPE ağında etkileşim örüntüleri ara. Bütün sık geçen etkileşimlerin kümesi M olsun.
3. Her m∈M için m.frekans, m.z-puan ve z-kenar'ı hesapla. M'deki örüntüler arasından m.frekans>F, m.z-puanı>Z, m.kenar>N ve m.oy<T olanları seç. Diğerlerini M'den çıkar.
4. PPE ağındaki etkileşmeyen her bir protein çifti için;
   $C_{i,j}$(M) güven değerini hesapla
   $C_{i,j}$(M) güven değeri sıfırdan büyük olanları seç
5. Aday etkileşimleri güven değerlerine göre büyükten küçüğe sırala.
6. Sıralı etkileşimlerden en üstte olan %K'lık kesimini algoritmanın çıktısı olarak rapor et.

Bir örüntünün Z-puanının hesaplanması ile ilgili olarak Bölüm 3.2.2'ye bakınız.

3.2.3.2. Örüntülerin benzeşmezlik ve örtüşme ölçütleri

Örüntü benzeşmezliği ölçüsü olarak, en büyük ortak alt-çizge kavramına dayalı olan bir çizge uzaklık ölçütü kullanılmıştır (Bunke ve Shearer 1998). İki boş olmayan örüntü *p* ve *q* arasında tanımlı uzaklık $D_{p,q}$:

$$D_{p,q} = 1 - |mcs(p,q)| / max(|p|,|q|)$$

Burada $|M_i|$, M_i örüntüsünün içindeki düğüm sayısı olup, mcs(p,q) ise *p* ve *q* örüntülerinin en büyük ortak alt-çizgeleridir. Muntazam olarak;

$$mcs(p,q) = \{s \subseteq p,q \mid \neg\exists t, t \subseteq p,q \wedge |t| > |s|\}$$

Hem *p* ve hem de *q*'nun alt-çizgesi olan bütün çizgeler içindeki en fazla sayıda düğüm olan çizge *mcs(p,q)* olarak tanımlanır. Şekil 3.7.b'deki M1 ve M2 örüntüleri

için |M1|=5, |M2|=5 ve mcs(M1,M2)=3'dür. Dolayısıyla M1 ve M' arasındaki benzeşmezlik $D_{M1,M2}$=0.4 . Aynı şekilde $D_{M3,M4}$=0.25'dir .

Örüntüler p ve q arasındaki örtüşmenin ölçütü olarak *e* kenarını kapsayan gömüntülerinin örtüşme oranı ölçüt olarak alınmıştır (ve $O_{p,q,e}$ ile gösterdik). Her iki gömüntünün *i* ve *j* düğümleri arasındaki *e* aday kenarını kapsadığını varsayarak, $O_{p,q,e}$ şöyle tanımlıdır:

$$O_{p,q,e} = \left|K(göm(p,G,e) \cap göm(q,G,e))\right| \Big/ \left|K(göm(p,G,e) \cup göm(q,G,e))\right|$$

Burada *göm(p,G,e),* p örüntüsünün G ağındaki e kenarını kapsayan gömüntüsüdür. |K(s)| , s çizgesinin kenar sayısıdır. Kısaca G ağında her iki gömüntü tarafından kapsanan kenar sayısının iki gömüntünün kapsadığı farklı kenarların toplam sayısına bölümüdür. Şekil 3.7b'deki M1 ve M2 örüntülerinin Şekil 3.7a'daki G çizgesindeki 3 ve 5 numaralı düğümler arasında tanımlı örtüşme değeri $O_{M1,M2,e3,5}$=2/8=0.25'dir. Benzer şekilde $O_{M3,M4,e4,7}$=3/5=0.60'dır. Bir aday kenarı için bir örüntünün birden fazla gömüntüsü var ise en yüksek O değerini veren gömüntü kombinasyonu kullanılarak hesaplanır. Bu ölçütler kullanıldığında $e_{3,5}$ ve $e_{4,7}$ kenarları için güven puanları şöyle olur:

$C^{ag}_{3,5}(M) = 0.5(1-3/5)(1-2/8) = 0.15$ ve

$C^{ag}_{4,7}(M) = 0.5(1-3/4)(1-3/5) = 0.05$.

3.2.4. Altçizge eşitliği

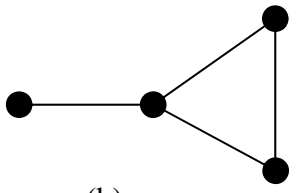
Birkaç taneden fazla sayıda kenarı olan çizgelerin aynı olup olmadığını belirlemek veya iki çizgenin tüm bağlantılarının yapısal olarak eşit olup olmadığını tespit etmek elle de olsa bilgisayar yardımıyla da olsa oldukça zor bir işlemdir. Bu kararı verebilen pratik bir algoritma tasarlanması "çizge eşitlik problemi" adı verilen meşhur bir problem olagelmiştir. Önceki bölümlerde bahsedilen örüntü bulma algoritması ağırlıklı olarak altçizge eşitlik sınamasının kullanımını gerektirmektedir. Bu bölümde altçizge eşitlik sınamasını gerçekleştiren algoritma açıklanmaktadır.

İki çizge arasındaki yapısal eşitliğin belirlenmesi çizgelerin düğüm ve kenarlarının birbirleriyle doğru biçimde eşlenmeleriyle mümkündür. Bu husus Şekil 3.8'deki iki çizge örneğiyle anlatılmıştır.

G	a	b	c	d
a	0	1	1	0
b	0	0	1	1
c	1	0	0	0
d	0	1	0	0

H	e	f	g	h
e	0	0	1	0
f	0	0	1	1
g	1	1	0	1
h	0	1	1	0

(a)



(b)

Şekil 3.8. (a)Bağlantı matrisleri verilmiş iki çizge (H ve G), (b) H ve G çizgelerinin temsil ettiği çizgenin bir çizimi.

G çizgesinin tanımı H çizgesinin tanımına aşağıdaki düğüm ve kenar eşleşmeleri (ing. bijection) ile dönüştürülebilir:

Düğüm	Kenar
a → f	{a,c} → {f,h}
c → h	{a,b} → {f,g}
b → g	{b,c} → {g,h}
d → e	{b,d} → {g,e}

Eğer n≥6 ise, n düğümlü ve q kenarlı çizgeler için böyle eşleştirme çiftlerinin bulunması işlemi hesaplamsal olarak oldukça zorlaşmaya başlar. Bir tanımın diğerine dönüştürülmesi işlemi, basitçe zor kullanma yöntemi ile *n!* düğüm eşleştirmesi ile *q!* kenar eşleştirmesinin denenmesini gerektirir.

Bir çizge eşitliği f: G → H şeklinde ifade edilir ve bir çift eşleştirmedir:

$f_V : V_G \rightarrow V_H$ ve $f_E : E_G \rightarrow E_G$

f_V fonksiyonu her bir e ∈ E kenarının uçlarını $f_E(e)$ kenarının uçları ile bağıntılandırmaktadır. Bu durumda G ve H'nin eşit (isomorphic) oldukları söylenir ve G ≅ H şeklinde gösterilir.

$A_G[v,w]$, v ve w düğümleri arasındaki kenar sayısı olsun. Bizim uygulamamız için bu 0 veya 1'dir. Bütün v,w∈G için $A_G[v,w]=A_H[f(v),f(w)]$ ise, G ve H çizgelerindeki $f : V_G \rightarrow V_H$ düğüm eşlemesi kenar çokluğunu korumaktadır. G ≅ H 'dir ancak ve ancak kenar çokluğunu koruyan bir $f : V_G \rightarrow V_H$ düğüm eşlemesi var ise. Elle çizge eşitliği sınaması için çizge eşitliği tanımını kullanmak yerine bu gerçekten yararlanmak daha kolaylaştırıcıdır, çünkü kenarlara isim verilerek kenar eşleştirmesi yapılması gereğini ortadan kaldırmaktadır. Dolayısıyla çizilmiş iki çizge arasındaki eşitliği göstermek için , önce birinin kenarları $v_1 \ldots v_n$'e kadar isimlendirilir, sonra ikinci çizge şekline, düğüm isimleri karşılık gelen kenarlar aynı kenar çokluğuna sahip olacak şekilde $f(v_1)$, $f(v_2)\ldots$, $f(v_n)$ olarak atanır.

Altçizge tanımına göre, H'nin G'nin altçizgesi olabilmesi için V_H ve E_H'nin V_G ve E_G'lerin alt kümesi olmaları gerekir. H, G'nin bir altçizgesine eşit ise H G'nin altçizgesidir şeklinde söylemek yaygındır. Çalışmada bu tanımlar kullanılmıştır.

Bir çizgenin başka bir çizgenin alt çizgesine eşit olup olmadığının bulunması problemi NP-complete sınıfından bir problemdir. Altçizge eşitlik problemini çözen bilinen tüm algoritmaların çalışma zamanı çizgenin ebatı cinsinden üstseldir. Polinom zamanlı bir algoritmanın bulunması (aksi ispatlanamadığı halde) oldukça zor görünmektedir. Bu çalışmada aşağıdaki kabullerden yararlanılmıştır.

Kabul 1: G ≅ H ise bütün x ∈ V_G için derece(x) ≤ derece(f_V(x)) 'dir.

Kabul 2: G=(V_G,E_G) ve H=(V_H,E_H) olsun. V⊆ V_G ve W⊆ V_H olsun. G'nin V'den müteşekkil altçizgesi(GV) ile H'nin W'dan müteşekkil altçizgesinin (HW)

birbirine eşit olduğunu varsayalım (GV ≅ HW). $v \in V_G \backslash V_{GV}$ ve $w \in V_H \backslash V_{HW}$ olan tüm düğümler için; G'nin $V_{GV} \cup \{v\}$'den müteşekkil altçizgesi, H'nin $V_{HW} \cup \{w\}$'dan müteşekkil altçizgesi ile ancak ve ancak aşağıdaki şart sağlandığında eşittir:

Bütün $x \in V_{GV}$ ve $y \in V_{HW}$ için, $\{v,x\} \in V_G$ ise $\{w,y\} \in V_H$'dir (Gross 2003).

Dikkat edilmelidir ki, iki etiketli çizge arasındaki altçizge eşitliği olabilmesi için bu şartlara ek olarak karşılıklı eşleştirilen düğümlerin etiketlerinin de uyuşması gerekmektedir. Bu çalışmada H çizgesi PPE ağına karşılık geldiğinden; G'deki bir düğümün etiketinin H'deki eşinin etiketlerinden en az biri ile uyuşması şartı aranmıştır.

Bir örüntünün gömüntüsünü tespit etmek için kullandığımız altçizge eşitlik sınamasını gerçekleştiren mapG3 algoritması aşağıda verilmiştir. Algoritmanın 5 parametresi şunlardır:

G: Düğümleri etiketli yönsüz çizge biçiminde ifa edilmiş PPE ağıdır

s: aranılan örüntü çizgesidir. Düğümleri 0'dan başlayarak numaralandırılmıştır.

p: $p \in V_G$ olup, $v \in Vs$ ile eşleştirilecektir. v'nin etiketi p'ninkiler arasında vardır.

v: $v \in Vs$ olup, $p \in V_G$ ie eşleştirilecektir. v'nin etiketi p'ninkiler arasında vardır.

map[]: o zamana dek eşleştirilen düğümleri içeren dizidir. map[i]==0 ise s'nin i'inci düğümü henüz G'deki bir protein ile eşleştirilmemiştir. Map[i]<>0 ise map[i], s'nin i'inci düğümünün eşleştirildiği proteinin numarasını taşımaktadır. İlk mapG3 çağrısında map[] elemanlarının tümü sıfırdır.

mapG3 algoritması v'nin p ile olan eşlemesinin o zamana dek bulunan eşleştirme kümesine eklenmesi ile s'deki tüm düğümlerin eşlendiğini tespit ettiğinde algoritma sonlanmaktadır. Aksi takdirde v'nin komşuları tek tek taranarak her biri için p'nin komşuları içinde bir eş aranmaktadır. Bu eş bulunduğunda mapG3 kendisini yapılan eşleme küme si ile birlikte yeniden çağırmaktadır (özyineleme). Bu durum yukarıda verilen 2 numaralı kabulün algoritmaya uyarlanmasıdır. Her adımda mevcut eşleştirme kümesine kabulde belirtilen şartı sağlayan yeni bir düğüm eklemek suretiyle eşleştirme kümesi genişletilmektedir. Bu genişletme s'deki tüm düğümler eşlenene kadar devam etmektedir. Bu durumda altçizge eşitlik sınaması başarı ile

souçlanmış olduğundan mapG3 bulunan eşleştirme kümesini döndermektedir. Şayet hiçbir genişletme imkânı bulunamazsa, başarılı bir altçizge eşitliği tespit edilemediğinden mapG3 sıfır döndermektedir. mapG3 algoritması derinlik öncelikli arama ağacı ile arama yapmaya ve aynı yere geri dönmeye (backtracking) güzel bir örnek teşkil etmektedir.

mapG3(v, map[], p, s, G)

```
1.  map[v]=p;
2.  map.length=map.length + 1;
3.  eğer (map.length== s'in düğüm sayısı) ise
4.      doğru bir işleştirme bulunmuştur, sonuç olarak map[]'i dönder.
5.  v'nin henüz eşleştirilmemiş (map[i]==0 olan) her bir i komşusu için
6.      p'nin henüz eşleştirilmemiş (j∉ map[] olan) her bir j komşusu için
7.          eğer i'nin etiketi j'nin bir etiketi ile aynı ise,
8.              düğüm_uygun=doğru;
9.              i'nin map[] içinde yer alan s'deki her bir k komşusu için
10.                 eğer map[k] j'nin G'de komşusu değil ise veya
11.                 k'nın etiketi map[k]'nın etiketlerinin içinde bulunmuyorsa
12.                     düğüm_uygun =yanlış;
13.                     kır (adım 14'e git);
14.             eğer düğüm_uygun ise
15.                 newmap[]=map[]'in bir kopyesi
16.                 eğer (yeni_map=mapG3(i,newmap[],j,s,G) > 0) ise
17.                     yeni_map'i dönder
18. Eşleştirme bulunamadığı için 0 dönder
```

Altçizge_say (s,G)

```
1.  sayı=0
2.  G'deki her bir i düğümü için
3.      eğer s'nin sıfırıncı elemanını etiketi i'nin bir etiketi ile aynı ise
4.          m[] listesinin elemanlarına 0 ata.
5.          Eğer (mapG3(0,m,i,s,G)>0) ise
6.              sayı'yı bir artır
7.  sayı'yı dönder.
```

altçizge_say_ gömüntü_ile(s,G)

```
1.  sayı=0
2.  eğer s'in bir gömüntü listesi var ise
3.      s'in gömüntü listesindeki her bir map[] gömüntüsü için
4.          v=s'in DÖA kod ağacındaki son eklenen düğümü veya
5.                                  son eklenen kenarın bir düğümü
6.          p=map[v];
7.          map[v]=0;
8.          map[].length=map[].length-1;
9.          eğer (yeni_map[]=mapG3(v,map,p,s,G)>0) ise
10.             sayı'yı bir artır
11.             map[]=yeni_map[];
12.         değil ise
13.             map[]'i gömüntü istesinden çıkar
14. aksi takdirde
15.     G'deki her bir i düğümü için
16      eğer s'nin 0. elemanını etiketi i'nin bir etiketi ile aynı ise
17.         m[] listesinin elemanlarına 0 ata.
18.         Eğer (yeni_map[]=mapG3(0,m,i,s,G)>0) ise
19.             sayı'yı bir artır
20.             yeni_map[]'i s'in gömüntü listesine ekle
21. sayı'yı dönder.
```

altçizge_say_ gömüntü_ile ve altçizge_say algoritmaları mapG3 algoritmasını kullanarak s örüntüsünün G çizgesindeki örtüşmeyen gömüntü sayısını bulan algoritmalardır. altçizge_say algoritması G'ki her bir i proteini için s'deki ilk düğümle etiketi uyuşuyor ise mapG3(0,m,i,s,G) çağrısını yapmaktadır. Böylece olabilecek tüm örüntülerin başlangıç eşleştirmeleri kapsanmış olmaktadır. altçizge_say_ gömüntü_ile algoritması altçizge_say algoritması ile aynı işleve sahip olmakta olup sadece kullandığı yöntem değişiktir. altçizge_say_gömüntü_ile algoritması gömüntü bilgilerini bir sonraki sayım için saklamak suretiyle hız kazanmaktadır. Her altçizge eşitlik sınaması sadece bir düğüm eşleştirmesine malolduğundan örüntü arama algoritmasının çalışma zamanında çok önemli miktarda kazanç sağlanmaktadır. Dikkat edilmelidir ki s'in DÖA kod ağacına göre bir çocuğu oluşturulur iken s'in bütün gömüntü listesi çocuğunkine kopyalanmaktadır. Bu sayede ana örüntü için oluşturulan gömüntüler çocuk örüntünün gömüntülerini tespit edip saymakta kullanılmaktadır. Altçizge eşitlik sınamalarında gömüntülerin kullanımı ile ilgili daha ayrıntılı bilgi için Bölüm 3.2.1'e bakınız.

4. ARAŞTIRMA SONUÇLARI

PPISpan algoritması C++ dilinde gerçekleştirildi. Deneyler Intel Pentium-IV 2.0 GHz işlemcili ve 1GB hafızalı kişisel bilgisayarda çalıştırıldı. *Saccharomyces Cerevisiae* organizmasına ait 3 farklı umumi PPE ağında örüntü araması yapıldı: DIP, STRING ve WI-PHI. PPE ağının düğümleri olan proteinleri Hamur Mayası için verilmiş GO Slim moleküler fonksiyon dipnotları/sınıflamaları ile etiketlendirildi. Kullanılan veri kümeleriyle ilgili ayrıntılı bilgi "Materyal ve Metot" bölümünde bulunabilir.

Desteği 15 ve üzeri olan sık geçen örüntüler arandı. Değişik asgari destek (eşik) değerleri ile yaptığımız deneyler sonucunda asgari destek değeri olarak 15'in makul sayıda örüntüyü makul çalışma süreleri içinde verdiğine kanaat getirdik. Çizelge 2, bu üç PPE ağı için bulunan örüntü sayılarını ve bunlar içinde önemli olanları vermektedir. Bu örüntüler "moleküler fonksiyonu bilinmemektedir" şeklindeki GO terimi ile etiketlenmiş örüntüleri içermemektedir. Sağdaki sütunda sık geçen örüntülerin içinde Z-puanı 2,3'ün üstünde olan örüntülerin sayıları verilmiştir.

Çizelge 2. Bulunan örüntülerin sayısı: PPE ağlarında en az 15 gömüntüsü bulunan örüntülerin sayısı.

PPE Ağı	Sık geçen örüntü sayısı	Z-puanı 2,3'den büyük olan örüntü sayısı
DIP	205	199
STRING	287	17
WI-PHI	378	321

PPISpan DIP ağında 205 adet örüntü bulunmuştur. Bunlardan 199 adeti Z-puanı 2.3'ün üzerinde olanlardır. Örüntülerin gömüntüleri DIP ağının %37.06'sını (1828 proteini) kapsamaktadır. STRING ağında, 287 etkileşim örüntüsü bulunmuştur ve bunların sadece 17'sinin Z-puanı 2.3'ün üzerindedir. Bu örüntülerin gömüntüleri STRING ağının %40.79'unu kapsamaktadır (1204 protein). WI-PHI ağında 321'i

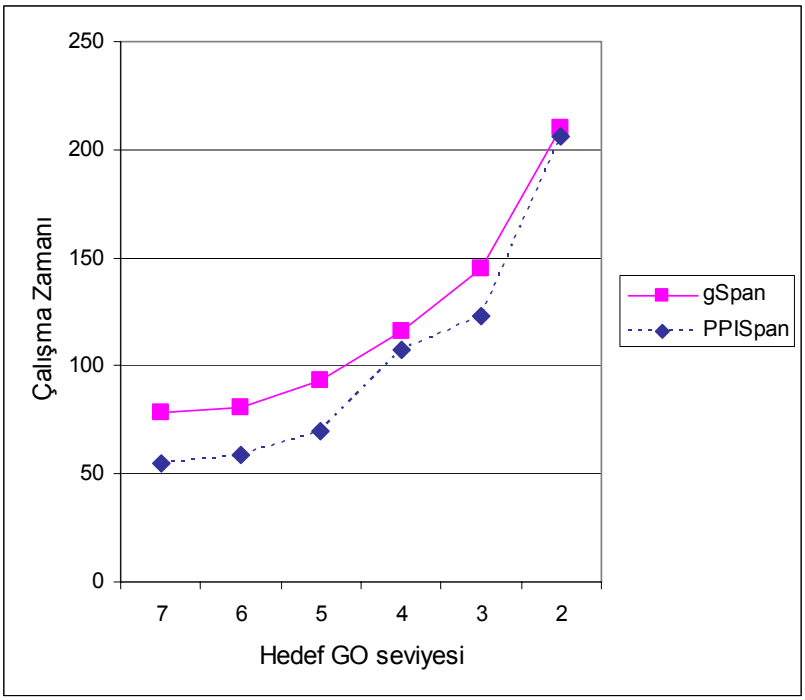
önemli olan toplam 378 adet sık geçen örüntü bulunmuştur. Bu örüntüler ağdaki 1734 proteini yani ağın %37.27'sini kaplamaktadır. Bir örüntünün iki gömüntüsü birbirleriyle örtüşmemekle birlikte örüntüler birbirleriyle örtüşebilirler. Sadece bir örüntü bir diğerinin alt-çizgesi olamaz. Bulunan örüntülerin büyük çoğunluğu ağaçtırlar. Yıldızlar çoğunluktadır. Çemberler az görülmektedir. Bu bulgumuz, yaklaşık ağaç yapılarını bulan ve dolayısıyla daha hızlı (anında etkileşimli) örüntü bulma algoritmalarının kullanılabileceğini göstermektedir.

GO Slim terim kümesine ek olarak orijinal terimlerle ve çeşitli GO hedef seviyelerinde de arama yaptık (bkz. 3.2.Metot). Hedef seviyeyi yukarı doğru çıkardıkça çalışma zamanının beklenildiği üzere üstsel olarak arttığını gördük. DIP verisi ile yapılan bu deneylere ilişkin sonuçlar gSpan algoritmasının çıktıları ile birlikte Çizelge 5'de gösterilmiştir. Çizelgede istatistik öneme sahip olsun olmasın algoritma tarafından bulunan tüm örüntülerin sayısı verilmiştir. Bulunan örüntü sayısı genel olarak hedef seviyenin yukarı çekilmesiyle birlikte artmaktadır. Bunu GO seviyelerinde yukarı doğru çıktıkça bir terimin altına düşen PPE ağındaki protein sayısının artmasına bağlamak mümkündür.

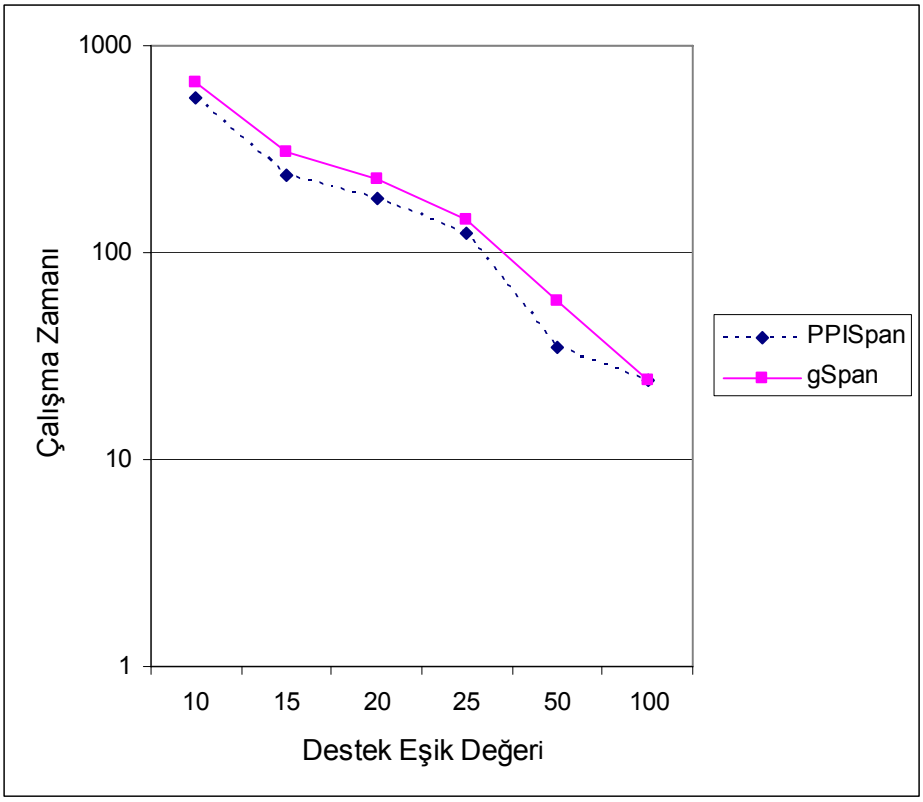
Çizelge 5. Bazı eşik destek değerleri için PPISpan ve gSpan algoritmalarının farklı GO seviyelerinde ve GO Slim terimleri kümesiyle verdikleri çıktı örüntü sayısı ve çalışma süreleri (saniye).

	Eşik Destek	10		15		25		50		100	
Algoritma	***GO Seviyesi***	Sayı	Süre	Sayı	Süre	Sayı	Süre	Sayı	Süre	Sayı	Süre
PPISpan	Orijinal	159	248	32	119	12	46	0	18	0	6
	3. seviye	180	561	79	235	24	123	2	35	0	24
	GOslim	442	8025	205	2460	26	566	9	163	0	38
gSpan	Orijinal	159	294	32	156	12	78	0	27	0	6
	3. seviye	180	669	79	303	24	145	2	58	0	24
	GOslim	442	9190	205	2572	26	759	9	182	0	36

Çalışma zamanı açısından PPISpan algoritmasının gSpan'e göre genel olarak daha verimli olduğunu söyleyebiliriz. Bu durum, Şekil 4.1'de L=3 hedef seviyesi için çizilmiş olan performans grafiğinde görülmektedir. Bu performans farkının eşik destek seviyesinin artmasıyla azalma eğilimi gösterdiği çıkarımını yapmak mümkündür. Şekil 4.2'de, farklı destek eşik değerleri için gSpan ve PPISpan'in performans grafikleri çizilmiştir. Burada çalışma zamanı ekseninin yelpazesi logaritmiktir.



Şekil 4.1. Farklı GO seviyelerinde PPISpan algoritmasının performansı



Şekil 4.2. Farklı destek eşik değerleri için PPISpan algoritmasının performansı.

Sonraki bölümde bulduğumuz örüntüleri doğrulamak için bilinen moleküler bileşiklerden ve patikalardan nasıl yararlandığımızı açıklıyoruz. Daha sonra ise biyologlara ilginç gelebilecek bazı örüntüleri sunuyoruz.

4.1. Bilinen Moleküler Bileşikler ve Patikalar ile Karşılaştırma

Bir gen ölçekli PPE ağı, moleküler bileşikler, sinyal iletim ve nakliye patikaları gibi fonksiyonel parçalardan oluşmaktadır. Öte yandan, PPISpan algoritmamızın buldukları gibi fonksiyonel etkileşim örüntüleri, PPE ağının değişik ortamlarında geçen belli bazı tip düğümlerden müteşekkil alt-çizgelerdir. Bu bölümde, var olan biyolojik bilgilerle, bulduğumuz örüntüleri yorumlamaya ve doğrulamaya çalışıyoruz. Tekrar vurgu yapmak isteriz ki, örüntü bulmadaki amacımız özgün bileşikler veya patikalar bulmak değildir. Amacımız, ağın altında

yatan fonksiyonel etkileşim mekanizmalarını ve bunların PPE ağlarının farklı ortamlarında nasıl tekrar kullanıldığını anlamaktır.

Bulunan sık geçen örüntüleri analiz etmenin makul bir yolu onların gömüntüleri ile bilinen moleküler bileşikleri ve patikaları karşılaştırmaktır. Bu amaçla kurguladığımız deneyde, gömüntülerin içerdikleri proteinler (düğümler) ile moleküler bileşik ve patikaları, düğümlerin aralarındaki etkileşimleri temsil eden kenarları dikkate almaksızın, karşılaştırdık. İdeal şartlarda topolojiyi de karşılaştırmaya dâhil etmemiz gerekir ancak mevcut moleküler bileşik veritabanı sadece bileşiğin üyesi olan proteinleri içermekte, ikili etkileşimleri listelememektedir. Dolayısıyla bu bölümde örüntülerdeki topolojiyi göz ardı ederek, örüntüyü bir protein kümesi olarak ele alıyoruz.

Moleküler bileşikleri MIPS bileşik veritabanından aldık. Sinyal iletim, nakliye ve düzenleme patikalarını ise KEGG veritabanından aldık. Yüksek miktarda veri sağlayan deneylerin çıktılarını dışarıda bırakarak, geride kalan yüksek kalitedeki 267 adet MIPS bileşiğini bilinen moleküler bileşikleri olarak kabul ettik ve kullandık. KEGG veritabanından kullandığımız patikalar ise şunlardır: ABC taşıyıcıları, MAPK sinyal iletim patikası, phosphatidylinositol sinyal iletimi sistemi, kabarcıklı taşıyıcı patikalarındaki SNARE etkileşimleri ve otofaj patikalarının düzenlenmesi.

Sık geçen etkileşim örüntülerinin yorumlanması ve doğrulanması için iki ölçüt önermekteyiz. Birinci ölçüt, bir örüntünün gömüntülerinin MIPS bileşikleri ve KEGG patikaları ile ortalama örtüşme sayısıdır. Bunu *cpsayısı* olarak isimlendiriyoruz. Bu ölçütün amacı örüntünün yeri konusunda bilgi edinmek, özel olarak da bileşiklerin içinde mi yoksa aralarında mı olduklarını tespit etmektir. *cpörtüşme* olarak isimlendirdiğimiz ikinci ölçütümüz ise bir gömüntüdeki proteinlerin hangi oranda bilinen bileşikler ve patikalarla örtüştüğünü niceler. Bir *e* gömüntüsünün örtüşme ölçütü, *e*'deki proteinlerin bilinen fonksiyonel modüller ile örtüşme oranı olarak hesaplanır:

$$cpörtusme(e) = \frac{|\{p \mid p \in e \wedge p \text{ fonksiyonel parça içindedir}|}{|e|}$$

Sık görülen bir etkileşim örüntüsü, fonksiyonel parçaların içindeki etkileşimleri içermeye, dışındakileri içermeye göre daha yatkındır. Fonksiyonel

modüller, MIPS'deki gibi moleküler bileşikler ve KEGG'deki gibi düzenleme patikaları olabilir. Dolayısıyla bir örüntü birden çok fonksiyonel parça ile örtüşecektir. Ancak, bilinen bileşikler tam olmaktan oldukça uzak ve DIP PPE ağının sadece %23,9'unu kapsamaktadırlar. Dolayısıyla örüntülerin bir kısmı bilinen fonksiyonel parçalar ile örtüşmüyor olacaktır. Bütün örüntüler üzerinde sistematik analiz yaptığımızda, düğümlerin etiketleri göz ardı edilerek yapılan deneylerimiz göstermiştir ki, fonksiyonel parçalar ile olan örtüşme, benzer topolojilerin rastgele gömüntüleriyle olan örtüşmeyle önemli farklılık göstermemiştir (bkz. Çizelgeler 3 ve 4). Bu gözlemin bizce nedeni bazı örüntülerin genel fonksiyonel terimler içermeleri yani fonksiyon açısından yeterince özel olmamalarıdır. Bir başka deyişle, bazı örüntüler için topoloji, altında yatan fonksiyonel betimlemeden daha önemlidir ve bu nedenle rastgele gömüntülere benzemektedirler.

Çizelge 3'de DIP, STRING ve WI-PHI ağlarında bulunan tüm örüntülerin ortalama *cpörtüşme* değerleri ve karşılarında aynı topolojiye sahip rastgele örüntülerin *cpörtüşme* değerleri verilmiştir. Ölçüm MIPS bileşiklerine göre yapılmıştır. Çizelge 4'de ise aynı ölçüm nakliye ve sinyal iletme patikalarına göre yapılmıştır.

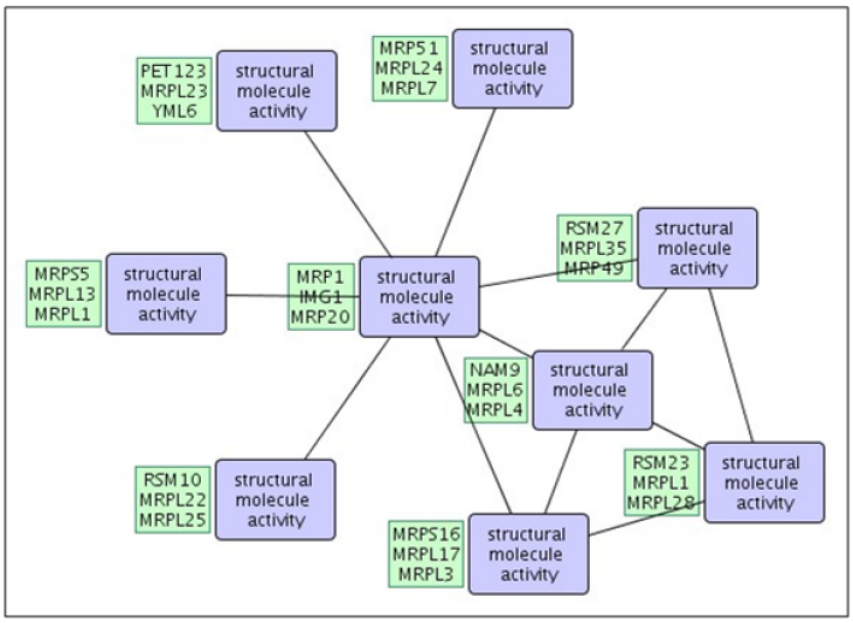
Çizelge 3. Tüm örüntülerin rastgele örüntüler ile MIPS bileşikleri kullanılarak elde edilmiş cpörtüşme değerleri açısından karşılaştırılması.

PPE Ağı	Sık geçen örüntülerin cpörtüşme değerleri		Rastgele örüntülerin cpörtüşme değerleri	
	Ortalama	Standart Sapma	Ortalama	Standart Sapma
DIP	0,729	0,111	0,718	0,049
STRING	0,835	0,056	0,73	0,036
WI-PHI	0,609	0,138	0,657	0,054

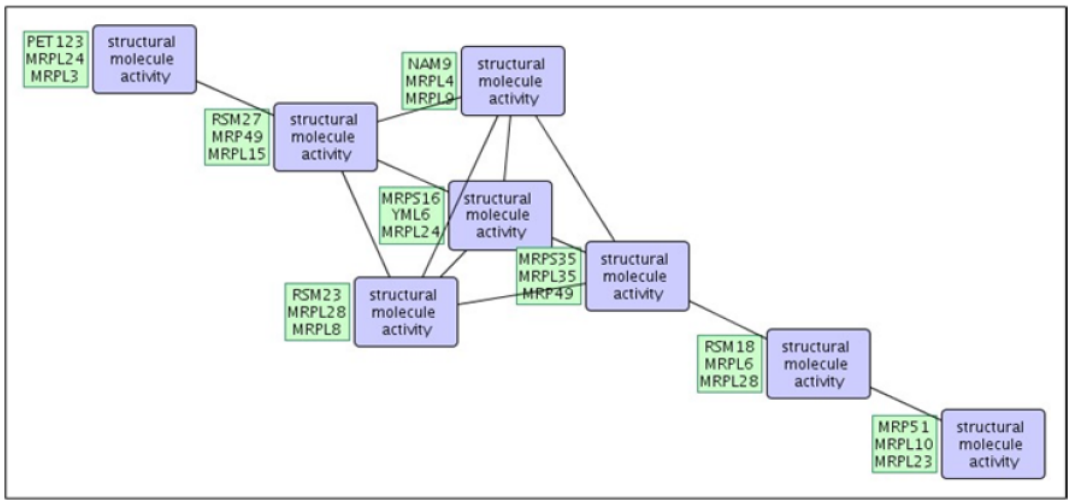
Çizelge 4. Tüm örüntülerin rastgele örüntüler ile nakliye ve sinyal iletme patikaları kullanılarak elde edilmiş cpörtüşme değerleri açısından karşılaştırılması.

PPE Ağı	Sık geçen örüntülerin cpörtüşme değerleri		Rastgele örüntülerin cpörtüşme değerleri	
	Ortalama	Standart Sapma	Ortalama	Standart Sapma
DIP	0,982	0,050	0,959	0,013
STRING	0,979	0,033	0,991	0,014
WI-PHI	0,982	0,043	0,948	0,017

DIP ağındaki sık geçen etkileşim örüntülerinden MIPS bileşikleriyle karşılaştırıldığında en yüksek *cpörtüşme* değerine sahip olan iki tanesi Şekil 4.4(a) ve Şekil 4.4(b)'de verilmiştir. Bunlar metindeki, Şekil 4.6 ve Şekil 4.7'deki 1 numaralı ve 2 numaralı örüntülere karşılık gelmektedirler. WI-PHI ağındaki sık geçen etkileşim örüntülerinden MIPS bileşikleriyle karşılaştırıldığına en yüksek *cpörtüşme* değerine sahip olan iki tanesi Şekil 4.4(c) ve Şekil 4.4(d)'de verilmiştir. Bunlar metindeki, Şekil 4.6 ve Şekil 4.7'deki 5 numaralı ve 6 numaralı örüntülere karşılık gelmektedirler. Örüntülerin ilk üç gömüntüsü yeşil/köşeli kutucuklarda gösterilmiştir. Mavi/yuvarlak kenarlı kutucuklarda düğümlerin GO moleküler fonksiyonel sınıfları verilmiştir.

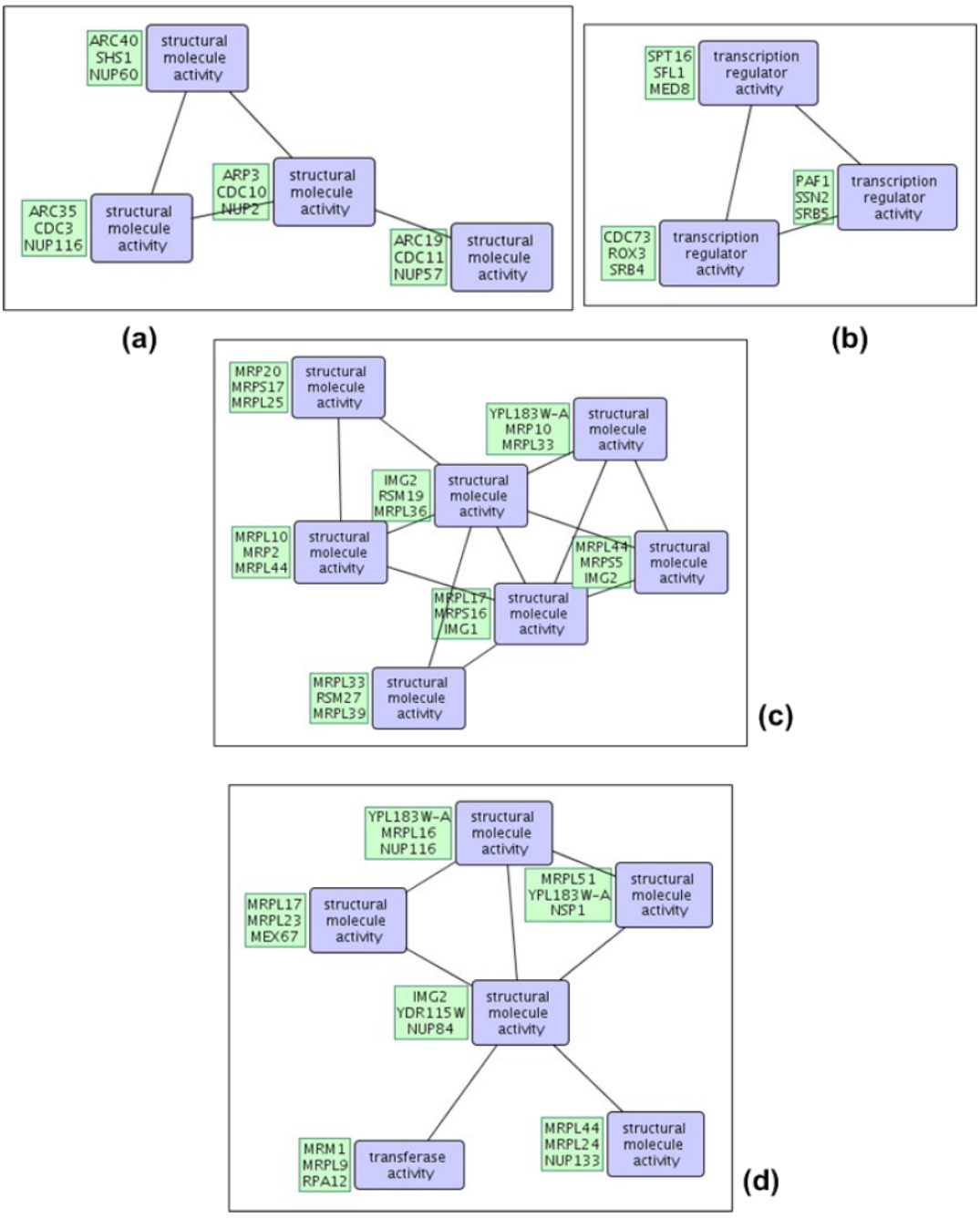


(a)



(b)

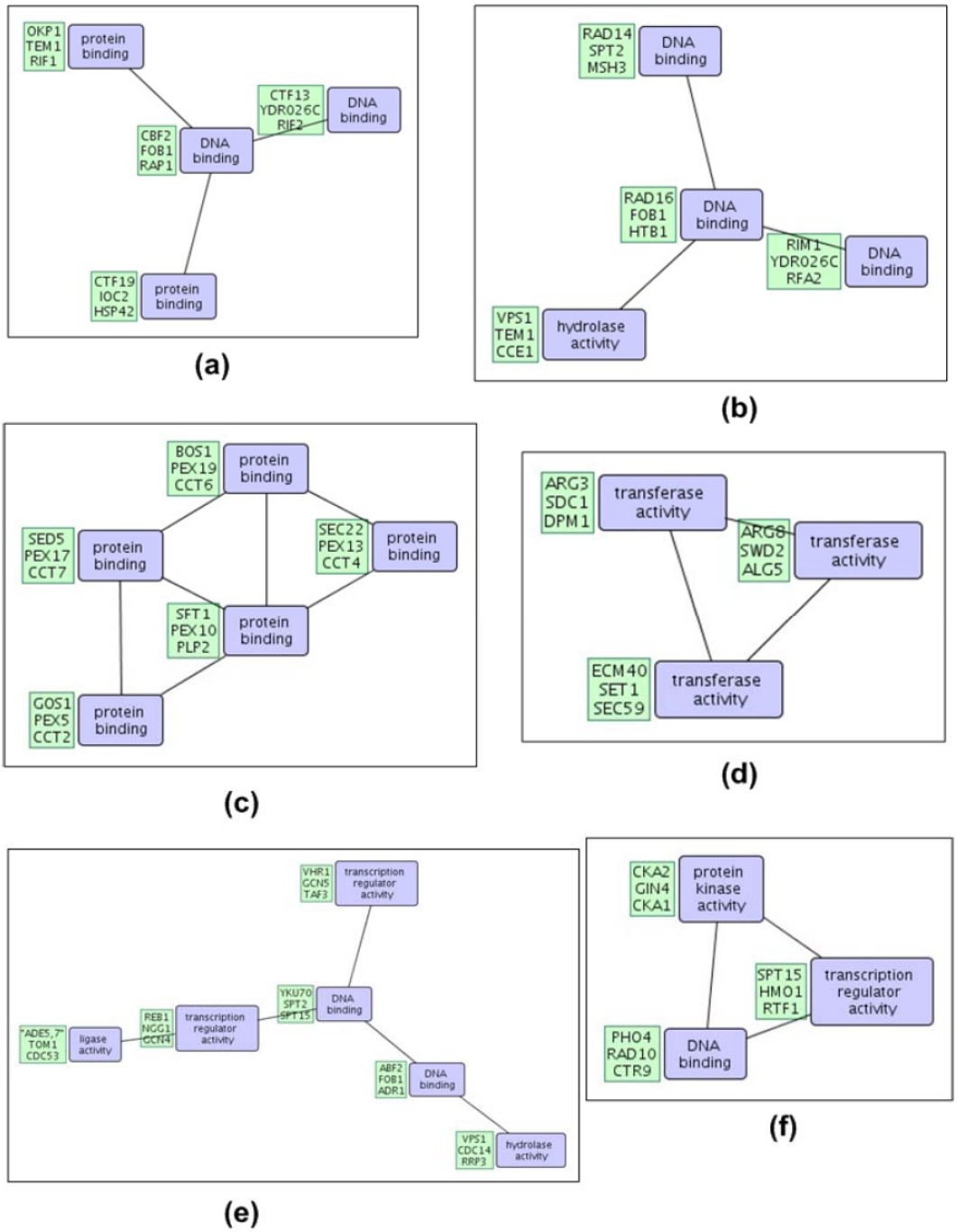
Şekil 4.3. STRING ağında bulunuş örüntülerden iki örnek.



Şekil 4.4. DIP ve WI-PHI ağlarında bulunmuş örüntülerden dört örnek.

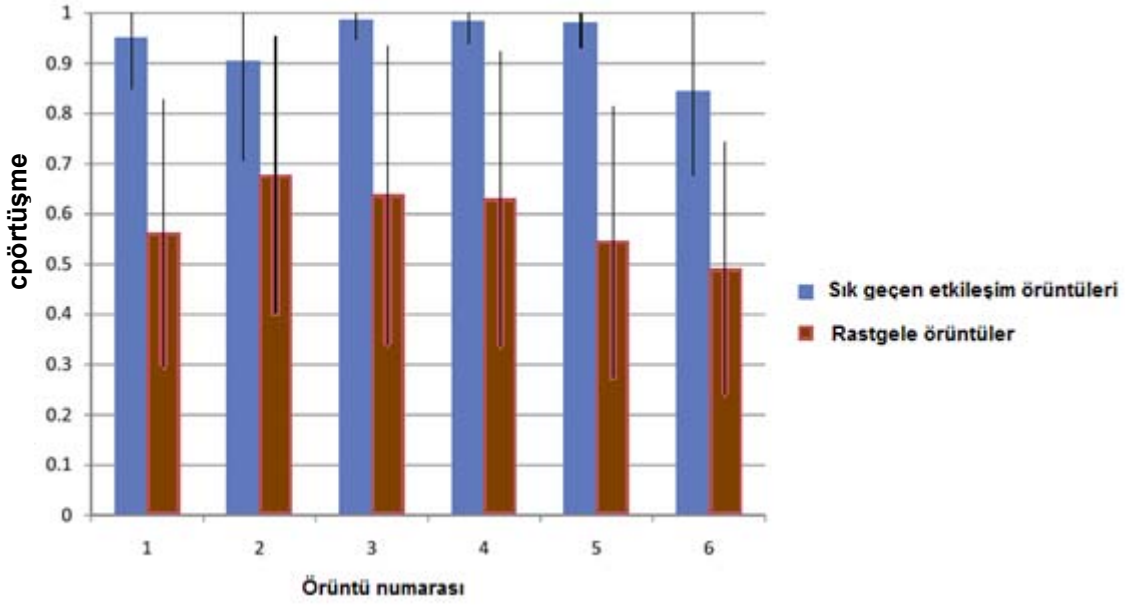
Aşağıda, DIP, STRNG ve WI-PHI ağlarındaki en yüksek *cpörtüşme* değerlerine sahip iki örüntüyü değerlendiriyoruz. Bu örtüşmenin aynı topolojiye sahip rastgele gömüntülerinki ile karşılaştırıldığında önemli olduğunu gösteriyoruz. Seçilen

örüntüler 4.4 ve 4.3 şekillerinde gösterilmişlerdir. MIPS bileşikleriyle karşılaştırıldığına en yüksek cpörtüşme değerine sahip olan iki sık geçen etkileşim örüntüsü Şekil 4.3(a) ve 4.3(b)'de verilmiştir. Bunlardan metinde 3 numaralı ve 4 numaralı örüntüler olarak bahsedilmekte, aynı zamanda bunlar Şekil 4.6 ve Şekil 4.7'deki 3 ve 4 numaralı örüntülere karşılık gelmektedirler. Örüntülerin ilk üç gömüntüsü yeşil/köşeli kutucuklarda gösterilmiştir. Mavi/yuvarlak kenarlı kutucuklarda düğümlerin GO fonksiyonel sınıfları verilmiştir. Rastgele örüntüler PPISpan tarafından fonksiyonel betimlemeler görmemezlikten gelerek bulunmuşlardır. Şekil 4.6'de MIPS bileşiklerine göre 6 örüntü için hesaplanmış cpörtüşme değeri verilmektedir. Şekil 4.7'de MIPS bileşiklerine göre 6 örüntü için hesaplanmış *cpsayısı* değeri verilmektedir. Şekil 4.6'en anlaşıldığı üzere, sık geçen örüntülerin *cpörtüşme* değerleri aynı topolojiye sahip rastgele gömüntülerin ortalama *cpörtüşme* değerlerinden önemli oranda yüksektir (p-değeri<10^{-10}). Şekil 4.7'e göre, DIP ağında bulunmuş olan 1 ve 2 numaralı örüntülerin gömüntüleri ortalama olarak 2.4 ve 1.9 adet MIPS bileşiği ile örtüşmektedirler. 1 numaralı örüntünün cpsayısı rastgele örüntülerin cpsayısından önemli oranda fazladır (p-değeri=0.0372), ancak 2 numaralı örüntü için bu doğru değildir (p-değeri=0.3268).



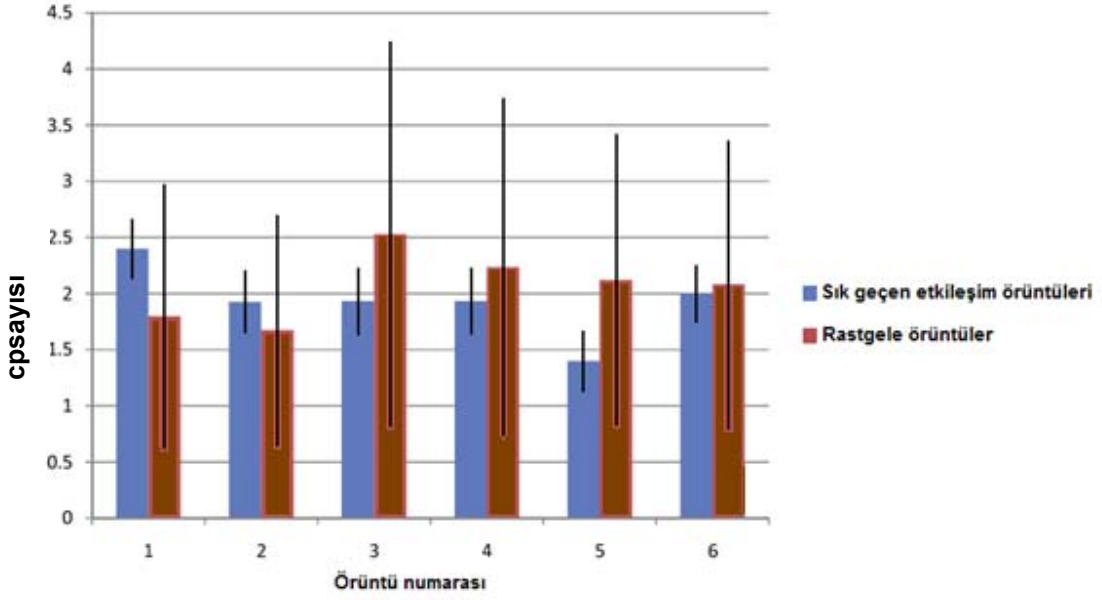
Şekil 4.5. DIP, STRING ve WI-PHI ağlarından seçilmiş örüntüler.

Şekil 4.5'de KEGG patikaları ile karşılaştırıldığında en yüksek *cpörtüşme* değerini veren 6 örüntü verilmektedir. Bu örüntüler algoritmamız tarafından DIP, STRING ve WI-PHI ağlarında bulunmuşlardır. Şekil 4.5(a)'dan (f)'ye kadar olan örüntüler aynı sırayla, Şekil 4.8 ve Şekil 4.9'deki 7 numaralı örüntüden 12 numaralı örüntüye kadar karşılık gelmektedir. Örüntülerin ilk üç gömüntüsü yeşil/köşeli kutucuklarda gösterilmiştir. Mavi/yuvarlak kenarlı kutucuklarda düğümlerin GO moleküler fonksiyonel sınıfları verilmiştir. Bu örüntülerin tümü kopyalama düzenleme ile ilgilidir. Şekil 4.8, bu örüntüler KEGG patikaları ile karşılaştırıldığında elde edilen cpörtüşme değerlerini, Şekil 4.9 ise aynı şekilde elde edilen *cpsayısı* değerlerini içermektedir. Sonuçlar, MIPS bileşikleri ile elde edilen örtüşme sonuçlarından oldukça farklı görünmektedir. Bunun en önemli sebebi deneyde kullanılan KEGG patikası sayısının (5) MIPS bileşik sayısından (267) çok az olmasıdır. Nakliye ve sinyal iletme patikaları PPE ağının çok küçük bir bölümünü kapsamaktadır. Ancak, Şekil 4.8 göstermektedir ki; seçilen örüntülerin *cpörtüşme* değerleri rastgele seçilen gömüntülerinkinin önemli derecede üstündedir (p-değeri=0,4165 olan 10 numaralı örüntü hariç). Şekil 4.9'de görüldüğü üzere, bir gömüntü içinde bulunan ortalama KEGG patikaları sayısı 1'dir. Ancak patikalarla kesişen bu tür örüntülerin sayısının azlığı ortalama örtüşme sayısı ile ilgili bir sonuç çıkarmamızı engellemektedir.



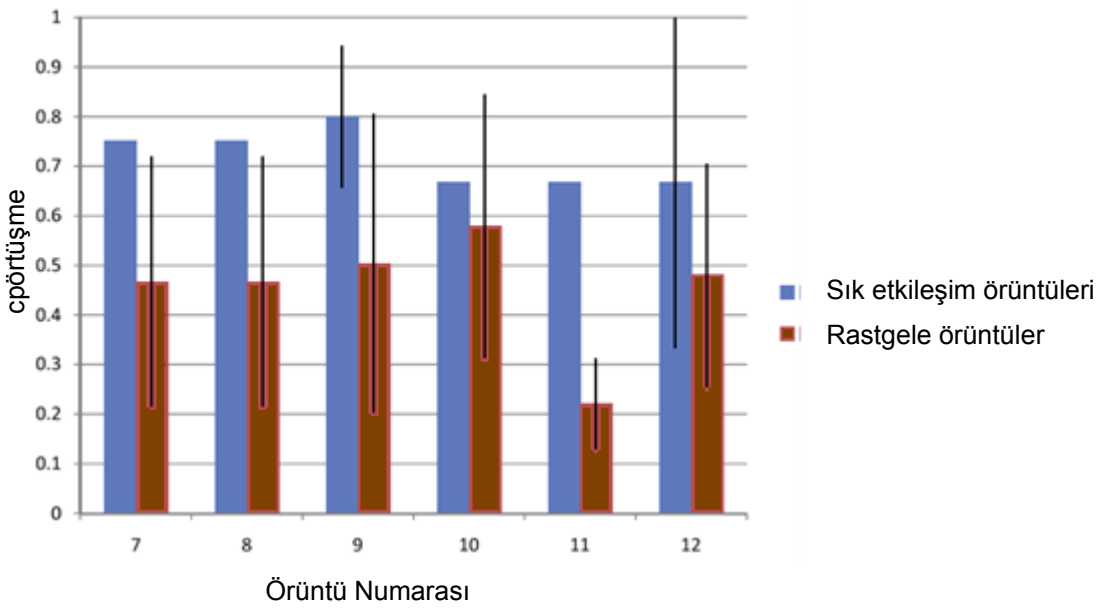
Şekil 4.6. Seçilmiş örüntülerin MIPS bileşikleriyle karşılaştırılarak elde edilen *cpörtüşme* değerleri.

Şekil 4.6'de, seçilmiş sık geçen örüntülerin ve aynı topolojiye sahip rastgele gömüntülerin ortalama *cpörtüşme* değerleri bar grafiği şeklinde gösterilmiştir. Bir gömüntünün cpörtüşme değeri onun proteinlerinden bilinen fonksiyonel parçalar ile örtüşenlerinin oranıdır. Bu değerler MIPS bileşiklerine göre hesaplanmıştır. Bir örüntünün farklı gömüntüleri farklı *cpörtüşme* değerlerine sahip olabilir. Bir örüntünün gömüntülerinin *cpörtüşme* değerlerindeki standart sapma, barların üstünde hata çizgileri olarak iki yöne de bir standart sapma uzayacak kadar çizilerek gösterilmiştir. Yatay eksendeki örüntü numaraları Şekil 4.4 ve 4.3'deki örüntüleri temsil etmektedir.



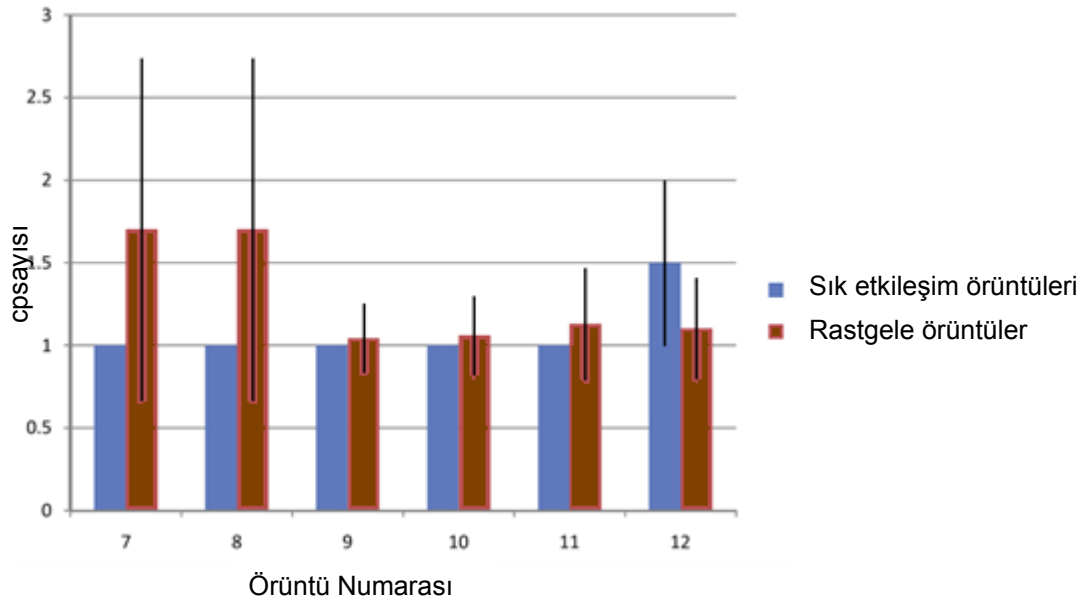
Şekil 4.7. Seçilmiş örüntülerin MIPS bileşikleriyle karşılaştırılarak elde edilen *cpsayısı* değerleri.

Şekil 4.7'de, seçilmiş sık geçen örüntülerin ve aynı topolojiye sahip rastgele gömüntülerin ortalama *cpsayısı* değerleri bar grafiği şeklinde gösterilmiştir. cpsayısı, bir etkileşim örüntüsünün bir gömüntüsünün kesiştiği farklı bileşik veya patikalar sayısıdır. Bu değerler MIPS bileşiklerine göre hesaplanmıştır. Bir örüntünün farklı gömüntüleri farklı *cpsayısı* değerlerine sahip olabilir. Bir örüntünün gömüntülerinin *cpsayısı* değerlerindeki standart sapma, barların üstünde hata çizgileri olarak iki yöne de bir standart sapma uzayacak kadar çizilerek gösterilmiştir. Yatay eksendeki örüntü numaraları Şekil 4.4 ve 4.3'deki örüntüleri temsil etmektedir.



Şekil 4.8. Seçilmiş örüntülerin nakliye ve sinyal iletme patikalarıyla karşılaştırılarak elde edilen *cpörtüşme* değerleri.

Şekil 4.8'da, seçilmiş sık geçen örüntülerin ve aynı topolojiye sahip rastgele gömüntülerin ortalama *cpörtüşme* değerleri bar grafiği şeklinde gösterilmiştir. Bir gömüntünün *cpörtüşme* değeri onun proteinlerinden bilinen fonksiyonel parçalar ile örtüşenlerinin oranıdır. Bu değerler KEGG patikalarına göre hesaplanmıştır. Bir örüntünün farklı gömüntüleri farklı *cpörtüşme* değerlerine sahip olabilir. Bir örüntünün gömüntülerinin *cpörtüşme* değerlerindeki standart sapma, barların üstünde hata çizgileri olarak iki yöne de bir standart sapma uzayacak kadar çizilerek gösterilmiştir. Yatay eksendeki örüntü numaraları Şekil 4.5'deki örüntüleri temsil etmektedir.



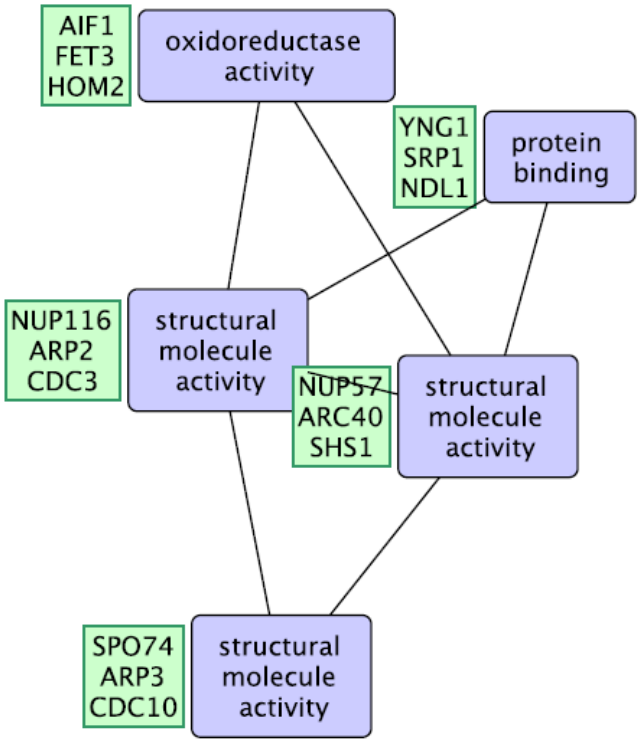
Şekil 4.9. Seçilmiş örüntülerin nakliye ve sinyal iletme patikaları ile karşılaştırılarak elde edilen *cpsayısı* değerleri.

Şekil 4.9'de, seçilmiş sık geçen örüntülerin ve aynı topolojiye sahip rastgele gömüntülerin ortalama *cpsayısı* değerleri bar grafiği şeklinde gösterilmiştir. c*psayısı*, bir etkileşim örüntüsünün bir gömüntüsünün kesiştiği farklı bileşik veya patikalar sayısıdır. Bu değerler KEGG patikalarına göre hesaplanmıştır. Bir örüntünün farklı gömüntüleri farklı *cpsayısı* değerlerine sahip olabilir. Bir örüntünün gömüntülerinin *cpsayısı* değerlerindeki standart sapma, barların üstünde hata çizgileri olarak iki yöne de bir standart sapma uzayacak kadar çizilerek gösterilmiştir. Yatay eksendeki örüntü numaraları Şekil 4.5'deki örüntüleri temsil etmektedir.

Özetle, bulunan örüntüleri doğrulama çabalarımız göstermektedir ki, bazı örüntüler moleküler bileşiklerle ve metabolik patikalarla önemli ölçüde kesişim göstermektedir ve örüntüler çoğunlukla moleküler bileşiklerin arasında ve metabolik patikaların içinde bulunmaktadırlar.

4.2. Bazı İlginç Etkileşim Örüntüleri

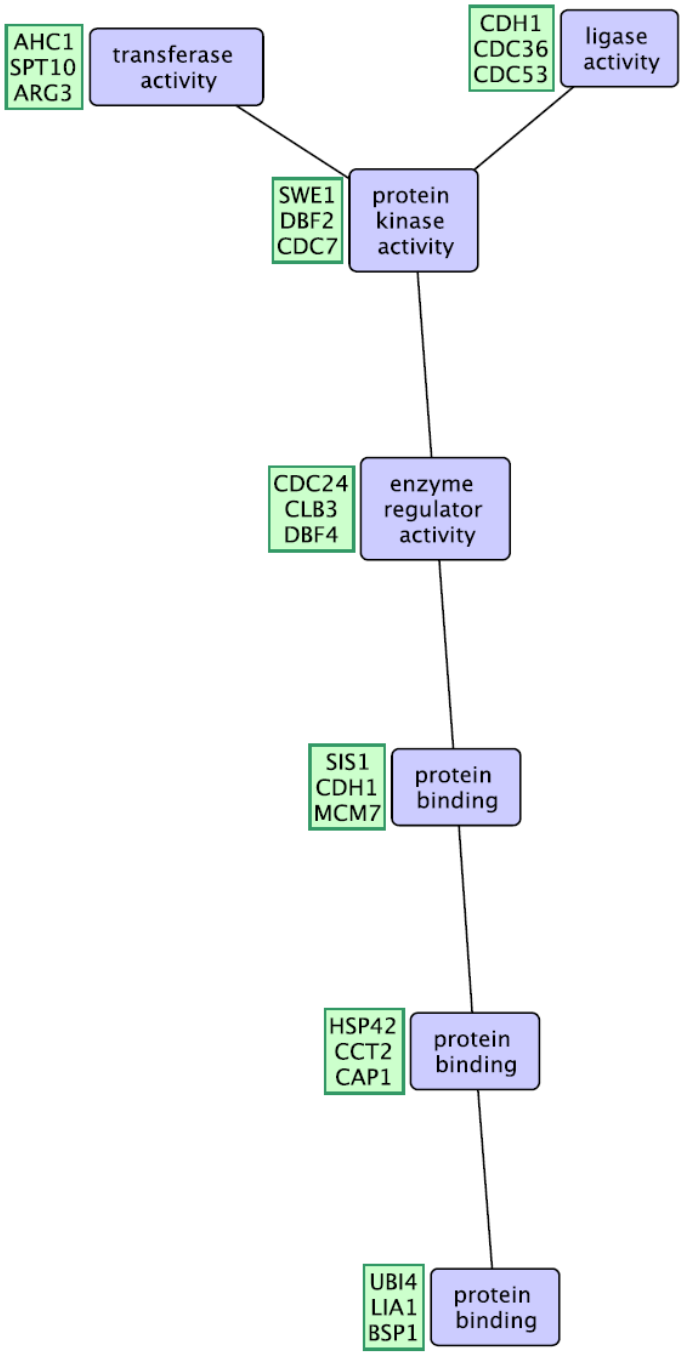
Bu bölümde biyologlar için ilginç olabilecek bazı örüntüleri sunuyoruz. Şekil 4.10, DIP ağında keşfedilmiş olan ve çemberler içeren bir örüntüyü göstermektedir. Örüntü 5 düğümlüdür, PPE ağında 15 kez görülmektedir ve Z-puanı 7,4'dür. Örüntüde 3 tane yapısal molekül etkinliği terimi, bir adet protein bağlanma terimi ve bir adet oksidoreduktaz etkinliği terimi bulunmaktadır. Bu örüntünün 15 gömüntüsünden 3'ü düğümlerin solunda yeşil kutucuklarda verilmiştir.



Şekil 4.10. DIP ağında bulunan bir fonksiyonel etkileşim örüntüsü.

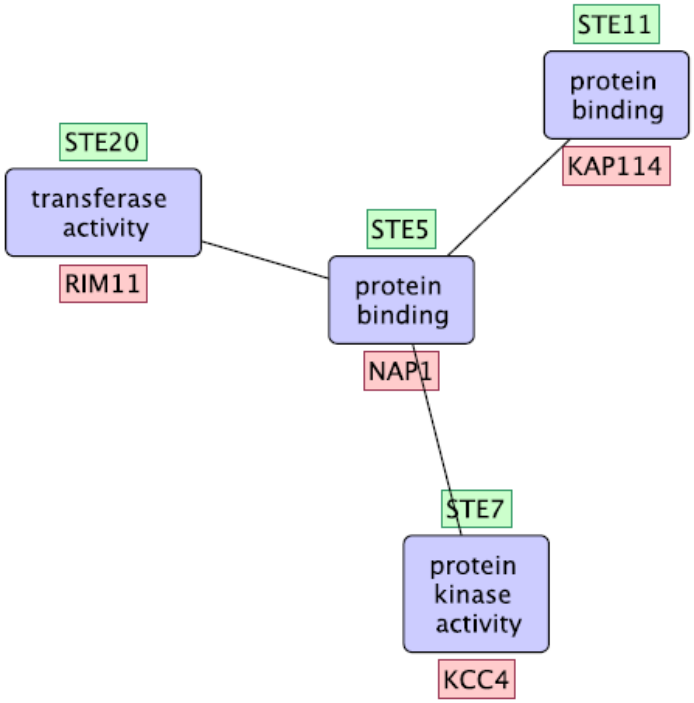
Örneğin YNG1 proteini, protein bağlanması etkinliğini gerçekleştiren bir proteindir ve bu örüntünün bir gömüntüsünde oksidoreduktaz etkinliği gösteren AIF1 proteini ile birlikte yer almıştır. Aynı gömüntünün diğer proteinleri NUP116, NUP57 ve SPO74'dür. Bu örüntüde görülen etkinlikler pek çok hücre olayında temel etkinliğe

sahiptirler dolayısıyla bu örüntünün Hamur Mayası canlısının PPE ağında çok sayıda geçmiş olması sürpriz değildir.



Şekil 4.11. WI-PHI ağında bulunan bir fonksiyonel etkileşim örüntüsü.

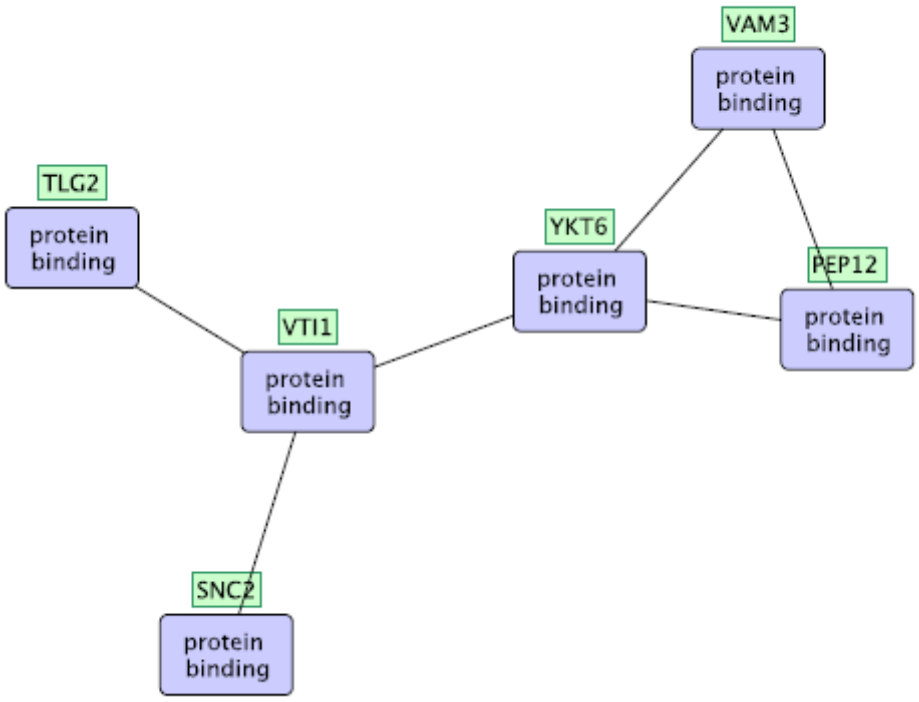
Pek çok veri kaynağından alınan etkileşim bilgilerinin birleştirilmesi ile oluşturulmuş olan WI-PHI PPE ağında daha büyük örüntüler bulunmuştur. Şekil 4.11'de, 7 farklı fonksiyonel terimden oluşmuş bir sık geçen örüntü görülmektedir. Bu uzun bir doğrusal patikanın sonundaki çatallanmadan oluşmaktadır. Örüntü çeşitli protein etkinliklerini barındırmaktadır: ligaz, transferaz, kinaz, enzim düzenleyici ve protein bağlanma etkinlikleri. Ağaç yapısında olan bu örüntünün 7 düğümü olup, ağda 15 gömüntüsü vardır.



Şekil 4.12. MAPK sinyal iletme patikası ile ilgili bir fonksiyonel etkileşim örüntüsü.

Hamur Mayası canlısının STRING ağında, yeterli desteği bulmuş ve istatistiksel olarak önemli olduğu gösterilmiş (Z-puanı=3,03) olan bir başka örüntü Şekil 4.12'de verilmiştir. Bu örüntünün bu şekilde verilen bir gömüntüsü MAPK patikası ile tamamen örtüşmektedir. Örüntünün GO terimleri küçük mavi kutucukların içinde verilmiştir. MAPK sinyal iletim patikasının üyesi olan 4 genin isimleri düğümlerin üstündeki yeşil kutucuklarda gösterilmiştir. Bu örüntünün 15

tane gömüntüsü olup, bunlardan bir tanesi düğümlerin altındaki kırmızı kutucuklarda gösterilmiştir. Bu gömüntü bilinen patikalarla kesişmemektedir. KCC4, çevresel hücre iskeleti ile birlikte hücre döngüsünün ilerlemesini düzenleyen bir kinazdır. KCC4 NAP1 ile birleşmekte ve NAP1 diğer iki protein ile etkileşmektedir. Bu tür analizler biyologların örüntüleri değişik ortamlar için çalışmalarına ve yorumlamalarına imkân verir.



Şekil 4.13. Vasküler iletim patikasındaki SNARE etkileşimleri ile ilişkili bir fonksiyonel etkileşim örüntüsü.

Şekil 4.13'de kabarcıklı nakliye patikasında bulunan SNARE etkileşimleri ile tamamen örtüşen bir gömüntüye sahip olan bir örüntü verilmiştir. Bu gömüntünün proteinleri yeşil kutucuklarda gösterilmiştir.

Keşfedilen bazı etkileşim örüntüleri daha önceden nitelendirilmemiş fonksiyonel parçalara karşılık geliyor olabilir, çünkü kullandığımız ağlar çok miktarda veri çıktısı veren deneylerden elde edilmişlerdir. Bu çalışmanın devamında,

biyolojik deneyler ile özgün gömüntülerin analiz edilmesi ve bunlardan elde edilen bulguların doğrulanması uygun olacaktır.

4.3. Etkileşim Kestirimi Sonuçları

Örüntülerin oylaması esasına dayalı etkileşim kestirimi algoritmamızın performansını ölçmek için *S. cerevisiae* organizmasının DIP veritabanından Ocak 2007'de aldığımız PPE ağını kullandık (Salwinski 2004). Ağı GO kütüphanesinin moleküler fonksiyon sınıfından terimlerle etiketledik. GO hedef seviyesini orijinal olarak bıraktık. Bu ağda sık geçen örüntüleri bulduktan sonra, bunları etkileşim tahmininde kullandık (yöntemimizin detayı için Metot bölümüne bakınız).

Algoritmamızın ikinci adımında 8 ve daha fazla sayıda gömüntüye sahip 186 adet örüntü bulundu. Üçüncü adımda bunlardan 38'inin önemli olduğu görüldü. Bu örüntülerden dört tanesi örnek olarak Şekil 4.14'de gösterilmiştir. Algoritma toplam 964 tane etkileşim tahmini yaparken, en yüksek doğru pozitif oranına şu parametrelerle erişmiştir: K=25, Z=5, N=4, T=10 ve F=8.

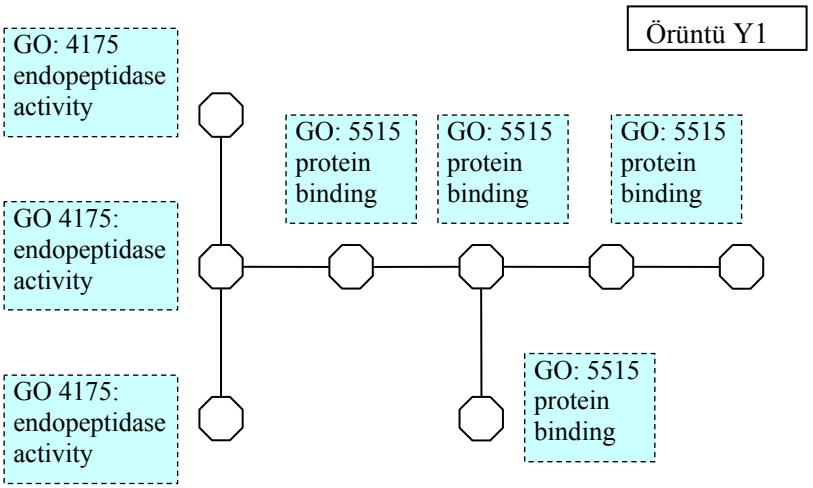
Yöntemimizin etkinliğini göstermek amacıyla Jansen (2003) tarafından tanımlanmış olan "Olasılık Oranını" kullandık:

$$L = (P+ / P-) / (G+ / G-)$$

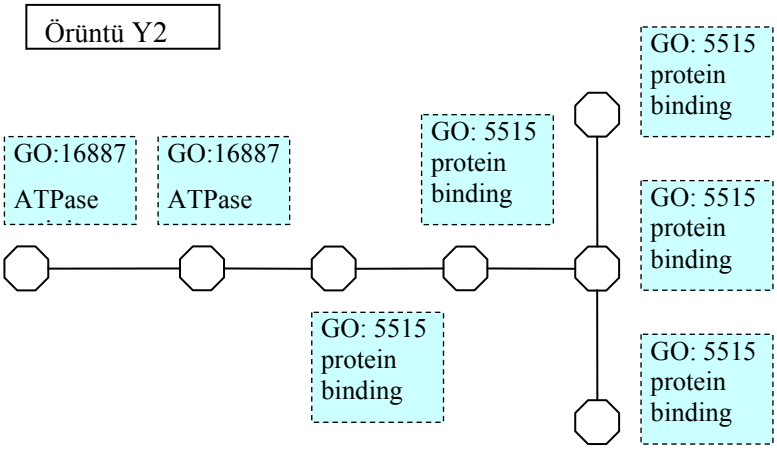
Burada L Olasılık Oranı, P+ doğru pozitiflerin sayısı, P- yanlış pozitiflerin sayısı, G- altın negatif kümesinin büyüklüğü, G+ ise altın pozitif kümesinin büyüklüğüdür. Çalışmamızda altın standart kümesi olarak, MIPS bileşik fihristi kullanılmıştır. Bu katalogda aynı bileşik içinde bulunan protein çiftleri etkileşenler yani altın pozitifler olarak alınmıştır.

Altın negatif kümesi olarak Jansen tarafından da kullanılan hücresel mahal bilgisini kullandık (Kumar 2002). Farklı hücresel mahallerde bulunan proteinleri etkileşmiyor kabul ederek altın negatif kümesini oluşturduk. Altın pozitif kümesinde 8265 adet, altın negatif kümesinde ise 2790366 adet etkileşim bulunmaktadır. Algoritmamız 395'i altın standart kümesi içinde bulunan toplam 964 adet etkileşim tahmininde bulundu. Bunlardan 287 tanesi pozitif, 108 tanesi ise negatiftir. Altın pozitif kümesi ile bizim tahmin kümemiz arasındaki örtüşme istatistiksel olarak

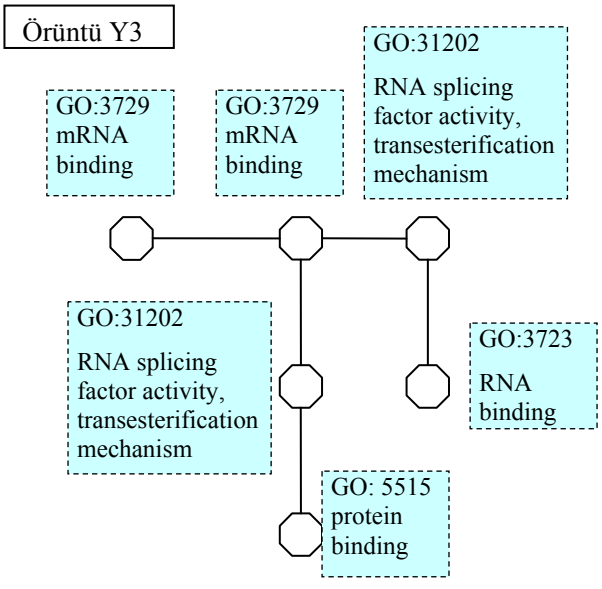
önemlidir (p-değeri<10^{-10}). Bu sonucun Olasılık Oranı 897 olmaktadır. Bu rakam Jansen ve ark. tarafından erişilen orandan (400) oldukça yüksektir.


Örüntü Y1
GO: 4175 endopeptidase activity
GO: 5515 protein binding
GO: 5515 protein binding
GO: 5515 protein binding
GO 4175: endopeptidase activity
GO: 5515 protein binding
GO 4175: endopeptidase activity

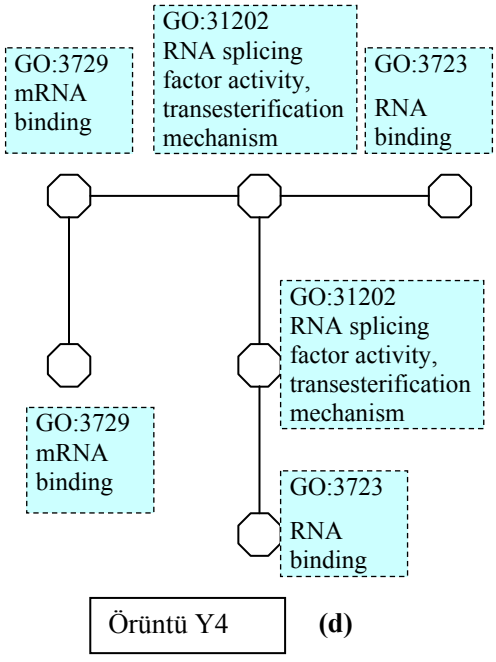
(a)


Örüntü Y2
GO: 5515 protein binding
GO:16887 ATPase
GO:16887 ATPase
GO: 5515 protein binding
GO: 5515 protein binding
GO: 5515 protein binding
GO: 5515 protein binding

(b)


Örüntü Y3
GO:3729 mRNA binding
GO:3729 mRNA binding
GO:31202 RNA splicing factor activity, transesterification mechanism
GO:31202 RNA splicing factor activity, transesterification mechanism
GO:3723 RNA binding
GO: 5515 protein binding

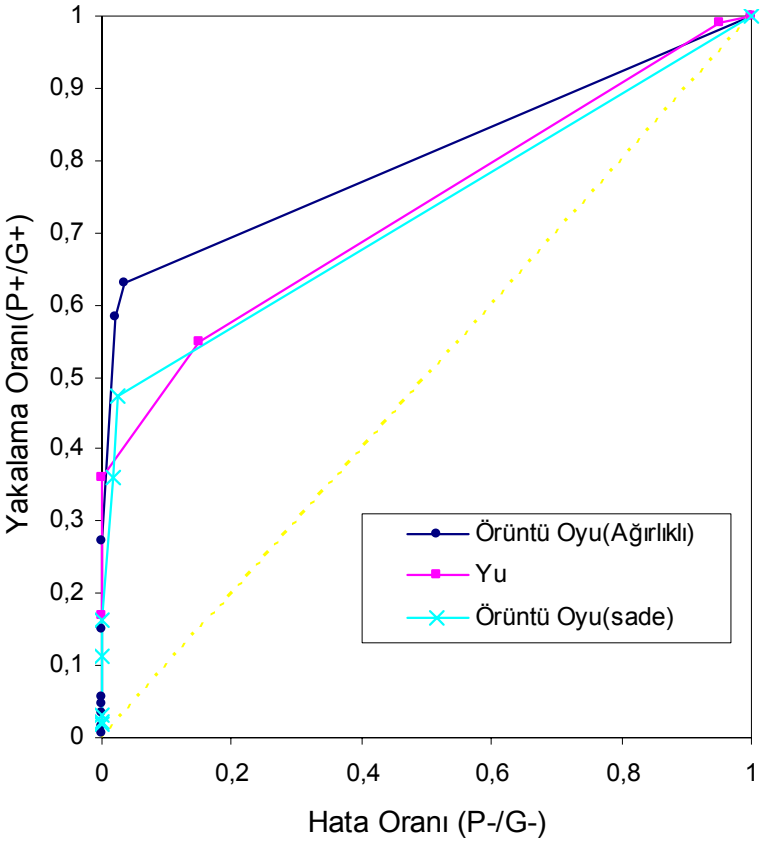
(c)


GO:3729 mRNA binding
GO:31202 RNA splicing factor activity, transesterification mechanism
GO:3723 RNA binding
GO:31202 RNA splicing factor activity, transesterification mechanism
GO:3729 mRNA binding
GO:3723 RNA binding
Örüntü Y4

(d)

Şekil 4.14. PPE ağında bulunan bazı önemli örüntüler.

Düğümlerin etiketleri olan GO moleküler fonksiyon terimleri kutucuklar içinde verilmiştir. Yöntemimizin etkinliğini göstermek için ikinci bir kıstas olarak ROC eğrisi (Egan 1975) çizilmiş ve Yu ve ark. (2006)'nın topoloji temelli takım tamamlama yönteminki ile karşılaştırılmıştır. Yöntemimizin performansı onlarınkini geçmektedir(Şekil 4.15).

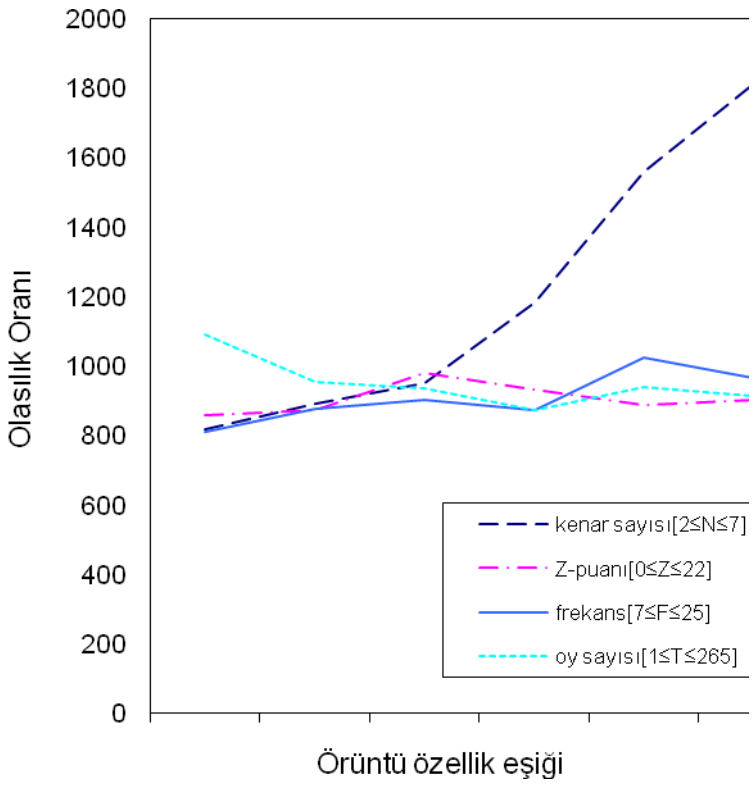


Şekil 4.15. Yu'nun yöntemi ile bizim örüntü oyuna dayalı yöntemimizin ROC eğrileri.

Çeşitli örüntü özelliklerinin yöntemimizin kestirim doğruluğunu ne derece etkilediğini anlamak için bir dizi deney yaptık. Bu örüntü özellikleri kenar sayısı, frekans (örtüşmeyen gömüntü sayısı), z-puanı ve oy sayısı(örüntünün olumlu oy verdiği kenar sayısı). Her deneyde bu özelliklerden birinin örüntü seçme eşik

değerini değiştirerek, diğer özellikleri sabit tuttuk. Bütün özellikler içinde sadece kenar sayısı özelliği Olasılık Oranı değerinde önemli ölçüde değişikliğe neden oldu. Farklı örüntü kümeleri kullanıldığında algoritmamız tarafından elde edilen farklı Olasılık Oranları Şekil 4.16.'da ifade edilmektedir. Burada, örüntü kümeleri örüntülerin özelliklerinin eşik değerleri değiştirilerek oluşturulmuştur (N,Z,F ve T). Her bir özellik için *x* ekseni belirtilen değer aralıklarında normalleştirilmiştir. Olasılık Oranı diğer özelliklerin eşik değerlerinin değişmesiyle neredeyse hiç değişmezken, gittikçe büyük örüntülerin (kenar sayısı yüksek olanların) oy kullanmasıyla çok belirgin biçimde artmaktadır. Sadece 7 ve daha üzeri sayıda kenarı olan örüntülerin oy vermesine izin verildiğinde Olasılık Oranı 1800'ü geçmektedir.

Bundan çıkarılabilecek sonuç, örüntünün kestirim doğruluğunun onun kenar sayısı ile birlikte artmakta olduğudur. Bir başka deyişle, büyük örüntüler tarafından kendilerine oy verilen kenarların doğru çıkma olasılığı daha küçükler tarafından oy verilenlere göre genel olarak daha yüksek olmaktadır.



Şekil 4.16. Farklı örüntü kümeleri kullanıldığında algoritmamız tarafından elde edilen farklı Olasılık Oranları.

Kestirimlerimizin kalitesini ve yöntemimizin etkinliğini daha fazla sınamak için Kiemer ve ark. (2007) tarafından derlenmiş farklı bir altın pozitif kümesi kullandık. Bu veri, literatürden elle derlenmiş 1777 etkileşimden oluşmaktadır. Bunlar az sayıda çıktı veren ve her deneyde en fazla 10 adet etkileşim belirtebilen deneylerden alıntılanmıştır. Yaptığımız 964 kestirimden 84 tanesi altın pozitif kümenin içinde kalmış olup (doğru pozitif), bu sonuç 2648 Olasılık Oranı değerine karşılık gelmektedir. Bu oran daha önce rapor edilmiş oranlardan yüksektir (Jansen 2003, Saito 2002, Albert 2004, Yu 2006).

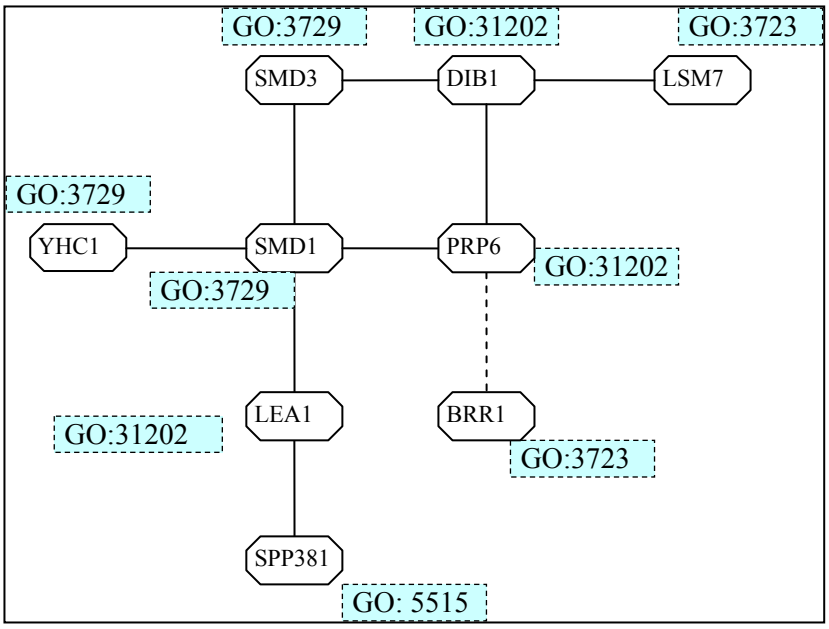
4.4. Etkileşim Tahmini – Biyolojik Örnekler

Bu bölümde 3.2.3.1 Bölümünde detayları verilen ve 4.3 Bölümünde deneysel sonuçları açıklanan yöntemle deşifre edilen bazı etkileşimlerden örnekler verilmektedir. Algoritmamız tarafından *S. Cerevisiae* organizması için tahmin edilen 964 etkileşim, PPE haritasının geneline yayılmış, pek çok farklı bileşikler ve moleküler fonksiyon sınıflarında bulunmuştur. Burada tahmin listemizin en üst sırasındaki ilk iki etkileşimi örnek olarak vereceğiz.

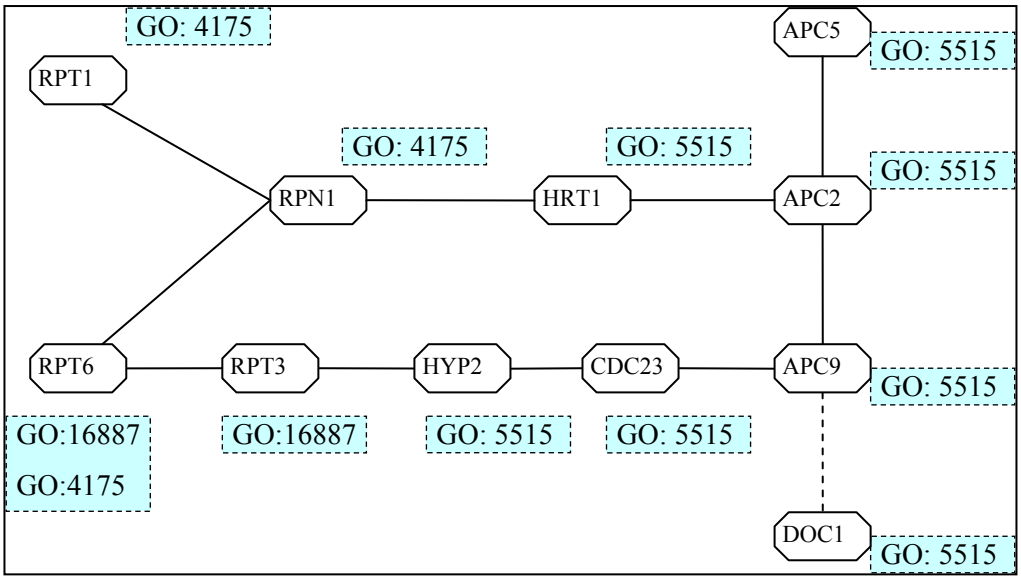
İlk örnek (PRP6 - BRR1) snRNP biyogenesis ile ilişkilidir. BRR1 bir RNA'ye bağlanan proteindir. PRP6, snRNP (U4/U6) ilişkili bir uç ekleme faktörüdür. Uç ekleme faktörleri mRNA dizilerindeki intron ve eksonları çıkarma işlevini yerine getiren proteinlerdir. Bunlar aynı protein bileşiğine aittirler. Büyük çaplı etkileşim deneylerin sonuçları olarak DIP veritabanında bu (PRP6-BRR1) etkileşim yer almamaktadır. Metodumuz, bu etkileşime oy veren 28 adet örüntü tespit etmiştir. Şekil 4.17'de PPE ağının bu etkileşime oy veren en büyük iki örüntünün gömüntülerinin bulunduğu parçası verilmiştir. Bu örüntüler Şekil 4.14'deki Y3 ve Y4 örüntüleridir.

Diğer örnek etkileşimimiz APC9-DOC1 çifti arasındadır. Bunlar ubiquitin-protein ligase bileşiğinin parçasıdırlar (Zachariae 1998). Şekil 4.18'de PPE ağının bu etkileşime oy veren en büyük iki örüntünün gömüntülerinin bulunduğu parçası

verilmiştir. Bu örüntüler Şekil 4.14'deki Y1 ve Y2 örüntüleridir. Dikkat edilirse her iki örnekte de örtüşme ($O_{Y_1,Y_2} = 0.12$ ve $O_{Y_3,Y_4} = 0.06$) ve benzeme miktarı ($D_{Y_1,Y_2} = 0.87$ ve $D_{Y_3,Y_4} = 0.96$) çok düşüktür.



Şekil 4.17. PRP6 ve BRR1 arasındaki etkileşime oy veren Y3 ve Y4 örüntülerinin durumunu gösteren PPE parçası.



Şekil 4.18. APC9 ve DOC1 arasındaki etkileşime oy veren Y1 ve Y2 örüntülerinin durumunu gösteren PPE parçası.

5. TARTIŞMA

Bu bölümde PPISpan'in kullanımını etkileyen bazı noktalara ve PPE ağlarında mümkün olabilecek başka uygulama biçimlerine değineceğiz. Öncelikle PPISpan'in örüntü arama işleminin sonuçlarını en büyük oranda etkileyen hususun girdi PPE ağının kalitesi olduğunu söylemeliyiz. Bilinmektedir ki, mevcut PPE ağlarının hepsi bol miktarda yanlış pozitif barındırmaktadır ve tamam olmaktan uzaktır (Mering ve ark. 2002). Gürültünün etkisini azaltmak amacıyla deneylerimizi hem deneysel olarak belirlenmiş PPE ağlarında hem de kestirimlerle ortaya konmuş PPE ağlarında denedik. Bu çalışmanın devamında başka PPE ağlarında PPISpan denenerek, değişik ağlardaki örüntülerin ne derece örtüştüğü değerlendirilebilir.

Bilinmelidir ki, PPISpan en uygunluğu garantilemeyen bir yapıya sahiptir. Özellikle, bir örüntünün mümkün olan tüm örtüşmeyen gömüntü kümeleri ortaya çıkartılır ve en yüksek gömüntü sayısına sahip olan küme ele alınacak olursa, bu örüntü için örtüşmeyen gömüntülerinin sayısının daha fazla olduğu durumlar görülebilir.

PPISpan tam eşleştirme yapılabilecek yani eksiksiz olan gömüntüleri PPE ağında saymak için tasarlanmıştır, dolayısıyla (yanlış eksi olması nedeniyle) bir kenarı eksik olan gömüntüleri atlamaktadır. Bu da doğal olarak yanlış eksilere yol açar. Öte yandan yanlış pozitif etkileşimler biyolojik deneylerle görülmeyecek örüntülerin bulunmasına da yol açıyor olabilir. Bu şekilde gürültü barındıran ağlar için yaklaşık ve benzer örüntü bulan veya yaklaşık ve benzer gömüntüleri de tarayan algoritmalar geliştirilebilir.

PPISpan algoritmasının örüntü çıktılarının kalitesini etkileyen bir başka unsur da proteinlerin etiketlendiği GO terimlerinin doğruluğunun kesinliği ve belirliliğidir. Bu çalışmamızda PPE ağına fazladan gürültü katmamak için elektronik olarak çıkarımlanmış terimlere yer verilmemiştir. Topolojinin yanında düğüm etiketleri de bir örüntünün anlamını ve belirliğini etkileyen bir unsurdur. Bu çalışmamızda GO Slim kütüphanesinin -aslında kaba bir sınıflandırması olan - moleküler fonksiyon ontolojisi kullanılmıştır. Bu kaba sınıflandırma çok belirgin olmayan örüntüler açığa çıkarmıştır. Dolayısıyla, detaylı bir biyolojik yorum yapılması zor olabilir. Öte

yandan bu çalışmamızla, değişik belirginlik seviyelerinden GO terimlerini kullanarak değişik seviyelerdeki etkileşim örüntülerini araştırabilecek bir yapı ortaya koymuş olduk.

GO moleküler fonksiyon sınıflandırmaları haricinde başka etiket kümeleri kullanılarak da örüntü araması yapmak ilginç sonuçlar verebilir. Örneğin, Hücresel parça betimlemelerini protein etiketlemede kullanarak, faydalı örüntüler (örneğin bir hücredeki birden fazla bölmeden geçen sinyal iletim patikaları) bulunabilir. Başka gen ölçekli sınıflamalar ve protein özellikleri de, bunlarla PPE ağını etiketlemek suretiyle, önemli etkileşim örüntülerini bulmak için kullanılabilir.

PPISpan birden fazla canlıda ortak olarak bulunan örüntüleri keşfetmek için kolaylıkla uyarlanarak kullanılabilir. Bu durumda bütün organizmalara ait PPE ağlarının bütünü GO terimleriyle etiketlenerek algoritmaya girdi olarak verilir. En düşük destek parametresinin yanı sıra bir örüntünün sık geçen sayılabilmesi için en az kaç farklı organizmada belirtilen desteğe sahip olması gerektiğini belirten ikinci bir destek parametresi girdi olarak verilir. Bu destek değerinden daha fazla sayıda PPE ağında sık olarak görülen örüntüler bu canlılar arasında paylaşılan örüntü olarak kabul edilip, yorumlanabilirler. GO terimleri canlıya özel olmadığı için, PPE ağlarının bu amaçla etiketlenmeleri için gayet uygundurlar.

Protein fonksiyon tahmininde örüntüler

Protein fonksiyonu, biyofiziksel olarak (ör. kinaz), biyokimyasal olarak (ör. Bir patikanın parçası olmak), biyolojik açıdan (ör. gelişim), patolojik olarak (ör. Tümör baskılayıcılık) veya başka açılardan proteinin hücresel işlevi demektir. Proteinlerin fonksiyonlarının belirlenmesi biyoloji biliminin önemli problemleri arasındadır ve bu yolda henüz alınması gereken çok mesafe vardır. Örneğin en fazla çalışılmış Hamur Mayası organizmasının bile proteinlerinin ancak %60'ı fonksiyonel olarak sınıflandırılabilmiştir. Proteinlerin fonksiyonlarının belirlenmesi ihtiyacının baskısıyla, bu problemi çözmeye yönelik geliştirilmiş hesaba dayalı tekniklerde büyük artış görülmüştür. Bunun nedeni canlının içinde yapılan fonksiyon belirleme deneylerinin çok pahalı olmasıdır. Geliştirilen hesaba dayalı yöntemler çoğunlukla, protein dizisi, protein yapısı, genom bilgisi, filogenetik yanay, protein etkileşim ağları, biyomedikal literatür bilgilerini kullanmışlardır. Bunların yanında, etkileşim

örüntülerinin protein fonksiyonunu belirlemede kullanılabileceği ifade edilmiştir (Alon 2003). Bu bölümde, tezimizde ifade ettiğimiz yöntemle elde edilecek örüntülerin bu amaçla nasıl kullanılabileceği üzerinde duracağız. Nitekim önceden yapılmış topolojiye dayalı çalışmalardan bazıları protein dizisine dayalı olanlara göre, özellikle dizi benzerliğinin olmadığı durumlarda, daha iyi performans sergilemişlerdir (Schwikowski 2003, Deng 2003). Dolayısı ile proteinlerin topolojik çevrelerinin benzerliklerini fonksiyon tahmininde kullanmak, tahminlerin keskinliğini artırabilir.

Önceki bölümlerde önerdiğimiz protein etkileşimlerini tahmin etmede kullandığımız örüntü oylamasına dayalı mekanizmanın benzerini proteinlerin fonksiyonlarını tahminde de kullanabiliriz. Bu şekilde, proteinlerin topolojik çevrelerini yansıtan örüntü bilgisini kullanarak benzer topolojili proteinlerin fonksiyonlarının de benzer olacağından hareketle kestirimde bulunabiliriz. Nitekim bir örüntü içinde barındırdığı her bir düğüm için ağda sık görülen bir topolojiyi belirtmektedir. Bütün örüntülerin şahitliği (oy vermesi) ile yapılacak fonksiyon kestirimi ile önceden bilinmeyen pek çok proteinin fonksiyonu ortaya çıkarılabilir. Bu amaçla, "Örüntü oylarıyla etkileşim kestirimi" başlıklı Bölüm 3.2.3'de açıklanan algoritma ve güven ölçütlerinden esinlenerek yeni bir yöntem ortaya konabilir.

Belirtmeliyiz ki, protein fonksiyonu tahmini, protein etkileşim tahmininden temel farklılıklar içermektedir. Örneğin protein etkileşimleri ikilidir. Yani, belli bir protein çifti arasında etkileşim vardır veya yoktur. Öte taraftan protein fonksiyonu (belki de yüzlerce) aday etiketler arasından bir alt kümenin bulunması işlemini gerektirir. Bu ve benzeri farklılıklar, protein fonksiyonu kestirim algoritmasının karmaşıklığını ve performansını etkileşim tahmini algoritmasınınkine göre önemli ölçüde değiştirebilir.

6. SONUÇ VE ÖNERİLER

Bu doktora tezinde, çizge madenciliği yöntemleriyle protein-protein etkileşim ağlarında fonksiyonel etkileşim örüntülerini keşfetmek maksadıyla, PPISpan adını verdiğimiz yeni bir sık geçen örüntü bulma algoritması önerilmiştir. PPISpan, Gen Ontolojisi kütüphanesinden alınan moleküler fonksiyon terimleriyle ağdaki proteinleri tekil olmayacak şekilde etiketlemekte ve bu ağda sık bulunan ve önemli olan fonksiyonel etkileşim örüntülerini tespit etmektedir. PPISpan algoritması Hamur Mayası (*Saccharomyces cerevisiae*) canlısının deneyler ile ortaya çıkarılmış ve kestirimlenmiş protein-protein etkileşim ağlarında denenmiştir. GO moleküler fonksiyon terimleriyle etiketlenmiş bu ağlarda potansiyel olarak ilginç olabilecek pek çok örüntü tespit edilmiştir. Bu örüntüler hücrenin protein etkileşim ağının yapısal düzenine yeni bir bakış açısı önermektedir. Bulunan örüntülerin çoğu ağaç yapısında olup çemberler nadirdir. Bu bulgumuz, yaklaşık ağaç yapılarını bulan ve dolayısıyla daha hızlı (anında etkileşimli) sonuçlar veren ağaç madenciliği algoritmalarının aynı amaçla kullanılabileceğini göstermektedir.

Bu çalışmanın ilerideki aşamalarında, (insan gibi) başka organizmaların da protein etkileşim ağlarında örüntü bulmayı hedefliyoruz (Mathivanan ve ark. 2006). Ayrıca daha önce kullanılan sık eleman küme madenciliği teknikleriyle ilgili yöntemleri (Inokuchi 2004) kullanarak "genel örüntüleri" araştırmayı düşünüyoruz. Nitekim bu çerçevede yapılmış çalışmalar GO aşamalı sisteminin ağaç benzeri yapısıyla uyuşmaktadırlar.

Bu çalışmada etkileşim tahmini için geliştirilen yöntemle, sıkça geçen örüntüler gibi PPE ağları ile ilgili fazladan bilgileri kullanarak, doğruluk oranı yüksek etkileşim tahmini yapılabileceği gösterilmiştir. Belirtmeliyiz ki bu sonuç, ikili etkileşim bilgileri dışındaki ağla ilgili bilgilerin protein özelliklerini kestirmede pek işe yaramadıkları yönündeki hükümle (Chen 2007) farklılık arz etmektedir. Çizge örüntülerinin faydalı olduğunu, onları protein-protein etkileşim tahmininde başarılı bir şekilde kullanarak gösterdik. Hamur Mayası organizmasının etkileşim ağında bulduğumuz örüntüleri kullanarak, bu örüntülerin eksik kenarlı olan gömüntülerini belirleyerek, eksik kenarları aday etkileşimler olarak ilan ettik.

Bunları geliştirdiğimiz ölçütler yardımıyla sıralayarak aday güven listemizi oluşturan yeni bir kestirim algoritması önerdik. Bu algoritmayı kullanarak yaptığımız deneyler sonucunda, GO terimleri ile etiketlenmiş PPE ağında pek çok tahminde bulunduk. Tahminlerimizi altın standardı veri kümeleri ile sınayarak yöntemimizin iyi bir kestirici olduğunu gösterdik. Bu performansı hiçbir fazladan biyolojik veriye başvurmadan sadece topoloji bilgilerinden yararlanarak yakaladık. Buna ek olarak, kestirim sürecimizin mekanizması tamamen şeffaf ve yoruma açık bir yapıya sahiptir.

7. KAYNAKLAR

Albert I, Albert R: Conserved network motifs allow protein–protein interaction prediction. Bioinformatics 2004, 20:3346-3352.

Albert R, Jeong H, Barabasi A-L. (2000) Error and attack tolerance of complex networks. Nature 406, 378-382.

Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33: D418–D424.

Alon U: Biological networks: the tinkerer as an engineer. Science, 301:1866–67, 2003.

Aloy P, Russell, RB: Interprets: protein interaction prediction through tertiary structure. Bioinformatics, 19(1), 161–2, 2003.

Asthana S, King OD, Gibbons FD, Roth FP: Predicting protein complex membership using probabilistic network reliability. Genome Res 2004, 14(6):1170-1175.

Bader G, Hogue C: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003, 4:2.

Bader GD, Hogue CWV: Analyzing yeast protein-protein interactions data obtained from different sources. Nature Biotechnology 2002, 20:991-997.

Bader JS, Chaudhuri A, Rothberg JM, Chant J : Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol (2004) 22: 78–85.

Bartel PL, Roecklein JA, SenGupta D, Fields S (1996) A protein linkage map of Escherichia coli bacteriophage T7. Nat Genet 12: 72–77.

Bartel PL, Roecklein JA, SenGupta D, Fields S: A protein linkage map of Escherichia coli bacteriophage T7. Nature Genet 1996. 12: 72–77.

Ben-Hur A, Noble WS : Kernel methods for predicting protein– protein interactions. Bioinformatics (2005) 21: i38–46.

Berg J, Lassig M: Local graph alignment and motif search in biological networks. PNAS, 2004, 101 14689-14694, Sept. 2004

Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, et al. (2000) The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 7: 957–959.

Besemann C, Denton A, Yekkirala A: Differential association rule mining for the study of protein-protein interaction networks. BIOKDD04: 4th Workshop on Data Mining in Bioinformatics (with SIGKDD Conference) 2004:72-80.

Bork P, Jensen LJ, Mering C, Ramani AK, Lee I, Marcotte EM: Protein interaction networks from yeast to human. Current Opinion in Structural Biology 2004, 14:292–299

Bradford JR, Westhead DR : Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics (2005) 21: 1487–1494.

Brohee S, van Helden J: Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 2006, 7:488.

Bunke H, Shearer K: A graph distance metric based on the maximal common subgraph. Pattern Recognition Letters 1998, 19: 255–259.

Cakmak A, Ozsoyoglu G: Mining biological networks for unknown pathways. Bioinformatics 2007, 23(20):2775-2783.

Chen J, Hsu W, Lee ML, Ng SK: Increasing confidence of protein interactomes using network topological metrics. Bioinformatics 2006, 22(16):1998-2004.

Chen P, Deane CM, Reinert G: A statistical approach using network structure in the prediction of protein characteristics. Bioinformatics 2007, 23(17):2314-2321.

Chen XW, Liu M : Prediction of protein–protein interactions using random decision forest framework. Bioinformatics (2005) 21: 4394–4400.

Chou KC: Prediction of protein structural classes and subcellular locations. Curr Protein Pept Sci, 1(2), 171–208, 2000.

Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ: Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics 2007, 6(3):439-450.

Cordella LP, Foggia P, Sansone C, Vento M: A (sub)graph isomorphism algorithm for matching large graphs. IEEE Trans. Pattern Anal Mach Intell 2004, 6(10):1367-1372.

Deane CM, Salwinski L, Xenarios I, Eisenberg D: Protein Interactions: Two methods for Assessment of the Reliability of High Throughput Observations. Mol. Cell. Proteomics 2002, 1: 349-356.

Dohkan S, Koike A, Takagi T: Prediction of Protein-Protein Interactions Using Support Vector Machines. Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering 2004.

Duan XJ, Xenarios I, Eisenberg D (2002) Describing biological protein interactions in terms of protein states and state transitions: The LiveDIP database. Mol Cell Proteomics 1: 104–116.

Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M: Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends in Genetics 2002, 18(10):529-536.

Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 1998, 95:14863-14968

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: Protein function in the post-genomic era. Nature 405: 823–826 (2000).

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86–90 (1999).

Ermolaeva MD, White O, Salzberg SL: Prediction of operons in microbial genomes. Nucleic Acids Res 29: 1216–1221 (2001).

Fariselli P, Pazos F, Valencia A, Casadio R : Prediction of protein– protein interaction sites in heterocomplexes with neural networks. Eur J Biochem (2002) 269: 1356–1361.

Finley RL Jr, Brent R: Interaction mating reveals binary and ternary connections between Drosophila cell cycle regulators. PNAS 1994 91: 12980–12984.

Finn RD, Marshall M, Bateman A (2005) iPfam: Visualization of protein– protein interactions in PDB at domain and amino acid resolutions. Bioinformatics 21: 410–412.

Gardiner WP: Statistics for the Biosciences, data analysis using minitab software. Prentice Hall 1997

Gavin A, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick J, Michon A, Cruciat C, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier M, Copley R, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002, 415(6868):141-7.

Gene Ontology Consortium: Gene Ontology: tool for the unification of biology. Nature Genet 2000, 25:25-29.

Gertz J, Elfond G, Shustrova A, Weisinger M, Pellegrini M ve ark. : Inferring protein interactions from phylogenetic distance matrices. Bioinformatics 19:2039–2045 (2003) .

Ghazizadeh S, Chawathe S: SEuS: Structure Extraction using Summaries. In proceedings of the 5th International Conference on Discovery Science. November 2002, Lecture Notes in Computer Science (LNCS) 2534.

Gilchrist MA, Salter LA, Wagner A : A statistical framework for combining and interpreting proteomic datasets. Bioinformatics (2004) 20: 689– 700.

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. (2003) A protein interaction map of Drosophila melanogaster. Science 302: 1727– 1736.

Girvan M, Newman MEJ: Community structure in social and biological networks. PNAS, June 11, 2002 , vol. 99, no. 12, 7821-7826

Girvan M, Newman MEJ: Community structure in social and biological networks. PNAS 2002, vol. 99, no. 12, 7821-7826

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE : Co-evolution of proteins with their interaction partners. J Mol Biol (2000) 299: 283–293.

Gross J, Yellen J: Graph Theory and its applictions. CRC Press 1998

Gross J, Yellen J: Handbook of Graph Theory, CRC Press 2003.

Gudes E, Shimony SE, Vanetik NE: Discovering Frequent Graph Patterns Using Disjoint Paths. IEEE Transactions on Knowledge and Data Engineering, VOL. 18, NO. 11, NOV 2006

Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, et al. (2006) MPact: The MIPS protein interaction resource on yeast. Nucleic Acids Res 34: D436–D441.

Hirsh E, Sharan R: Identification of conserved protein complexes based on a model of protein network evolution. Bioinformatics 2007, 23(2):e170-176.

Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T. (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast, 18, 523–531.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sørensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CWV, Figeys D, Tyers M: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 2002, 415:180-183.

Hu H, Yan X, Huang Y, Han J, Zhou XJ: Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics 2005, 21(Suppl 1):i213-221.

Inokuchi A: Mining generalized substructures from a set of labeled graphs. In ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04) Washington, DC, USA: IEEE Computer Society; 2004:415-418.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences 2001, 98(8):4569-4574.

Jansen R, Greenbaum D, Gerstein M: Relating whole-genome expression data with protein–protein interactions. Genome Res 2002 12: 37–46.

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: A bayesian networks approach for predicting protein-protein interactions from genomic data. Science 2003, 302(5644):449-453.

Jingchun Chen and Bo Yuan: Detecting functional modules in the yeast protein–protein interaction network. Bioinformatics , Vol. 22 no. 18 2006, pages 2283–2290

Jones RB, Gordus A, Krall JA, MacBeath G (2006) A quantitative protein interaction network for the ErbB receptors using protein microarrays. Nature 439: 168–174.

Jothi R, Cherukuri PF, Tasneem A, Przytycka TM : Co-evolutionary analysis of domains in interacting proteins reveals insights into domain– domain interactions mediating protein–protein interactions. J Mol Biol (2006) 362: 861–875.

Kanehisa M, Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucl Acids Res 2000, 28:27-30.

Kannan R, Tetali P, Vempala S: Simple Markov-chain algorithms for generating bipartite graphs and tournaments. Random Struct. Algorithms 14, 4 (Jul. 1999), 293-308

Kashtan N, Itzkovitz S, Milo R, Alon U: Mfinder tool guide. Technical report, Department of Molecular Cell Biology and Computer Science and Applied Mathematics, Weizman Institute of Science, Israel; 2002.

Kashtan N, Itzkovitz S, Milo S, Alon U: Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs; Bioinformatics, Vol. 20 no. 11 2004, 1746–1758

Kiemer L, Costa S, Ueffing M, Cesareni G: WI-PHI: A weighted yeast interactome enriched for direct physical interactions. Proteomics 2007, 7(6):932-943.

Kirac M, Ozsoyoglu G, Yang J: Annotating proteins by mining protein interaction networks. Bioinformatics, Vol. 22 no. 14 2006, pages e260–e270

Koike A, Takagi T : Prediction of protein–protein interaction sites using support vector machines. Protein Eng Des Sel (2004) 17: 165–173.

Koyuturk M, Kim Y, Subramaniam S, Szpankowski W, Grama A: Detecting conserved interaction patterns in biological networks. Journal of Computational Biology 2006, 13(7):1299-1322.

Koyuturk M, Kim Y, Subramaniam S, Szpankowski W, Grama A: Detecting conserved interaction patterns in biological networks. Journal of Computational Biology 2006, 13(7):1299-1322.

Koyutürk M, Grama A, Szpankowski W: An efficient algorithm for detecting frequent subgraphs in biological networks. Bioinformatics 2004, 20(Suppl 1):200-207.

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637–643.

Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y, et al: Subcellular localization of the yeast proteome. Genes Dev. 2002, 16:707–719

Kuramochi M, Karypis G: Finding frequent patterns in a large sparse graph. Data Min Knowl Discov 2005, 11(3):243-271.

Lanckriet GRG, Deng M, Cristianini N, Jordan MI, Noble WS: Kernel- based data fusion and its application to protein function prediction in yeast. Proceedings of the Pacific Symposium on Biocomputing 2004:300-311.

Lee I, Date SV, Adai AT, Marcotte EM: A probabilistic functional network of yeast genes. Science 2004, 306(5701):1555-1558.

Lee I, Date SV, Adai AT, Marcotte EM:A Probabilistic Functional Network of Yeast Genes. Science VOL 306 26 Nov 2004, p.p. 1555-1558

Letovsky S, Kasif S: Predicting protein function from protein/ protein interaction data: a probabilistic approach. Bioinformatics 2003, 19(Suppl 1):i197-204.

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al. (2004) A map of the interactome network of the metazoan C. elegans. Science 303: 540–543.

Lin CY, Chen CL, Cho CS, Wang LM , Chang CM, Chen PY, Lo CZ, and Hsiung CA: "hp-DPI: Helicobacter pylori Database of Protein teractomes--embracing experimental and inferred interactions". Bioinformatics, April 1, 2005; 21(7): 1288 - 1290.

Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP: Molecular Cell Biology. W.H. Freeman and Company 2004.

Lourdes Peña-Castillo, Murat Tasan, Chad L Myers, Hyunju Lee, Trupti Joshi, Chao Zhang, Yuanfang Guan, Michele Leone, Andrea Pagnani, Wan Kyu Kim, Chase Krumpelman, Weidong Tian, Guillaume Obozinski, Yanjun Qi, Sara Mostafavi, Guan Ning Lin, Gabriel F Berriz, Francis D Gibbons, Gert Lanckriet, Jian Qiu, Charles Grant, Zafer Barutcuoglu, Olga G Troyanskaya, ve ark.: A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biology 2008, 9(Suppl 1):S2 (27 June 2008)

Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH: Modular organization of protein interaction networks. Bioinformatics
2007, 23(2):207-214.

M. Deng, F. Sun, and T. Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. PSB, 2003.

MacBeath G, Schreiber SL (2000) Printing proteins as microarrays for high-throughput function determination. Science 289: 1760–1763.

Mahdavi M, Lin YH: False positive reduction in protein-protein interaction predictions using gene ontology annotations. BMC Bioinformatics 2007, 8:262.

Marcotte CJ, Marcotte EM : Predicting functional linkages from gene fusions with confidence. Appl Bioinformatics 1: 93–100 (2002).

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO ve ark. : Detecting protein function and protein–protein interactions from genome sequences. Science 285: 751–753 (1999).

Martin S, Roe D, Faulon J: Predicting protein–protein interactions using signature products. Bioinformatics 2005, 21(2):218-226

Maslov S, Sneppen K: Specificity and Stability in Topology of Protein Networks Science 296, 910 (2002).

Mathivanan S, Periaswamy B, Gandhi T, Kandasamy K, Suresh S, Mohmood R, Ramachandra Y, Pandey A: An evaluation of human protein- protein interaction data in the public domain. BMC Bioinformatics 2006, 7(Suppl 5):S19.

McKay BD: Practical graph isomorphism. Congressus Numerantium 1981, 30:45-87.

Mendenhall W, Sincich T: Statistics for Engineering and the Sciences. Prentice Hall 1994

Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P: STRING 7-recent developments in the integration and prediction of protein interactions. Nucleic Acids Research 2007:D358-D362.

Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, et al. (2005): STRING: Known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res 33: D433–D437.

Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 2002, 417:399-403.

Montgomery DC, Runger GC: Applied statistics and probability for engineers. John Wiley and Sons Inc. 2003

Moreno-Hagelsieb G, Collado-Vides J: A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 18: S329–S336 (2002).

Nabieva E, Jim K, Agarwal A, Chazelle B:Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics Vol. 21 Suppl. 1 2005, pages i302–i310

Oyama T, Kitano K, Satou K, Ito T: Extraction of knowledge on protein-protein interaction by association rule discovery. Bioinformatics 2002, 18(5):705-714.

Pandey J, Koyuturk M, Kim Y, Szpankowski W, Subramaniam S, Grama A: Functional annotation of regulatory pathways. Bioinformatics 2007, 23(13):i377-386.

Pandey J, Koyutürk M, Szpankowski W, Grama A: Annotating pathways of interaction networks. Proceedings of the Pacific Symposium on Biocomputing 2008, 13:153-165.

Patil A, Nakamura H: Filtering high-throughput protein-protein interaction data using a combination of genomic features. BMC Bioinformatics 2005, 6:100.

Pazos F, Valencia A : Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng (2001) 14: 609–614.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285–4288 (1999).

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, et al. (2003) Development of human protein reference database as an initial platform for

approaching systems biology in humans. Genome Res 13: 2363–2371.
Przulj N, Wigle D, Jurisica I: Functional topology in a network of protein interactions. Bioinformatics 2004, 20(3):340-348.
Qi Y, Bar-Joseph Z, Klein-Seetharaman J : Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins (2006) 63: 490–500.
Qi Y, Bar-Joseph Z, Klein-Seetharaman J: Evaluation of Different Biological Data and Computational Classification Methods for Use in Protein Interaction Prediction. PROTEINS: Structure, Function, and Bioinformatics 2006, 63:490–500.
Qi Y, Klein-Seetharaman J, Bar-Joseph Z : Random forest similarity for protein–protein interaction prediction from multiple sources. Pac Symp Biocomput(2005): 531–542.
R. Milo, S. Itzkovitz, and N. Kashtan. Superfamilies of designed and evolved networks. Science, 303(5663):1538– 1542, 2004.
Ramani AK, Marcotte EM : Exploiting the co-evolution of interacting proteins to discover interaction specificity. J Mol Biol (2003) 327: 273–284.
S. Itzkovitz, R. Milo, and N. Kashtan. Coarse-graining and self-dissimilarity of complex networks. Phys. Rev. E, 71(016127), 2005.
Saito R, Suzuki H, Hayashizaki Y: Construction of reliable protein-protein interaction networks with a new interaction generality measure. Bioinformatics 2002, 19:756–763.
Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J: Operons in Escherichia coli: Genomic analyses and predictions. Proc Natl Acad Sci U S A (2000) 97: 6652–6657.
Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32: D449–D451.
Scholkopf B, Tsuda K, Vert JP: Kernel methods in computational biology. Cambridge, Massachusetts: MIT Press 2004.
Schwikowski B, Uetz P, Fields S: A network of protein-protein interactions in yeast. Nature Biotechnol,18:623–627, 2003.
Scott J, Ideker T, Karp RM, Sharan R: Efficient algorithms for detecting signaling pathways in protein interaction networks. RECOMB 2005:1-13.
Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: From the Cover: Conserved patterns of protein interaction in multiple species. Proceedings of the National Academy of Sciences 2005, 102(6):1974-1979.
Shoemaker BA, Panchenko AR: Deciphering Protein–Protein Interactions.Part I. Experimental Techniques and Databases. PLOS Computational Biology, March 2007, 3-3-e42
Shoemaker BA, Panchenko AR: Part II. Computational Methods to Predict Protein and Domain Interaction Partners. PLoS Computational Biology 2007, vol. 3, 4-595.
Smith GP (1985) Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. Science 228: 1315–1317.
Sole RV, Pastor-Satorras R, Smith E, Kepler TB: A model of large-scale proteome evolution. Advances in Complex Systems 5, 43 (2002)
Spirin V, Mirny LA: Protein complexes and functional modules in molecular

networks. Proceedings of the National Academy of Sciences 2003, 100(21):12123-12128.

Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksöz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE: A human protein-protein interaction network: a resource for annotating the proteome. Cell. 2005 Sep 3;122(6):957-68

Strogatz, S. (2001) Exploring complex networks. Nature 410, 268-276.

Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D: Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization: A combined computational approach. Genome Biol 4: R59 (2003).

Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker T: A direct comparison of protein interaction confidence assignment schemes. BMC Bioinformatics 2006, 7:360.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: A comprehensive analysis of protein- protein interactions in Saccharomyces cerevisiae. Nature 2000, 403:623-627.

Ullmann JR: An algorithm for subgraph isomorphism. J ACM 1976, 23:31-42.

Valente AXCN, Cusick ME: Yeast Protein Interactome topology provides framework for coordinated-functionality. Nucl Acids Res 2006, 34(9):2812-2819.

Valiente G: Algorithms on trees and graphs. Springer 2002, s. 351

Vanetik N, Gudes E, Shimony SE: Computing Frequent Graph Patterns from Semistructured Data : Proc. of the 2002 IEEE International Conference on Data Mining (ICDM'02), Page: 458

Vanetik N, Gudes E: Mining Frequent Labelled and Partially labelled graph patterns Proceedings of the 20th International Conference on Data Engineering (ICDE' 2004)

Vazquez A, Flammini A, Maritan A, Vespignani A: Modeling of Protein Interaction Networks. Complexus 2003;1:38-44

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF ve ark.: Protein interaction mapping in C. elegans using proteins involved in vulval development. Science 2000 287: 116–122.

Washio T, Motoda H: State of the Art of Graph based Data Mining. ACM SIGKDD Explorations Newsletter , Volume 5 , Issue 1 (July 2003)

Wernicke S, Rasche F: Fanmod: a tool for fast network motif detection. Bioinformatics, 22(9):1152–1153, 2006.

Wernicke S: A faster algorithm for detecting network motifs. WABI 2005:165-177.

Worlein M, Meinl T, Fischer I, Philippsen M: A Quantitative Comparison of the Subgraph Miners MoFa, gSpan, FFSM, Gaston. PKDD 2005, LNAI 3721, pp. 392–403, 2005.

Wu X, Zhu L, Guo J, Zhang D, Lin K: Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations. Nucleic Acids Res. 2006, 34(7):2137-2150.

Yamanishi Y, Vert JP, Kanehisa M : Protein network inference from multiple

genomic data: A supervised approach. Bioinformatics (2004) 20: I363– I370.

Yan X, Han J: gSpan: graph-based substructure pattern mining. In ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02) Washington, DC, USA: IEEE Computer Society; 2002:721.

Yanai I, Derti A, Delisi C : Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes. Proc Natl Acad Sci U S A 98: 7940–7945 (2001).

Yang G: Computational aspects of mining maximal frequent patterns. Elsevier, Theoretical Computer Science 362 (2006) 63 – 85

Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, et al. (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. Mol Syst Biol 1: 0026.

Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H: Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. Proceedings of the National Academy of Sciences 2004, 101(16):5934-5939.

You CH, Holder LB, Cook DJ: Application of graph-based data mining to metabolic pathways. In ICDMW '06: Proceedings of the Sixth IEEE International Conference on Data Mining – Workshops Washington, DC, USA: IEEE Computer Society; 2006:169-173.

Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S,Vidal M, M Gerstein: Annotation Transfer Between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs. Genome Res. 2004, 14, 1107–1118.

Yu H, Paccanaro A, Trifonov V, Gerstein M: Predicting interactions in protein networks by completing defective cliques. Bioinformatics, Vol. 22 no. 7 2006, 823–829

Zachariae W, Shevchenko A, Andrews PD, Ciosk R, Galova M, Stark MJ, Mann M, Nasmyth K. “Mass spectrometric analysis of the anaphase-promoting complex from yeast: identification of a subunit related to cullins.” *Science* 1998, 279(5354):1216-9.

Zhang LV, Wong SL, King OD, Roth FP : Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics (2004) 5: 38.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, et al. (2001) Global analysis of protein activities using proteome chips. Science 293: 2101–2105.