TÜBİTAK

Proteomların ve Genomların Bilgi İçeriklerinin Modellenmesi

Program Kodu: 3001

Proje No: 114F504

Proje Yürütücüsü:
Doç. Dr. Sertaç EROĞLU

AĞUSTOS 2017
ESKİŞEHİR

ÖNSÖZ

Bu sonuç raporu, TUBİTAK 3001 “Başlangıç Ar-Ge Projeleri Destekleme Programı” tarafından 01/05/2015 tarihinde onaylanarak yürürlüğe giren; 114F504 numaralı ve “Proteomların ve Genomların Bilgi İçeriklerinin Modellenmesi” başlıklı projeye aittir. Projede araştırma sorusunu cevaplamak üzere kullanılan yaklaşım biyoinformatik yöntemleri içermektedir.

Canlı organizmalar, oldukça karmaşık yaşamsal faaliyetlerin gerçekleştiği mükemmel fiziksel sistemlerdir. Bu açıdan bakıldığında günümüzün en teknolojik cihazlarının işlevsel kompleksliği en ilkel bir canlı organizmanınki ile bile karşılaştırıldığında oldukça basit kalmaktadır. Bilim tarihi boyunca canlı organizmaların işleyiş prensiplerini anlamak için yapılan pek çok çalışma sonucunda uzunca bir yol kat edilmiş olmasına rağmen, bu sistemler hakkındaki bilgi düzeyimiz hala oldukça sınırlıdır. Canlı organizmaların işleyişine dair en büyük gizemlerden birisi; organizmaların her türlü metabolik faaliyetlerini yöneten moleküler bilgi depolama ve bilgi iletimi süreçleridir. Gelişen bilgisayar ve bilgi işleme teknolojisi ile günümüzde moleküler bilgi depolama ve bilgi iletimi süreçlerini analiz edebilmek mümkün hale gelmiştir. Bu çalışmalar Biyoinformatik olarak bilinen modern bir disiplinlerarası çalışma alanının doğmasına neden olmuştur.

Biyoinformatik tekniklerin biyolojik sistemlerin çalışma prensiplerini anlamak üzere yeni yaklaşımlar geliştirme hususunda rüştünü ispatlamış olması, bu tekniklerin günümüzde klinik tedavi süreçlerinde de sıklıkla kullanılması sonucunu doğurmaktadır. Dolayısıyla, bu alanda yapılmakta ve yapılacak olan her çalışma ve bu çalışmaların literatüre olan katkıları biyolojik sistemlerin çalışma prensiplerinin daha iyi anlaşılması ve insan hayatının daha kaliteli sürdürülmesi açısından oldukça önemlidir.

Bu nedenle proje yürütücüsü olarak proje önerimi kabul ederek bu araştırmayı sonuçlandırmam için sağlamış olduğu finansal destek için TÜBİTAK’a ve projenin mali idaresi sürecine katkı sağlayan TÜBİTAK ve Eskişehir Osmangazi Üniversitesi İdari ve Mali İşler Daire başkanlığı personellerine teşekkür ederim. Aynı zamanda proje başvurusu değerlendirme, gelişme raporları ve sonuç raporu inceleme süreçlerinde değerli vakitlerini ayırarak görüşleriyle katkıda bulunan alanında uzman hakem hocalarıma teşekkürü borç bilirim.

Doç. Dr. Sertaç EROĞLU
Eskişehir Osmangazi Üniversitesi
Fizik Bölümü

Ağustos 2017

TÜBİTAK

İÇİNDEKİLER

ÖNSÖZ i

İÇİNDEKİLER ii

TABLO VE ŞEKİLLER LİSTESİ iii

ÖZET v

ABSTRACT vii

SONUÇ RAPORU ANA METNİ 1

1. GİRİŞ 1

1.1 Fiziksel Bir Sistem Olarak Canlı Organizmalar 1

1.2 Moleküler Biyolojinin Ana Öğretisi, Semiyotik, Doğal Diller ve Biyomoleküller 2

1.3 Bilgi Deposu Olarak DNA, Genom ve Proteom 4

1.4 Projenin Araştırma Sorusu: Genom ve proteom gibi biyomolekül kaynaklarında depo edilen bilgi miktarı modellenerek ölçülebilir ve yorumlanabilir mi? 5

2. PROJE ARAŞTIRMA KONUSUNUN LİTERATÜRDEKİ YERİ 6

3. GEREÇ VE YÖNTEM 8

3.1 Bilgi Teorisinde Genel Bir Mesajın ve Bir Mesaj Kaynağının İçerdiği Bilginin Ölçütü 9

3.2 Genom ve Proteom Kaynaklarının İçerdiği Bilginin Ölçütü Modelinin Elde Edilmesi 12

4. BULGULAR 15

5. TARTIŞMA/SONUÇ 19

KAYNAKLAR 20

TÜBİTAK

TABLO VE ŞEKİLLER LİSTESİ

Tablo 1.	Çalışmada kullanılan model organizmalar……....	**14**
Tablo 2.	Proteomik uzunluk dağılımlarının Menzerath-Altmann modeli için elde edilen en iyi uyum parametreleri…….	**15**
Tablo 3.	İncelenen model proteom setinin ölçülen ve model ile elde edilen entropi ve bilgi içeriği değerleri…….	**16**
Şekil 1.	İki büyük derlemede yer alan farklı kelimelerin sayısının kelime uzunluklarına göre dağılımı grafiği……...	**3**
Şekil 2.	Schizosaccharomyces pombe organizmasının (ökaryotik maya hücresi) ….	**4**
Şekil 3.	Bir gen dizininde intron ve ekson bölgelerinin temsili gösterimi …….....	**8**
Şekil 4.	Temsili bir S mesaj kaynağından yayınlanan mesajlar, $s_1 s_2 s_3 \ldots s_i \ldots s_N$ …..	**11**
Şekil 5.	İncelenen proteom setindeki proteinlerin uzunluklarının; **(a)** Ölçülen, **(b)** Model ile elde edilen olasılıklar için olasılık dağılımları…….	**15**
Şekil 6.	İncelenen proteom setindeki proteomların bilgi içeriklerinin protein sayısına göre değişiminin; ölçülen ve model ile simülasyonu yapılan değerleri…….	**17**
Şekil 7.	İncelenen proteom setindeki proteomların gelişigüzel protein uzunluğu simülasyonu ile elde edilen olasılık dağılımları……..	**17**

TÜBİTAK

ÖZET

Bir canlı organizmanın tüm kalıtımsal ve yaşamsal faaliyetlerinin kodları o organizmanın DNA'sındaki genlerde yer alır ve organizmanın gen envanteri o organizmanın genomu olarak bilinir. Genomdaki her gen, spesifik bir (ya da nadiren birkaç) biyolojik fonksiyondan sorumlu bilgi taşıyan moleküler bir mesaj olarak düşünülebilir. O halde bir organizmanın genomu o organizmada yürütülmesi gereken tüm metabolik faaliyetlere dair bilgileri içeren bir mesaj kaynağıdır. DNA' da kodlu genetik mesajlar biyolojik fonksiyonları direkt olarak yerine getiremezler. Önce bu moleküler mesajlar RNA molekülüne dönüştürülür, sonra her bir RNA molekülü bir protein molekülünün sentezinde kullanılır. Biyolojik faaliyetlerin yerine getirilmesinden direkt olarak sorumlu moleküller bu protein molekülleridir. Protein dizinlerinin monomer birimleri olan aminoasitlerin organizasyonu ile bu proteinlere karşı gelen gen dizinlerinin monomer birimleri olan nükleotidlerin organizasyonu arasında birebir eşleşme yoktur. Şu an için genetik kodun çözümlenmesinden başlayarak protein sentezlenmesi ile sona eren sürecin aşamaları sadece ana hatları ile bilinmekte, ancak bu sürecin tüm detayları bilinmemektedir. Bir organizmanın gen envanteri olan genoma benzer şekilde, o organizmanın protein envanteri proteom olarak isimlendirilir. DNA ve protein dizinlerinin içerdikleri moleküler bilginin nicel olarak elde edilmesi literatürde pek çok çalışmaya konu olmuştur. Ancak, birbirleri ile çelişen sonuçlara ilaveten moleküler dizinlerin bilgi içeriğini veren bilgi ölçütü modellemesi yaklaşımı üzerinde hala uzlaşma sağlanamamıştır. Hâlihazırdaki yaklaşımlarla çok sayıda protein ve gen dizini içeren proteom ve genom kaynaklarının içerdikleri işlem/veri yoğun bilgi miktarının elde edilmesi oldukça zordur. Bu nedenle literatürde proteomların ve genomların bilgi içeriğine dair herhangi bir araştırma ile karşılaşılmamıştır. Bu projede öncelikli amacımız, organizmaların biyomoleküler dizin kaynaklarının (proteomların ve genomların) içerdikleri bilgi miktarlarının nicel olarak elde edilmesine olanak sağlayan bir modelin tasarlanmasıdır.

Genel olarak bilgi, olayların/olguların elle tutulup gözle görülemeyen bir özelliğidir ve bilgiyi alan alıcının durumunda (halinde) değişiklik yaratmak için gereken şartlara sahiptir. Komünikasyon mühendisliğinde bilgi teorisinin matematiksel temeli 1948'de Claude Shannon tarafından atılmıştır. O zamandan günümüze bilgi teorisi, DNA ve proteinler gibi bilgi taşıyan sistemleri incelenmesinde de kullanılmaktadır. Bilgi teorisine göre, bir sistemin içerdiği yapısal bilgi, o sistemi oluşturan bileşenlerin yapısal organizasyonunda gizlidir. Semboller dizininden oluşan bir mesajın içerdiği bilginin miktarı, mesajda yer alan semboller arasındaki korelasyon/düzen ile ilişkilidir. Benzer şekilde bir mesaj kaynağının içerdiği bilginin miktarı, kaynakta yer alan mesajlar arasındaki korelasyon/düzen ile ilişkilidir. Sürdürmekte olduğumuz araştırma çalışmalarımız, sembollerden oluşmuş bilgi içeren dizinlerin

oluşturduğu kaynaklarda ortak bir yapısal düzen özelliğinin varlığı ortaya çıkarmış ve sonuçlar yayınlanmıştır. Tamamlanan bu projede proteom ve genom kaynaklarının içerdikleri bilginin modellenmesinde takip edilecek strateji, literatüre kazandırılmış olan bu ortak yapısal özelliğin modellenmesinde kullanılmıştır.

Bu projede, özgün bir yaklaşımla proteom ve genom kaynaklarının içerdikleri bilginin nicel ölçütünün teorik modellenmesi yapılmıştır. Model ile farklı üst âlemlerden seçilmiş model organizmaların proteom ve genomlarının bilgi içerikleri elde edilerek, sonuçlar fiziksel ve biyolojik bakış açılarıyla yorumlanmıştır. 10 proteomdan oluşan bir model proteom setinin sahip olduğu ölçülen, model ile tahmin edilen (dildeki uzunluk dağılım modeli) ve simüle edilen (gelişigüzel uzunluk dağılımı) bilgi içeriklerini analiz ettik ve karşılaştırdık. Proteinlerin ölçülen ve model ile tahmin edilen bilgi içeriklerinin mükemmel bir şekilde uyumlu oldukları gözlenmiştir. Proteomik protein başına ortalama bilgi ölçülen/modellenen durum için 8 ve simüle edilen durum için 7 bit olarak elde edilmiştir. Araştırma, biyolojik etkileşme mekanizmalarının temelde etkileşmeyi başlatan protein dizininin aminoasit düzeninden öteye protein dizinindeki aminoasitlerin sayısı ile daha ilişkili olabileceğini ortaya koyar. Bu çalışmada sunulan yöntem bir organizmada ya da bir grup organizmada gerçekleşen biyolojik süreçlerin karşılaştırılmasında pratik bir yaklaşım olmakla birlikte; birincil ve ikincil yapı gibi farklı hiyerarşik yapılarda proteomik ve genomik sayısız bilgi karakteristiği araştırmalarına olanak sağlama potansiyeline sahiptir.

Proje çıktılarının şu ana başlıklarla verilen yaygın etkilerinin olacağı öngörülmektedir: i) Proteom ve genomların bilgi ölçütünü veren özgün bir modelin literatüre kazandırılması, ii) Bu model ile literatürde cevap bekleyen bazı önemli soruların farklı bir açıdan ele alarak cevaplanması ve bu konudaki bazı boşlukların doldurulması, iii) Dolayısıyla canlı organizmalarda gerçekleşen moleküler bilgi depolama/iletimi süreçlerinin daha iyi anlaşılması, iv) Yeni birtakım araştırma sorularının bilimsel gündeme taşınması.

ABSTRACT

All the inherited and vital function codes of a living organism is found in that organism's genes which are encoded in its DNA, and the organism's gene inventory is known as that organism's genome. Each gene in the genome can be considered as an information carrying molecular message, which is responsible for one (or rarely more) specific biological function. Therefore, an organism's genome is a message source that contains all the information necessary for accomplishing all the metabolic activities. The genetic messages coded in DNA cannot directly achieve the metabolic activities. Firstly, these molecular messages are converted into RNA molecules, then, each RNA molecule is used to synthesize one protein molecule. The molecules that are directly responsible for carrying out biological activities are these protein molecules. There is no one-to-one correspondence between a protein's amino acid (monomer unit of protein sequences) organization, and the associated gene's nucleotide (monomer unit of gene sequences) organization. For the time being, the stages of the process, starting with genetic decoding and ending with protein synthesis, is known to some extent; yet, all the details of this process are not known. Similar to genome, the gene inventory of an organism, the protein inventory of that organism is called proteome. Acquiring quantitative information content of DNA and protein sequences is the subject of many studies in the literature. However, in addition to conflicting results, there is no well recognized information measure modelling approach for obtaining information content of molecular sequences. Using the current approaches, it is quite difficult to obtain the process/data-intensive information content of proteome and genome sources which are consisting number of protein and gene sequences. For that reason, we have not encountered any literature reports on the information content of proteomes and genomes. Our primary goal in this project is to design an information content model for acquiring quantitative information content of organism's biomolecular sequence sources (proteomes and genomes).

Information, in general, is an intangible property of things/events, and it fulfills the requirement for changing the state of the information receiver. The mathematical foundation for the information theory in communication engineering was developed by Claude Shannon in 1948. Since then the information theory has been utilized to investigate various information carrying systems including biomolecules such as DNA and proteins. According to the information theory, the structural information content of a system is concealed in the structural organization of the system's components. The amount of information contained in a message that is made out of a sequence of symbols is related to the correlation/order among the symbols within the message. Likewise, the amount of information contained in a message source that is made out of a number of messages is related to the correlation/order

among the messages within the source. Our ongoing research was revealed the existence of a common structural order property in message sources that is made out of information carrying message sequences, and the results were published. In the completed project, the pursued strategy in the modeling of the information content of proteome and genome sources was the use of this common structural property, which we introduced to the literature.

In this project, by using an innovative approach, a theoretical model for the quantitative measure of proteome and genome sources was derived. Information content of proteomes and genomes of the organisms belonging to different domains was obtained, and the results were interpreted from physical and biological point of views. We analyzed and compared the information content estimates of a representative proteome set of ten proteomes for measured, model-predicted (linguistic length distribution model) and simulated (random sequence length) cases. Excellent agreement was observed in the measured and model-predicted information contents of the proteomes. The overall average information per proteomic protein was obtained as 8 and 7 bits for the measured/model-predicted and the simulated proteomic collection data, respectively. The study reveals that the biological interaction mechanisms may primarily rely on the number of amino acids than the amino acid order of an interaction-initiating protein sequence. The approach presented here may serve as a practical tool for studying and comparing biological processes taking place in an organism or in a collection of organisms, and is anticipated to offer numerous promises for the exploration of proteomic and genomic information characteristics present in different structural hierarchies such as the secondary and tertiary structures.

The project outcomes are expected to have far reaching influence on the issues such as: i) An innovative model for the information content of proteomes and genomes will be introduced into the literature, ii) With this model, some important questions waiting to be answered in the literature will be responded from a different perspective, and some issues in this subject will be resolved, iii) Therefore, molecular information storage/transmission processes taking place in living organisms will be better understood, iv) Some new research problems will be brought to the attention of the scientific community.

TÜBİTAK

SONUÇ RAPORU ANA METNİ

1. GİRİŞ

Tamamlanan projenin konusu, proje yürütücüsünün proje başlamadan önceki yaklaşık altı senelik dönemde sürdürmekte olduğu teorik moleküler biyofizik araştırmalarının devamı niteliğindedir. Ön araştırmalarda bilgi taşıyan sembol dizinlerinin organizasyonel özellikleri hakkında oldukça önemli sonuçlar elde edilmiş olup bu sonuçlar yayınlanmıştır. Bu proje ile yürütücünün araştırmalarının ulaşmış olduğu seviye daha da ileri götürülerek; DNA ve protein gibi bilgi taşıyan moleküler sembol dizini kaynaklarının bilgi ölçütlerinin basit modellenmesi hedeflenmiştir. Elde edilen çıktıların yaşamın temel kavramları olan proteom ve genomların henüz bilinmeyen bazı özelliklerine ışık tutacak olması öngörülmüştür.

Proje konusu ve kapsamının; istatistik fizik, atom ve molekül fiziği, biyofizik, moleküler biyoloji, biyoinformatik ve gen mühendisliği gibi bilim dalları ile alakalı olması nedeniyle, bu proje gerek doğası ve gerekse çıktılarının yaygın etkisi açısından disiplinlerarası bir proje konumundadır. İzleyen başlıklar altında projenin konusu/kapsamı, yöntemi ve sonuçları literatüre de atıfta bulunarak detaylandırılarak sunulmaktadır.

1.1 Fiziksel Bir Sistem Olarak Canlı Organizmalar

Canlı sistemleri fiziksel bakış açısı ile inceleyen biyofizik bilim dalının bu sistemlerin işleyiş prensiplerinin anlaşılmasındaki rolü gün geçtikçe artarak sürmektedir. Bu gerçeklikten yola çıkarak projede hedeflenen sonuçlara ulaşmayı sağlayacak olan temel yaklaşım; canlı organizmaların fiziksel sistemler olarak ele alınarak incelenmesi ve bu sistemlerin yapısal organizasyon davranışları ile ilişkili bilgi miktarının (entropik bilgi) elde edilmesi olarak belirlenmiştir.

Fizik bilimi çerçevesinden bakıldığında; canlı sistemler, süregelen kompleks dinamik süreçler neticesinde kararlı ve yüksek derecede yapısal düzene sahip ve bu nedenle çevreleri ile termodinamik denge durumundan uzakta faaliyet gösteren yarı açık fiziksel sistemler olarak düşünülebilir. Canlı organizmalarda karşılaşılan bu karmaşık yapısal düzen; sistem, hücre ve hatta biyomoleküller gibi pek çok farklı boyutsal ölçekte kendisini gösterir. Sonuç olarak, maksimum entropiye sahip rassal (gelişigüzel) bir düzenden farklı olarak canlı organizmalar daha düşük entropiye sahiptirler. Çok daha derin anlamlar içermesine rağmen en temel tanımı “rassallığın ölçütü” olan entropi kavramı aynı zamanda bir sistemin içerdiği bilgi miktarı ile de ilişkilidir. Bu ilişki bilgi teorisinde sıklıkla kullanılan bilgi entropisi ile tanımlanır. Tarihsel olarak entropi kavramı ilk önce termodinamik ve istatistik mekanik çalışmalarında ileri sürülmüştür, ancak, içerdiği anlam nedeniyle kullanımı bu bilim alanlarıyla

sınırlı kalmamıştır. İncelenen fiziksel sistemin hal fonksiyonun önemli bir değişkeni olan entropi, bu özelliği nedeniyle modern bilgi teorisinin en önemli kavramlardan biri konumundadır (Shannon, 1948).

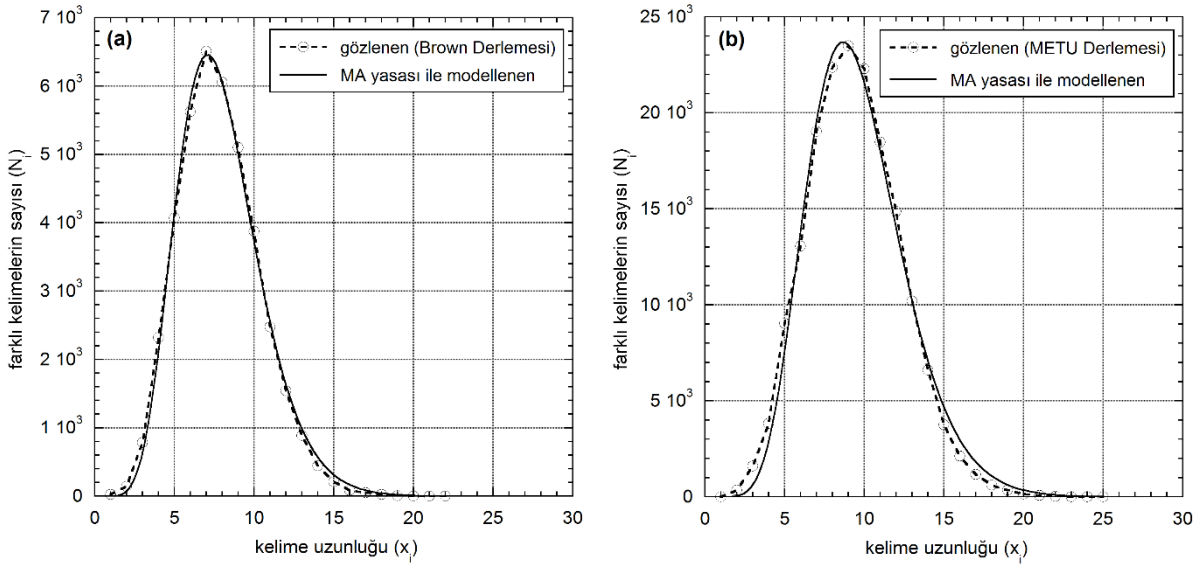
1.2 Moleküler Biyolojinin Ana Öğretisi, Semiyotik, Doğal Diller ve Biyomoleküller

Crick (1958) tarafından temelleri atılmış olan modern moleküler biyolojinin ana öğretisi; "DeoksiriboNükleik Asit (DNA) → RiboNükleik Asit (RNA) → protein → biyolojik fonksiyon" süreci ile özetlenebilir. Bu öğretiye göre, canlı bir organizmanın her türlü yaşamsal faaliyetini yerine getirmekle sorumlu olan proteinler (enzimler), o organizmanın genetik kodunun çözümlenmesi ile elde edilir (DNA, RNA'yı ve RNA' da proteini sentezler). Dolayısıyla bütün canlı organizmaların canlılıklarının devamı için gerekli olan her türlü fonksiyonun yerine getirilmesi, üç farklı sınıftaki oldukça büyük biyomolekül (DNA, RNA ve protein) ile sağlanır.
DNA, RNA ve protein molekülleri heteropolimerik yapıya sahip lineer makro moleküllerdir. Yani oldukça uzun dizin uzunluğuna sahip her bir molekül farklı "monomer" gruplarının bir araya gelmesi ile oluşur. DNA dört farklı monomer ya da nükleotid biriminden oluşurken (Adenin, A; Guanin, G; Sitozin, C; ve Timin, T); RNA ise, T'nin Urasil, U, ile yer değiştirdiği dört nükleotidden (A, G, C, U) oluşur. Proteinler ise en sıklıkla karşılaşılan 20 standart amino asit biriminden oluşan biyomoleküllerdir, Alberts vd. (2007).

Tamamlanan proje kapsamında, DNA, RNA ve protein moleküllerinin işlevleri ve yapısal özellikleri bir arada düşünüldüğünde; bu biyomoleküller spesifik bir fonksiyonu yerine getirmek üzere tasarlanmış ve dolayısıyla bu fonksiyonun yerine getirilmesine dair bilgiyi içeren birer simgeler dizini olarak ele alınmıştır. Semiyotik, simge bilimi, simgeler ya da işaretlerin yorumlanmasını ve üretilme süreçlerini inceleyen sosyal bir bilim dalıdır. İşaretleri harfler, notalar ya da görseller olan doğal diller, müzik eserleri ve trafik işaretleri gibi pek çok bilgi alış veriş mekanizması semiyotik biliminin ilgi alanına girer. Aslında doğada mevcut bütün bilgi alış veriş mekanizmaları kendine özgü simge sistemleri kullanır. Bu simge sistemleri harfler, notalar, şekiller ve renkler gibi görsel ya da duyusal olabilir. Ancak gündelik hayatta sıklıkla karşılaşılan bu bilgi transferi sistemlerinin dışında; DNA, RNA ve proteinler gibi biyomoleküllerin de birer bilgi depolama ve bilgi transferi sistemi olarak tanımlanması oldukça doğaldır (Smith 2000).

Özellikle insanlar arası yazılı ya da sözlü bilgi transferi sağlamak amacıyla kullanılan doğal diller ile canlı organizmalarda bilgi transferi sağlayan biyomoleküller arasındaki yapısal paralellik oldukça ilgi çekicidir. Bu paralellik literatürde proteinler üzerinde (Popov vd., 1996; Abe ve Mamitsuka, 1997; Przytycka vd., 2002; Ji ve Ciobanu, 2003; Gimona, 2006; Chiang vd., 2006; Motomura vd., 2012) ve DNA üzerinde (Brendel ve Busse, 1984; Mantegna vd., 1994; Mantegna vd., 1995; Searls, 1997; Ji, 1999; Ji, 2002) yapılan pek çok çalışmada ele

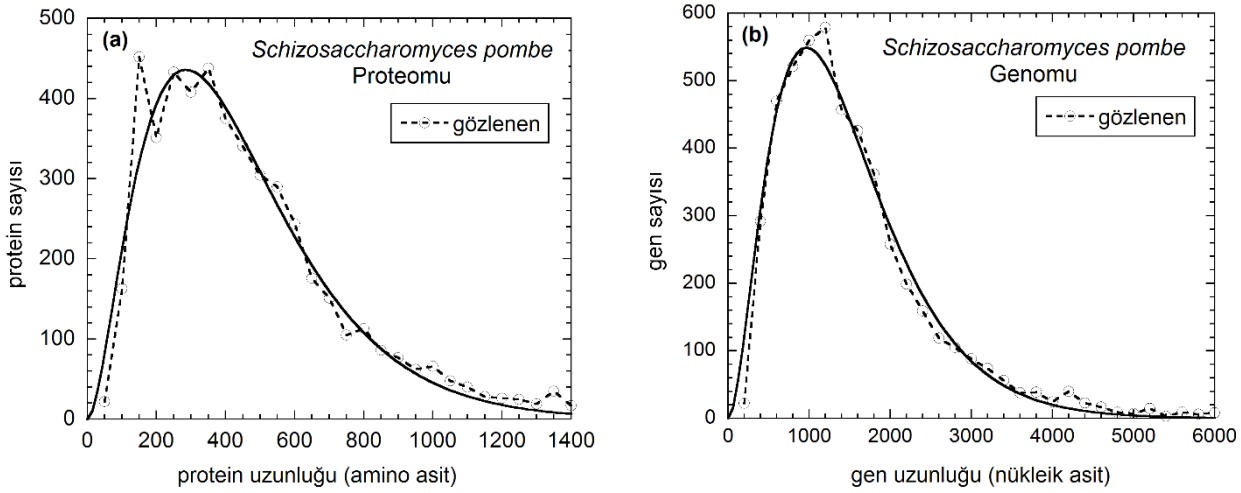
alınmıştır. Bu açıdan bakıldığında; DNA ve RNA dört harflik alfabeden, protein ise 20 harflik alfabeden oluşmuş, gramer kuralları henüz bilinmeyen birer doğal dil olarak düşünülebilir. Daha da ötesinde bu paralellikten yola çıkarak, kalıtımsal DNA dilinde yazılmış bir metnin önce RNA diline daha sonra protein diline tercüme edildiği ve en sonunda protein dilindeki bu metnin biyolojik fonksiyonların yerine getirilmesi üzerine tüm bilgileri (anlamları) içerdiği düşünülebilir.



Literatürde yer alan çalışmalarda, doğal diller ile biyomoleküller arasındaki paralelliğin sadece şekilsel düzeyde olmayıp, aynı zamanda organizasyonel davranış benzerliklerinin de olduğu ileri sürülmüştür. Bu kapsamda, proje yürütücüsü tarafından yapılan Eroglu (2013) çalışmasında insanlar arası iletişimi sağlayan doğal dillerin sahip oldukları ortak bir organizasyonel özelliğin biyomoleküllerde de var olduğu gösterilmiştir, Eroglu (2014b, 2014c). Bu sonuç, doğal diller ile biyomoleküller arasındaki organizasyonel paralellik fikrini

daha da güçlendirmektedir. Bilgi içeren sembol dizinlerinde karşılaşılan bu ortak yapısal özellik, projenin amaçlanan hedefe ulaşması için kullanılan yaklaşımın temelini oluşturmaktadır. Bahsi geçen sembol dizinlerinin ortak organizasyonel özelliği ve belirtilen referanslarda elde edilen sonuçların proje ile ilgili kısımları Şekil 1 ve 2' de özetlenmiştir.

Şekil 1. İki büyük derlemede yer alan farklı kelimelerin sayısının kelime uzunluklarına göre dağılımı grafiği. (a) İngilizce dilinde Brown Üniversitesi Derlemesi (Brown Derlemesi) ve (b) Türkçe dilinde Orta Doğu Teknik Üniversitesi Derlemesi (METU Derlemesi). Kesikli çizgi ile verilen eğriler gözlenen dağılımları gösterirken, sürekli çizgi ile verilen eğriler Menzerath-Altmann (MA) yasası ile modellenen dağılımları göstermektedir (Eroglu, 2013). Model tarafından kestirimi yapılan dağılım ile gözlenen dağılım arasındaki mükemmel uyum (R^2>0.99) oldukça ilgi çekicidir. Çalışmanın diğer bir önemli sonucu ise, aynı dağılım davranışının köken olarak birbirinden farklı iki ayrı dilde de gözlenmiş olmasıdır. Yani bu dağılım davranışı, doğal dillerde optimum bilgi transferi için karakteristik bir fiziksel mekanizmanın varlığını gösterir. Bu çalışmada kullanılan dağılım modeli, nicel dilbilimi çalışmalarında MA modeli olarak bilinen ve $N_i(x_i) = A\, x_i^b e^{-cx_i}$ şeklinde formülüze edilen dağılım fonksiyonudur. Fiziksel olarak MA dağılım modeli, güç yasasının eksponansiyel kesim (cut-off) ile kontrol edildiği dağılımdır ve iki-parametreli gamma dağılım fonksiyonu ailesindendir. Bu bağıntıda *A*, *b* ve *c* deneysel olarak belirlenen uyum (fit) parametreleridir.



Şekil 2. Schizosaccharomyces pombe organizmasının (ökaryotik maya hücresi). (a) Proteomundaki proteinlerin ve (b) Genomundaki genlerin (protein kodlayan dizinlerin), uzunluklarına göre dağılımı. Kesikli çizgiler ile verilen eğriler gözlenen ve sürekli çizgilerle verilen eğriler ise MA modeli ile kestirimi yapılan dağılımı göstermektedir (Eroglu, 2014b, 2014c). Gözlenen ve model ile tahmin edilen

dağılımlar arasındaki uyum dikkat çekicidir (R^2>0.97). Dolayısıyla MA dağılımı davranışı, doğal dillerde yazılmış metinlerde olduğu gibi biyomolekül kaynağı dizinlerinin organizasyonunda da karşılaşılan ortak bir yapısal özelliktir.

1.3 Bilgi Deposu Olarak DNA, Genom ve Proteom

Bir organizmanın DNA'sı; o organizmanın yaşamsal faaliyetlerini yerine getirmeye, sürdürmeye ve üremeye dair her bilginin (gen) tam bir setinin (genom) kodlanarak depolandığı bir biyomoleküldür. DNA'nın tamamı genlerden oluşmaz, genler arasındaki DNA dizinleri kodlamada kullanılmaz ve "faydasız DNA" (junk DNA) olarak isimlendirilir (Ohno, 1972). İnsan DNA'sının yaklaşık yüzde 3'ü kodlanabilen gen dizinlerinden oluşurken, geriye kalan yüzde 97'lik bölümü faydasız DNA dizinleridir. Adından da anlaşılabileceği gibi her ne kadar faydasız DNA dizinlerinin işlevsiz olduğu genel kabul görse de, bu dizinlerin bazı biyokimyasal süreçleri yönettiğine dair hipotezler de ileri sürülmektedir (Bieomont ve Vieira, 2006; Penisi, 2012).

Bir organizmanın genom setinde yer alan genler, DNA'nın transkripsiyonu süreciyle önce RNA moleküllerinin sentezlenmesinde ve sonrasında translasyon yolu ile o organizmanın ihtiyaç duyduğu proteinlerin tam bir setinin (proteom) sentezlenmesinde rol oynarlar. Her ne kadar protein dizinlerinin gen dizinleri ile birebir eşleştiği düşünülse de gerçekte durum bundan farklıdır. Özellikle gelişmiş organizmalarda gen dizinlerinin içerisinde protein sentezinde kullanılan dizinlere (ekson) ilaveten protein sentezinde kullanılmayan dizinlerle de (intron) sıklıkla karşılaşılır. Ancak, intron dizinlerinin gen regülasyonunda kullanıldığına dair bulgular gün geçtikçe artmaktadır (Rose, 2008).

Şu durumda bir organizmanın DNA molekülünün içerdiği bilginin tamamı genomik bilgi olmayıp, bilinmeyen miktardaki bir kısmı kodlamada kullanılmayan faydasız DNA ve intron dizinlerinde yer alır. Aynı zamanda o organizmanın genomunda depo edilen bilgi miktarı ile proteomunun içerdiği ve yaşamsal faaliyetlerini sürdürmek için ihtiyaç duyduğu bilgi miktarı arasında farklılık vardır.

1.4 Projenin Araştırma Sorusu: Genom ve proteom gibi biyomolekül kaynaklarında depo edilen bilgi miktarı modellenerek ölçülebilir ve yorumlanabilir mi?

Proje yürütücüsü proje başlangıcı öncesi yapmış olduğu ön araştırmalarda; canlı organizmaların genomlarındaki genleri ve proteomlarındaki proteinleri birer lineer dizin olarak ele alarak bu dizinlerin nükleotid ve amino asit sayılarına göre uzunluklarının dağılımını analiz etmiştir, Eroglu (2014b, 2014c). Bu çalışmalarda incelenen biyomoleküllerin uzunluklarının dağılımının gelişigüzel olmadığı, modellenebilir yapısal bir düzenin var olduğu ve ilginçtir ki bu yapısal organizasyonun Eroglu (2013) çalışmasında incelenen doğal dillerdeki farklı kelimelerin harf sayısı uzunluğuna göre dağılım davranışıyla büyük benzerliğe

sahip olduğu gösterilmiştir (Şekil 1 ve 2). Proje yürütücüsü Eroglu (2014a) çalışmasında; bilgi taşıyan sembolik dizinlerdeki bu ortak yapısal organizasyon davranışının doğal diller için fiziksel modellemesini yaparak, bu tür davranışın bilgi transferi esnasında alıcı ve vericinin "en az efor sarf etme (enerjinin minimizasyonu)" eğiliminden kaynaklandığı sonucuna ulaşmıştır.

Ön araştırmalar sonucunda elde edilen bulgular sembolik dizinlerin yapısal organizasyonu ve buna neden olan fiziksel mekanizmanın anlaşılmasında oldukça önemli sonuçlar sunarken, bu araştırma sonuçlarının genom ve proteomların henüz bilinmeyen daha pek çok özelliğinin anlaşılmasında önemli rol oynayabileceği yargısına ulaşılmıştır. Buna göre;

1. Proteom ve genomlardaki protein ve gen dizinlerinin yukarıda anlatılan araştırmalarda elde edilen karakteristik yapısal organizasyon özelliği kullanılarak (Eroglu (2014b, 2014c)); canlı organizmaların protein ve gen kaynaklarının sahip oldukları bilgi miktarının ölçütünü veren bir modelleme yapılabilir mi?

2. Farklı organizmaların protein ve gen kaynaklarının sahip oldukları bilginin nicel karşılaştırmasını yaparak sonuçları yorumlamak mümkün mü? Organizmanın kompleksliği ile bilgi miktarı arasında nasıl bir ilişki vardır?

3. Kodlamada kullanılmayan faydasız DNA dizinleri bilgi içerir mi? Eğer faydasız DNA bilgi içeriyorsa, içerdiği bilgi miktarı kodlamada kullanılan gen dizinlerinin içerdikleri bilgi ile karşılaştırılabilir mi? Bu inceleme, faydasız DNA dizinlerinin işlevinin olup olmadığına dair literatürdeki süregelen tartışmalara bir katkı sunabilir mi?

4. Moleküler mesajlarda yapısal bilgiye en fazla katkı moleküler dizinlerin içeriğinden mi yoksa uzunluğundan mı gelir?

soruları merak uyandıran ve proje ile cevaplandırılması hedeflenen temel sorular olmuştur.

Başlangıçta öngörüldüğü şekilde tamamlanan projede elde edilen sonuçlar, yeni araştırma sorularına ve bunların cevaplanmasına olanak sağlamıştır. Örneğin; Protein dizinlerinin birincil yapıları yani amino asit birimleri cinsinden uzunlukları göz önünde bulundurularak önerilen bu araştırma, proteinlerin ikincil ve üçüncül yapılarına göre bilgi içeriğinin elde edilmesine da uyarlanabilir mi?

2. PROJE ARAŞTIRMA KONUSUNUN LİTERATÜRDEKİ YERİ

İçeriğinin ne olduğundan bağımsız olarak, gen ve protein dizinlerinin içerdikleri bilgi miktarının elde edilmesi arzusu bu alanda çalışan pek çok bilim adamının araştırmalarında temel motivasyon olmaktadır. Bu konuda şimdiye kadar yapılan ve sonuçları literatürde yayınlanmış olan sayısız çalışmanın hemen hepsinde kullanılan ortak yaklaşım; bilgi

ölçütünün elde edilmesi istenen gen (ya da protein) dizinindeki nükleotid (ya da amino asit) birimlerini N birimden oluşan bloklara ayırarak (N-tuple) bu blokların dizin içerisindeki tekrarlama motiflerinin belirlenmesi ve bu motifler arasındaki korelasyonun analiz edilmesi şeklindedir (Herzel vd., 1994; Strait ve Dewey, 1996; Schmitt ve Herzel, 1997; Schneider, 1997; Grosse vd., 2000; Weiss vd., 2000; Sadovsky, 2006; Bose ve Chouhan, 2011; Machado vd., 2013). N=1 durumu, bir birim büyüklüğündeki blokları yani bireysel nükleotid ya da amino asit birimlerini ifade eder. Ancak literatürde kullanılan bu yöntem aşağıda sıralanan dezavantajlara sahiptir:

1. N-tuple analizi sadece eşit büyüklükteki (N birimden oluşan) blokların motiflerini inceler ve değişen N sayısı için birbiriyle uyuşmayan farklı sonuçlara ulaşılır,

2. N-tuple analizi işlem-yoğun bir analiz yöntemidir. Bu nedenle bu yöntem sınırlı N blok büyüklükleri için ve sınırlı sayıdaki dizinlere uygulanmaktadır (canlı organizmaların biyomolekül kaynaklarına, genom ve proteom, uygulanması oldukça güçtür),

3. N-tuple analizinin "sonlu büyüklükteki örnekleme etkisi (Basarin, 1959; Schmitt ve Herzel, 1997)" problemlerinden dolayı güvenilmez sonuçlar vermesi kaçınılmazdır,

Bu başlıca dezavantajlar nedeniyle literatürde sıklıkla birbiriyle çelişen sonuçlarla karşılaşılmaktadır. Örneğin; Strait ve Dewey (1996) gerçek protein dizinlerinde amino asit başına ortalama entropik bilgiyi 2.5 bit olarak elde etmiştir. Bu değer gelişigüzel amino asit dizilimine sahip kurgusal bir protein dizinindeki amino asit başına ortalama bilgi miktarının (bit/amino asit) yaklaşık % 58'i kadardır. Bu sonuç gerçek protein dizinlerindeki amino asitler arasında hatırı sayılır bir korelasyonun varlığını gösterir. Öte yandan, Weiss vd. (2000) gerçek protein dizinlerinde amino asit başına ortalama entropik bilginin (4.18 bit) gelişigüzel protein dizinlerindeki amino asit başına ortalama entropik bilginin yaklaşık % 97'sine eşit olduğunu göstermişlerdir. Bu sonuç ise gerçek protein dizinlerindeki amino asitlerin hemen hemen gelişigüzel bir dizilime sahip olduğunu ileri sürer. Bu iki çalışma arasında incelenen protein dizin setlerinin farklılığına ilaveten, Weiss vd. (2000) çalışmalarında sonlu büyüklükteki örnekleme etkisini minimize etmişlerdir. Her ne kadar Weiss vd. (2000)'nin elde ettikleri sonuç, protein dizinlerindeki hidrofobik ve hidrofilik grupların gelişigüzel olarak dağılmış olduğunu gösteren White ve Jacobs (1990)'ın çalışmasıyla kısmen desteklenmiş olsa da; literatürde protein dizinlerinin içermiş olduğu bilgi miktarına dair tartışma halen sürmektedir. Benzer belirsizlik DNA dizinleri üzerine yapılan çalışmalarda da mevcuttur.

Sonuç olarak, şu an itibarıyla literatürde genom ve proteom gibi biyomekül dizin kaynaklarının sahip oldukları bilginin ölçütünü veren kabul görmüş bir modelin eksikliği varlığını sürdürmektedir. Bu konuda atılacak her adım bilimsel farkındalığın artması açısından büyük öneme sahiptir.

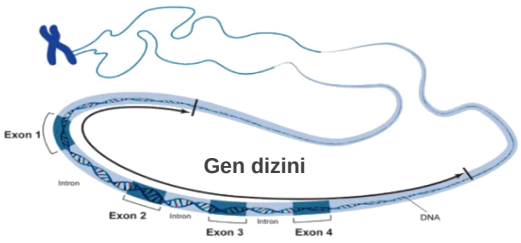
3. GEREÇ VE YÖNTEM

Proje önerisinin hazırlanmakta olduğu başlangıç aşamasında, proje konusu ile ilişkili ön literatür araştırması tamamlanmış olmakla birlikte, projenin kabul edilmesinden sonra detaylı literatür araştırmasına devam edilmiştir. Bu araştırmalar sonucunda, şimdiye kadar yapılan biyomoleküllerin bilgi içeriği çalışmalarının; seçilen gen ve protein dizinleri için bu dizinleri oluşturan bileşenlerin (nükleik asit ve amino asit) ya da bu bileşenlerin oluşturduğu farklı sayıdaki bileşen bloklarının dizin içerisindeki tekrarlama korelasyonları göz önünde bulundurularak yapıldığı belirlenmiştir. Bu projede yapılması önerilen çalışmanın ise literatürde yapılan çalışmalardan daha farklı bir yapısal organizasyonda (gen ve protein uzunluklarına göre) ve daha büyük bir ölçekte (genom ve proteom kaynakları için) bilgi içeriklerinin modellenerek bulunmasına olanak sağlayacağından; proje konusunun güncelliğini ve önemini koruduğu, literatür çalışmalarının proje önerisinde özetlenen mevcut çalışmalardan farklı olmadığı sonuçlarına varılmıştır.

Literatür araştırmalarıyla eş zamanlı olarak projenin konusu olan teorik çalışmalarımızda kullanılacak bilgisayar donanımı alımı gerçekleştirilmiş ve veri analizinde kullanılacak program bileşenlerinin yüklemesi tamamlanarak sistem kullanılır hale getirilmiştir.

Genom kaynaklarının içeriği olan genlerin ve proteom kaynaklarının içeriği olan proteinlerin projede önerilen yöntem ile analizi birebir benzerlik göstermemektedir. Bu durum, gen dizinlerinin protein dizinlerinden çok daha karmaşık kodlanmasından kaynaklanmaktadır. Protein dizinlerinin lineer yapısı (birincil yapısı) göz önünde bulundurulduğunda; protein dizinleri her birinin dizin içerisinde işlevsel olduğu kabullenilen (ikincil, üçüncül ve hatta dördüncül protein yapısının oluşumunda etkisi olan) amino asit birimlerinden oluşur.

Diğer yandan gen dizinlerinin hepsi için aynı şeyi söylemek mümkün değildir. Yani, gen dizinlerindeki her bir nükleik asit bileşeni (ya da bir grup nükleik asit dizini) o genin protein kodlamasında kullanılmaz. Bu durumla özellikle daha gelişmiş olan ökaryotik organizmalarda karşılaşılır. Gen içerisinde protein kodlanmasında kullanılan bölgeler ekson ve kodlamada kullanılmayan bölgeler ise intron olarak isimlendirilirler. İntron ve ekson bölgeleri Şekil 3'te bir organizmanın DNA'sında yer alan gen dizini içerisinde temsili olarak gösterilmektedir.



Şekil 3. Bir gen dizininde intron ve ekson bölgelerinin temsili gösterimi.

Bu durum, ökaryotik organizmaların genomlarının bilgi içeriklerinin doğru bir şekilde modellenmesi için yöntem üzerinde daha detaylı çalışılmasını gerektirmektedir. Bu çalışma projenin 2.nci döneminde sürdürülecektir. 1.nci dönemdeki başlangıç çalışmalarımız, önerilen yöntem ile nispeten daha uyumlu olan proteomların bilgi içeriklerinin modellenmesine odaklanmıştır.

3.1 Bilgi Teorisinde Genel Bir Mesajın ve Bir Mesaj Kaynağının İçerdiği Bilginin Ölçütü

Bu proje kapsamında öncelikle, proteom kaynaklarını oluşturan protein dizinlerinin organizasyonel özelliğinden dolayı sahip oldukları bilgi içeriğini veren en basit (temel) modelin oluşturulması çalışmaları tamamlanmıştır. Bu model ve sonuçları, yayına kabul edilen ve basım sırası bekleyen “Information Content Estimate of Model Proteomes: A Primary Structure Perspective” başlıklı bir yayında (Eroglu, 2017) yer almaktadır (Ek-1).

Bilgi kavramı her ne kadar soyut bir kavram olarak algılansa da, bir mesajın içerdiği bilgi miktarı ölçütünün elde edilmesi fikri 20. yüzyılın ilk çeyreğinden itibaren bilim adamlarının ilgisini çekmektedir. Temelde bilgi taşıyan bir mesajın sahip olduğu bilgi miktarı, mesajı oluşturan sembollerin istatistiksel dağılımın incelenmesi sonucunda bu semboller arasındaki korelasyon miktarı ile tanımlanır. İletişim teorisinde direkt olmasa da dolaylı şekilde bilgi kavramının ilk kez kullanımı Nyquist (1928) tarafından yapılmıştır. Nyquist bilginin ne olduğunu göz önünde bulundurmaksızın telgraf hatları ile iletilen sinyal formundaki karakterlerin nasıl en hızlı şekilde ancak bozulmaya uğramaksızın iletilebileceğinin teorisi üzerine yoğunlaşmıştır.

Sonrasında Hartley (1928) bilgi ölçütünü ifade eden bir bağıntıyı ileri süren ilk bilim adamı olmuştur. Hartley bir mesajın (mesaj terimi en genel anlamında bilgi ya da bilginin formu anlamını taşır) s tane mümkün sembol yardımıyla oluşturulabileceğini ileri sürmüştür. O halde, l sembol uzunluğuna sahip s^l tane birbirinden farklı mesaj yaratmak mümkündür. Hartley l sembol uzunluğuna sahip bir mesaj ile ilişkili bilginin ölçütünü

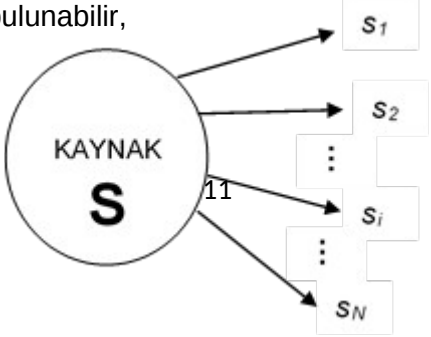
$$H_{Hartley}\left(s^{l}\right)=\ln\left(s^{l}\right) \tag{1}$$

şeklinde tanımlamıştır. Hartley’in bilgi ölçütü ifadesi (1) oldukça ilginç sonuçlar sunar. Öncelikle 1 sembol uzunluğunda bir mesajı düşünelim; bu mesajın içerdiği bilgi miktarı (1) bağıntısı kullanılarak $H_{Hartley}\left(s^{1}\right)=\ln\left(s^{1}\right)=\ln\left(s\right)$ şeklinde elde edilir. O halde, bu iki eşitlik şu şekilde birleştirilebilir:

$$H_{Hartley}(s^l) = \ln(s^l) = l \ln(s) = l\, H_{Hartley}(s^1) \quad (2)$$

Diğer bir deyişle, *l* tane sembol uzunluğuna sahip bir mesajın içerdiği bilgi 1 sembol uzunluğundaki bir mesajın içerdiği bilginin *l* katıdır. Bilginin bu tanımı sezgisel olarak oldukça tatmin edicidir. Hartley'in bilgi ölçütünü tanımlamada doğal logaritma fonksiyonunu seçmesindeki neden bu fiziksel sezgiye dayalı beklentiyi karşılayan tek fonksiyonun logaritma fonksiyonu olmasıdır. Hartley'in tanımına göre bir mesajın içerdiği bilginin birimi "*nat*" dır. Ancak *e* tabanı yerine 2 tabanına göre logaritma alınması bilgi ölçütünün sayısal değerinin anlamlandırılmasında kolaylık sağlaması nedeniyle 2 tabanına göre logaritma sıklıkla kullanılır ($\log_b p = \log_b a . \log_a p$ özelliği hatırlandığında taban seçiminin önemli olmadığı sadece seçilen taban için ilgili ölçek çarpanının kullanılması gerektiği açıktır). Bu durumda bilginin birimi "*bit*" cinsinden elde edilir. Sonuç olarak, Hartley'in bilgi ölçütü ifadesi (1) iki kabullenmeye dayanır: 1) Bilgiyi oluşturan her bir sembol mümkün olan q olası değerden herhangi birisini alabilir ve bu q olası değerin her biri eş olasılıkla gerçekleşebilir, 2) Mesajı oluşturan bütün semboller birbirinden bağımsızdır.

Modern bilgi teorisinin kurucusu olarak kabul edilen Shannon, Hartley'in bilgi ölçütü tanımlamasında yapılan iki kabullenmenin pek çok gerçek durumu yansıtmadığı ve bunların problematik olduğu sonucuna varmıştır. Bu problemlerden ilki, mesajların hepsinin eş olasılığa sahip olduğu kabullenmesinin yapılmış olmasıdır. İkinci problem ise, semboller arasındaki ilişkinin bilinmemesidir. Shannon bu problemlerin üstesinden gelecek şekilde, bilgi ölçütünü Nyquist ve Hartley'in tanımlamalarının ötesine taşımıştır. Bu amaçla Shannon öncelikle, $s_1 s_2 s_3 \ldots s_l$ sembolleri ile verilen ve *l* sembol uzunluğuna sahip bir mesajı oluşturan sembollerin her birinin, s_i, olası değerlerinin eşit olasılıklara sahip olması gerekmediği kabullenmesini yapmıştır. İkinci olarak Shannon, mesajı oluşturan sembollerin birbirinden bağımsız olmadığı birbirleri ile ilişkili olduğu kabullenmesini yapmıştır. Yani, s_2 sembolünün mesajda yer alması s_1 sembolünün mesajda yer almasına bağlı olabilir. Benzer şekilde s_3 'ün mesajda yer alması kendinden önceki s_1 ve s_2 sembollerinin mesajda yer almaları ile ilişkili olabilir. Bu kabullenmeler göz önünde bulundurulduğunda $s_1 s_2 s_3 \ldots s_l$ şeklinde gösterilen semboller dizini ile verilen bir mesajın içerdiği entropik bilgi miktarı Shannon'ın kurduğu bilgi teorisi ile bulunabilir, Shannon (1948).



Şekil 4. Temsili bir S mesaj kaynağından yayınlanan mesajlar, $s_1 s_2 s_3 \ldots s_i \ldots s_N$.

Şekil 4'de verilen bir kaynaktan yayınlanan *N* tane mesajdan *i* nci mesaj olan s_i ' nin gerçekleşme olasılığı p_i olsun; ' nin gerçekleşmesinin içerdiği bilginin ölçütü "*bilgi fonksiyonu*" $I(s_i)$ ile verilir:

$$I(s_i) = \log\left(\frac{1}{p_i}\right) \quad \text{ve} \quad 0 \le p_i \le 1. \tag{3}$$

Bu ifade, bilgi içeriği hakkındaki sezgisel beklentileri yerine getirir: Eğer $p_i = 1$ ise $I(s_i) \to 0$ ve $p_i = 0$ ise $I(s_i) \to \infty$ olur. Diğer bir deyişle, bir olayın gerçekleşme olasılığı p_i büyük ise içerdiği bilgi daha az, p_i olasılığı küçük ise içerdiği bilgi daha fazladır. En genelinde, $I(s_i)$ rassal bir değişken S'nin olası değerler seti içerisinden, $\{s_k\}$, bir değeri, s_i , alması olayının gerçekleşmesinin içerdiği bilgi miktarını (öz-bilgi, "*self-information*") tanımlar.

S kaynağından $s_1 s_2 s_3 \ldots s_i \ldots s_k$ şeklinde *k* tane mesaj yayınlandığı ve bu mesajların hangi sırada yayınlandığı bilinmiyorsa: S kaynağının alıcıya ilettiği ortalama öz-bilgi basitçe:

$$H(S) \equiv H(p_i) = \sum_i p_i I(s_i) = \sum_{i=1}^{k} p_i \ln\left(\frac{1}{p_i}\right) = -\sum_{i=1}^{k} p_i \ln(p_i) \tag{4}$$

şeklinde tanımlanır. $H(S)$ ifadesi S kaynağından yayınlanan mesajlar dizininin ortalama bilgi miktarı ya da literatürdeki tanımıyla kaynağın entropisi (Shannon Entropisi) olarak isimlendirilir. Bu basit incelemede kaynağın yayınladığı mesajların birbirinden bağımsız olduğu kabullenmesi yapılmıştır. Eğer semboller birbirine bağımlı ise o durumda bu ifadenin yeniden düzenlenmesi gerekir. Bu aşamada şu özelliği de vurgulamak oldukça önemlidir: Bilginin tanımında logaritma fonksiyonu kullanılmasının diğer bir fiziksel nedeni ise logaritma fonksiyonunun önemli bir özelliğinden kaynaklanır; çarpım terimlerinin logaritması terimlerin

toplamına eşittir. Dolayısıyla iki alt olayı içeren bir olayın gerçekleşmesi sonucu elde edilen bilgi, alt olayların her birinin gerçekleşmesi ile elde edilen bilgilerin toplamı olur. Bu durum ise fiziksel beklentimizle uyum içerisinde bir özelliktir. Sonuç olarak N tane mümkün mesaj içeren bir kaynağın sahip olduğu toplam öz-bilgi miktarı

$$H(p_i) = -N\sum_{i=1}^{k} p_i \ln(p_i) \tag{5}$$

ile verilir. Bilgi, doğası gereği soyut bir kavramdır. Dolayısıyla öz-bilginin sayısal değeri tek başına bir anlam ifade etmez. Bu durum fizikte potansiyel enerji kavramının tanımı ile benzerlik gösterir. Nasıl ki bir cismin konumundan dolayı sahip olduğu potansiyel enerji değeri ancak kesin bir referansa göre tanımlanması durumunda anlam ifade ediyorsa, öz-bilgi değeri de bir referansa göre tanımlanması durumunda anlam ifade eder. Yani, bir mesajın ya da mesaj kaynağının içerdiği bilgi miktarı o mesajın ya da mesaj kaynağının öz-bilgisinin kesin bir referansa göreceli olarak tanımlanması ile elde edilir. Bilgi içeriği için değişmez referans ise; bir mesajın ya da mesaj kaynağının maksimum entropisinin sahip olduğu öz-bilgi miktarıdır. *N* tane mesajın yayınlandığı bir kaynakta maksimum entropi (maksimum düzensizlik ya da maksimum gelişi güzellik) her bir mesajın yayınlanma olasılığının aynı olduğu durumda elde edilir, $H_{\max}(p) = N\log_2 N$.

3.2 Genom ve Proteom Kaynaklarının İçerdiği Bilginin Ölçütü Modelinin Elde Edilmesi

Bölüm 3.1'de özetlenen bilgi teorisi çalışmaları sonucunda; bir proteom kaynağını oluşturan *N* tane protein dizininin birincil yapıları yani amino asit sayısı cinsinden uzunluklarının dağılımı göz önünde bulundurularak sahip oldukları yapısal bilgi içeriğini veren en temel ifadenin,

$$I = N\log_2 N - \left(-N\sum_{i=1}^{n} p_i \log_2 p_i \right) \tag{6}$$

şeklinde verilebileceği açıktır. Burada $H_{\text{actual}}(p) = -N\sum_{i=1}^{n} p_i \log_2 p_i$ şeklinde ifade edilen terim; kaynağın mesajlarının yani proteinlerin uzunluklarının belirli bir dağılıma uymasından dolayı sahip olduğu öz bilgi içeriği iken, $H_{\max}(p) = N\log_2 N$ terimi ise kaynaktaki proteinlerin her birinin olası her uzunluk değerini eşit olasılıkla alabildiği duruma karşı gelen maksimum bilgi miktarına karşı gelir. Proje önerisinin yöntemler kısmında bahsedildiği gibi bilginin sayısal değeri belirli bir referansa göre tanımlanması durumunda anlam ifade eder. Proje çalışmaları ve detaylı literatür taraması sonucunda bilgi içeriğinin sayısal değerinin tanımlandığı

referansının $H_{\max}(p)$ ile ifade edilen maksimum bilgi içeriği olarak alınmasının önemli olduğu sonucuna varıldı.

Eşitlikteki p_i olasılığı, bir proteomun toplam N tane proteini içerisinden seçilen bir proteinin i nci olası uzunluk durumunda bulunma olasılığıdır. i nci olası uzunluk durumunda n_i tane protein varsa $p_i = n_i / N$ şeklinde elde edilir.

Bölüm 1.4'te açıklandığı şekliyle, proteomlardaki proteinlerin uzunluklarının

$$n_i(I_i) = A I_i^b e^{-cI_i} \tag{7}$$

ifadesi ile verilen dağılım fonksiyonuna uydukları, proje yürütücüsünün önceki çalışmalarıyla ulaşmış olduğu önemli bir sonuç olup önerilen projenin dayanak noktasıdır. Bu ifadede I_i, i nci uzunluk durumunu temsil ederken, $n_i(I_i)$ ise i nci uzunluk durumundaki proteinlerin sayısıdır. A, b ve c ise deneysel olarak belirlenecek olan parametrelerdir. Benzer şekilde, bir proteom içinden seçilen bir proteinin i nci uzunluk durumunda olma olasılığı $p_i = n_i(I_i)/N$ dir. O halde, proteomların bilgi içeriklerini eşitlik (6) kullanılarak ölçüm ile elde etmek mümkün olabildiği gibi, eşitlik (7) kullanılarak olasılıkların simülasyon modeli yoluyla elde etmek de mümkündür. Böylece, bilgi içeriği simülasyon modeli,

$$I = N \log_2 N - \left[- N \sum_{i=1}^{n} K I_i^b e^{-cI_i} \log_2 \left(K I_i^b e^{-cI_i} \right) \right] \tag{8}$$

ile verilir. Burada K parametresinin eşitlik (7)'deki A parametresi ile ilişkisi; $A=NK$ olacak şekilde, K bir ölçekleme parametresidir. Bu çalışmada proteomlarının bilgi içerikleri elde edilmiş olan model organizmalar Tablo 1'de verilmiştir.

Günümüzde organizmaların proteomlarında yer alan protein dizinleri, karmaşık karşılaştırmalı algoritmaları kullanan biyoinformatik teknikleri ile elde edilmektedir. Pek çok organizmanın sahip olduğu protein dizinleri halen tamamlanmamış olup bu yöndeki biyoinformatik araştırmaları sürdürülmektedir. Bu nedenle organizmaların veri bankalarındaki proteom ve genom bilgileri sürekli güncellenmektedir. Ancak proteomik ve genomik çalışmalarda sıklıkla kullanılan ve göreceli olarak proteom ve genom bilgileri tamamlanmış olan sınırlı sayıda organizma mevcuttur. Bu nedenle bu organizmalar literatürde ve proje yürütücüsün çalışmalarında sıklıkla "**model organizmalar**" olarak isimlendirilmektedir. Proje yürütücüsün önceki araştırmalarında ve yürütülmekte olan bu projede sürdürdüğü araştırmada aynı model organizmaların kullanılmasının nedeni ise; ileri sürülen yöntem ve

modellemelerden elde edilen sonuçların herhangi bir spekülasyondan uzak, mümkün olduğunca güvenilir olması arzusundan kaynaklanmaktadır.

Tablo 1. Çalışmada kullanılan model organizmalar.

Organizma	Kısaltması	Proteom büyüklüğü (protein sayısı)
ARKE		
Methanococcus jannaschii	Mja	1,771
Methanobacterium thermoautotrophicum	Mth	1,873
Archaeoglobus fulgidus	Afu	2,420
BAKTERİ		
Haemophilus influenza	Hin	1,610
Escherichia coli (K12)	Eco	4,141
Bacillus subtilis (str. 168)	Bsu	4,175
ÖKARYOT		
Schizosaccharomyces pombe	Spo	5,133
Saccharomyces cerevisiae	Sce	5,906
Caenorhabditis elegans	Cel	26,047
Mus musculus	Mmu	28,272

Bilgi ölçütü modelini uygulamak üzere seçilmiş olan model organizmalar bütün canlıları kapsayacak şekilde; Ökaryot, Bakteri ve Arke üst âlemlerinden seçilmiştir. Her bir üst âlemden en az üç farklı organizmanın genom ve proteom verilerinin incelenmesi amaçlanmıştır (Arke: 3 organizma, Bakteri: 3 organizma, Ökarya: 4 organizma). Bu sayede, örnek sayısı arttırılırken tüm canlı organizmaları kapsayan taksonomik sınıflandırmalar arasında karşılaştırma yaparak sonuçları yorumlamanın mümkün olması hedeflenmiştir. Bilgi ölçütü modelinin uygulanmasında, Nisan 2014 tarihinde “National Center for Biotechnology Information (NCBI)” organizasyonunun internet sayfasından (ftp://ftp.ncbi.nlm.nih.gov/genomes/) elde edilmiş olan proteom verileri kullanılmıştır.

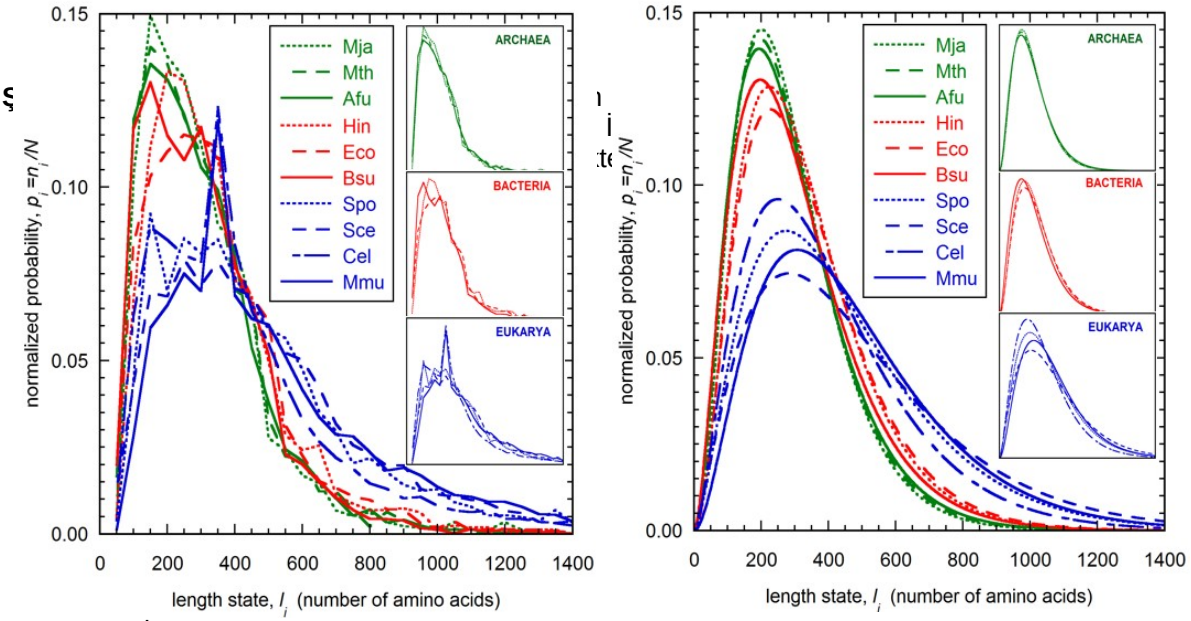
4. BULGULAR

Oluşturulan basit algoritma ile, eşitlik (8) kullanılarak proteomların bilgi içeriklerinin model simülasyonu yoluyla elde edilmesi için kullanılacak parametreler incelenen organizmalar için belirlenerek aşağıdaki tabloda verilmiştir. R^2 parametresinin yüksek değeri ($R^2 \geq 0.95$), ölçülen ve model ile simülasyonu yapılan proteomik proteinlerin uzunluklarının dağılımının oldukça uyumlu olduğunu göstermektedir.

Tablo 2. Proteomik uzunluk dağılımlarının Menzerath-Altmann modeli için elde edilen en iyi uyum parametreleri.

ORGANİZMA	MODEL PARAMETRELERİ			R^2
	K ($\times 10^{-5}$)	b ($\times 10^{0}$)	c ($\times 10^{-3}$)	
ARKE				
Mja	0.6 (0.3)	2.3679 (0.1195)	11.8 (0.5)	0.98
Mth	1.4 (0.7)	2.1535 (0.1177)	11.0 (0.5)	0.98
Afu	2.7 (1.1)	2.0024 (0.0953)	10.3 (0.4)	0.98
BAKTERİ				
Hin	0.8 (0.3)	2.1981 (0.0837)	9.8 (0.3)	0.99
Eco	1.1 (0.4)	2.1027 (0.0821)	9.3 (0.3)	0.99
Bsu	5.6 (2.7)	1.8094 (0.1066)	9.2 (0.5)	0.98
ÖKARYOT				
Spo	4.1 (1.9)	1.6581 (0.0950)	6.1 (0.3)	0.98
Sce	7.0 (1.7)	1.5119 (0.0511)	5.3 (0.1)	0.99
Cel	4.2 (2.6)	1.7092 (0.1331)	6.8 (0.4)	0.96
Mmu	0.8 (0.6)	1.9428 (0.1580)	6.3 (0.4)	0.95

Proteomların, ölçülen (Eşitlik (6)) ve model ile elde edilen (Eşitlik (8)) uzunluk dağılımı olasılıkları Şekil 5'te verilmiştir. Bu çalışmada uzunluk durumları için binleme yöntemi kullanılarak 50 amino asitten oluşan bin grupları oluşturulmuştur. Örneğin, i=1 nci durum 1-50 amino asit uzunluğu aralığı durumuna, benzer şekilde i=2 nci durum 51-100 amino asit uzunluğu aralığı durumuna karşı gelmektedir. Binleme işlemi 1400 amino asit maksimum protein uzunluğunda sonlandırılmıştır. Proteomların ölçülen (Eşitlik (6)) ve model ile elde edilen (Eşitlik (8)) bilgi içerikleri verileri Tablo 3'te özetlenmektedir (model ile elde edilen verilerde önceki Tablo 2'de yer alan model parametreleri kullanılmıştır).

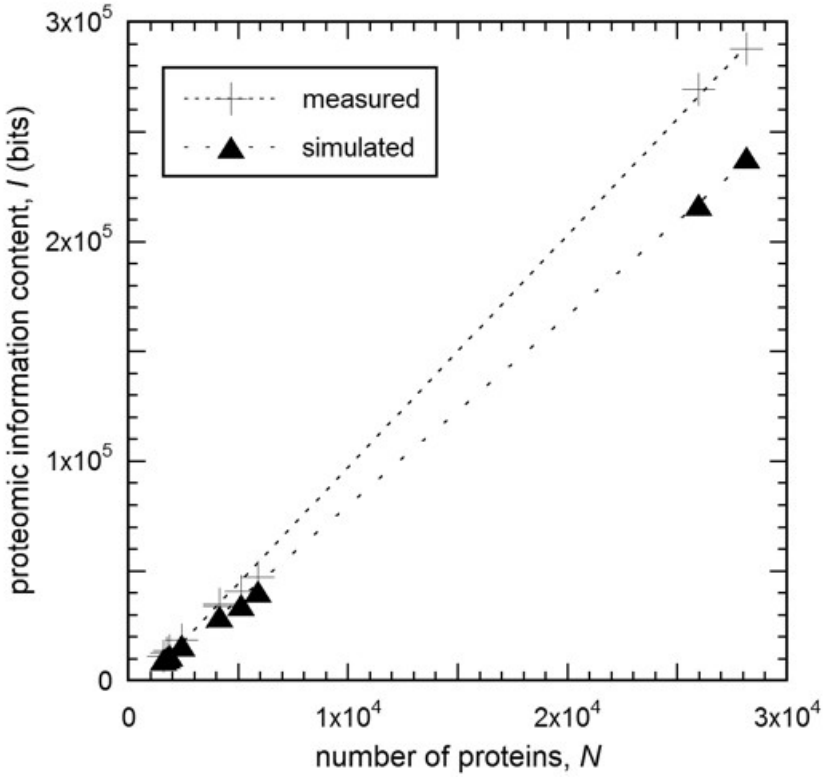


Tablo 3. İncelenen model proteom setinin ölçülen ve model ile elde edilen entropi ve bilgi içeriği değerleri.

ORGANİZMA	ÖLÇÜLEN # protein (N)	ÖLÇÜLEN H_{max} (bits)	ÖLÇÜLEN H_{actual} (bits)	ÖLÇÜLEN I (bits)	ÖLÇÜLEN I_{pp} (bits)	ÖLÇÜLEN $\frac{H_{actual}}{H_{max}}$	MODEL İLE ELDE EDİLEN # protein (N)	MODEL İLE ELDE EDİLEN H_{max} (bits)	MODEL İLE ELDE EDİLEN H_{actual} (bits)	MODEL İLE ELDE EDİLEN I (bits)	MODEL İLE ELDE EDİLEN $\frac{H_{actual}}{H_{max}}$
ARKE											
Mja	1,771	19,110	6,340	12,770	7	0.33	1,743	18,768	6,097	12,671	0.32
Mth	1,873	20,362	6,757	13,605	7	0.33	1,859	20,189	6,608	13,581	0.33
Afu	2,420	27,203	8,766	18,437	8	0.32	2,417	27,165	8,727	18,438	0.32
BAKTERİ											
Hin	1,610	17,151	6,035	11,116	7	0.35	1,629	17,381	6,153	11,228	0.35
Eco	4,141	49,757	15,847	33,910	8	0.32	4,128	49,582	15,676	33,906	0.32
Bsu	4,175	50,215	15,418	34,797	8	0.31	4,196	50,498	15,668	34,830	0.31
ÖKARYOT											
Spo	5,133	63,267	22,660	40,607	8	0.36	4,993	61,342	20,951	40,391	0.34
Sce	5,906	73,990	26,860	47,130	8	0.36	5,763	71,995	25,066	46,929	0.35
Cel	26,047	382,079	112,657	269,422	10	0.29	25,153	367,698	101,198	266,500	0.28
Mmu	28,272	418,061	130,216	287,845	10	0.31	26,560	390,351	109,819	280,532	0.28

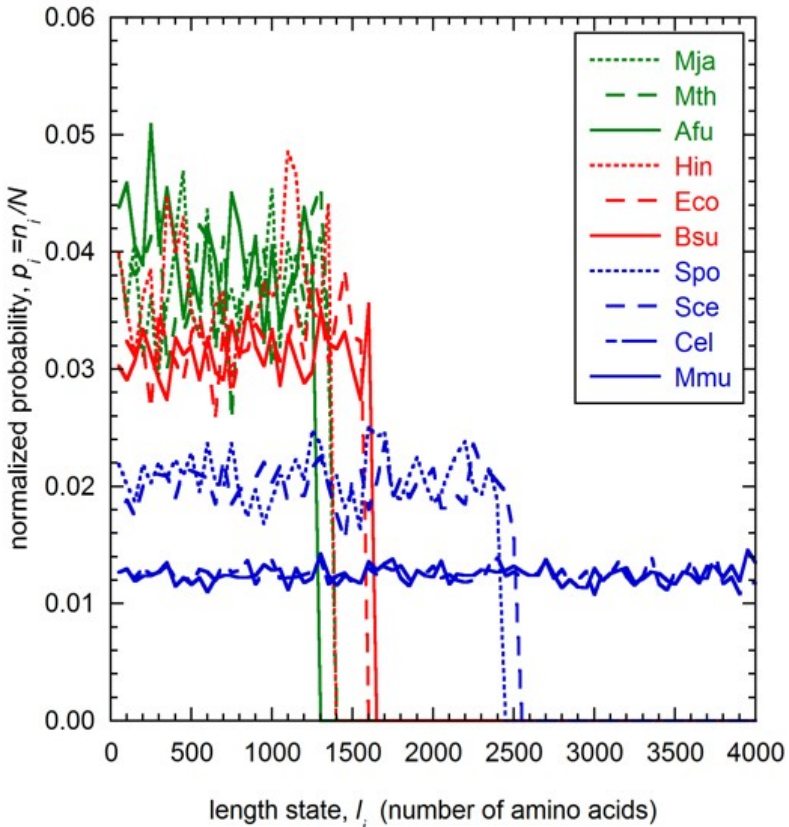
Her iki hesaplamada da proteomların bilgi içerikleri *I*'nın artan protein sayısı ile arttığı yani bilgi içeriğinin organizmanın kompleksliği ile orantılı olduğu (Şekil 6) ve proteomlardaki protein başına bilgi içeriği I_{pp}'nın ($I_{pp}=I/N$) 7-10 bit/protein aralığında olduğu gözlenmiştir (ortalaması 8 bit/protein). H_{actual}/H_{max} oranı organizmanın bulunduğu üst âlemden bağımsız olarak 0.29-0.36 aralığında değişir, ortalama değeri 0.33'dür. Bu sonuç, incelenen organizmaların gerçek entropilerinin maksimum entropilerinin yaklaşık 1/3'ine eşit olduğunu gösterir. Maksimum entropiden farklılık, protein uzunluğu durumlarının dejenereliği ile ilişkilidir. n_i *i* nci uzunluk durumunun girilebilir durumlarının sayısı yani n_i *i* nci uzunluk durumunun dejenerelik düzeyidir.

Caenorhabditis elegans ve *Mus musculus* gibi yüksek organizmaların olasılık dağılımlarında, 300-400 amino asit uzunluğu bölgesinde gözlenen ani şiddetli artışları oldukça ilginçtir ve gelecekte bu davranışın bilgi teorisi yöntemleriyle incelenmesi planlanmaktadır.



Şekil 6. İncelenen proteom setindeki proteomların bilgi içeriklerinin protein sayısına göre değişiminin; ölçülen ve model ile simülasyonu yapılan değerleri. Model ile elde edilen bilgi içeriği değerlerinin ölçülen bilgi içeriği değerlerinden ortalama %16 daha küçük olduğu gözlenmiştir.

Son olarak gereken durum; aynı model araştırılması incelenen proteinlerinin aynı yöntem ve uzunluklara kullanılarak olduklarını proteomların olasılık gelişigüzel davranışını sahip Bu sonuç Şekil simüle ederek 7'te verilmiştir. dağılımlarının incelemektir.



Şekil 7. İncelenen proteom setindeki proteomların gelişigüzel protein uzunluğu simülasyonu ile elde edilen olasılık dağılımları.

Proje kapsamında yürütülmüş olan proteomlar üzerine araştırmalar ışığında elde edilen sonuçlar; bizi bilimsel inanış ve beklentilerimizin dışında olması nedeniyle oldukça spekülatif sayılabilecek şu hipoteze yönlendirmiştir: "**Proteomlardaki proteinlerin birincil yapılarından (yani dizin uzunluklarından) dolayı sahip oldukları yapısal bilgi içerikleri protein dizinini oluşturan amino asit düzenlenişinden dolayı sahip oldukları bilgi içeriğinden daha kayda değerdir**". Aslında bu hipotezin keskin önermesi ise şu şekildedir: "**Bir protein dizininin hedeflenen bir metabolik fonksiyonu yerine getirmesinde dizin uzunluğu, dizinin amino asit düzenleniminden daha önemli bir faktördür!**". Kuşkusuz bizi bu hipoteze yönlendiren, literatürde karşılaşılan ve hipotezimizi destekleyen bazı sonuçlardır. Bu çalışmalardan birinde (Weiss vd., 2000) elde edilen sonuca göre; gerçek protein dizinlerindeki amino asit başına ortalama bilgi içeriği 4,18 bit olup bu değer gelişi güzel dizilime sahip bir proteinin amino asit başına ortalama bilgi içeriği olan 4,32 bit değerinin %97 sine tekabül eder. Bu sonuca göre gerçek proteinlerde amino asit dizilimleri arasında oldukça küçük bir korelasyon olup dizilim neredeyse gelişigüzeldir. Diğer bir çalışmada (White and Jacobs, 1990) elde edilen sonuç ise; "protein dizinlerinin üç boyutlu konformasyonunu sağlayan dizin içerisindeki hidrofobik (suyu sevmeyen) rezidü grupları protein dizinleri boyunca gelişigüzel bir şekilde dağılım gösterdiğini" yani proteinlerdeki amino

asit dizilimindeki gelişi güzelliği ileri sürmektedir. Sonuç olarak, literatürde yer alan bu çalışmalar, proje araştırmasında elde edilen sonuçlardan yola çıkarak ileri sürülen “proteomlardaki proteinlerin birincil yapılarından (yani dizin uzunluklarından) dolayı sahip oldukları yapısal bilgi içerikleri protein dizinini oluşturan amino asit düzenlenişinden dolayı sahip oldukları bilgi içeriğinden daha kayda değerdir” hipotezini destekler niteliktedir.

5. TARTIŞMA/SONUÇ

Şimdiye kadar anlatılan ve projeden elde edilen sonuçlar yayına kabul edilen ve basım sırası bekleyen “Information Content Estimate of Model Proteomes: A Primary Structure Perspective” başlıklı bir yayında (Eroglu, 2017) yer almaktadır (Ek-1).

Proteomların bilgi içeriği üzerine yapılan bu çalışmalara ilaveten genomlardaki genlerin uzunluklarının basit (ekson ve intronlar göz önüne alınmaksızın) entropisi çalışmaları yapılmış ve benzer sonuçlara ulaşılmıştır. Elde edilen sonuçlar bir poster bildirisi halinde sunulmuştur (Ek-2 ve Ek-3). Bu çalışmada, organizmaların kompleksliği ile genlerin uzunluk entropisinde artış olduğu sonucuna varılmıştır. Projenin tamamlanmasını takip eden süreçte entropi çalışmalarına ve daha karmaşık bir düzenlenim içeren genomların bilgi içeriği çalışmalarına odaklanılacaktır.

Projemizin en önemli sonucu; biyolojik etkileşme mekanizmalarının temelde etkileşmeyi başlatan protein dizininin aminoasit düzeninden öteye protein dizinindeki aminoasitlerin sayısı ile daha ilişkili olabileceğini ortaya koymasıdır. Bu çalışmada sunulan yöntem bir organizmada ya da bir grup organizmada gerçekleşen biyolojik süreçlerin karşılaştırılmasında pratik bir yaklaşım olmakla birlikte; birincil ve ikincil yapı gibi farklı hiyerarşik yapılarda proteomik ve genomik sayısız bilgi karakteristiği araştırmalarına olanak sağlama potansiyeline sahiptir.

Takip eden süreçte projede ulaşılan, “Bir protein dizininin hedeflenen bir metabolik fonksiyonu yerine getirmesinde dizin uzunluğu, dizinin amino asit düzenleniminden daha önemli bir faktördür” sonucunu destekleyen diğer araştırma metotları incelenerek bunların projedeki metot ile karşılaştırılmaları yapılarak yorumlanacaktır.

Proje sonrasında veri bankasından model genom ve proteom seti içeriklerinin arttırılması ve güncellemelerinin yapılarak çalışmaların sürdürülmesi planlanmaktadır. Aynı zamanda model parametrelerinin biyolojik anlamı üzerine yoğunlaşılarak bu konuda bilgi alışverişinde bulunabileceğimiz uzmanlarla işbirliği imkânları araştırılacaktır. Ön çalışmaların tamamlanması sonrasında moleküler bilgi içeriği çalışmalarının TÜBİTAK 1001 destek programına önerilecek bir proje başvurusu ile sürdürülmesi planlanmaktadır.

KAYNAKLAR

- Abe, N., Mamitsuka, H. 1997. “Predicting protein secondary structure using stochastic tree grammars”, Machine Learning, 29, 275-301.
- Alberts, B., Bray D., Lewis, J., Raff, M., Roberts, K., Watson, J. D. (2007). “Molecular Biology of the Cell (5th Ed.)”. New York: Garland Science.
- Basarin, G. P. 1959. “On a statistical estimate for the entropy of a sequence of independent random variables”, Theory of Probability and Its Applications, 4, 333-336.
- Bieomont, C., Vieira, C. 2006. “Genetics: Junk DNA as an evolutionary force”, Nature, 443 (7111), 521-524.
- Bose, R., Chouhan, S. 2011. “Alternate measure of information useful for DNA sequences”, Physical Review E, 83, 051918.
- Brendel, V., Busse, H.G. 1984. “Genome structure described by formal languages”, Nucleic Acids Research, 12, 2561–2568.
- Chiang, D., Aravind, K.J., Searls, D.B. 2006. “Grammatical representations of macromolecular structure”, Journal of Computational Biology, 13, 1077-1100.
- Crick, F.H.C. 1958. “On protein synthesis”, The Symposia of the Society for Experimental Biology XII, 139–163.
- Eroglu, S. 2013. “Menzerath-Altmann law for distinct word distribution analysis in a large text”, Physica A, 392(12) 2775–2780.
- Eroglu, S. 2014a. “Menzerath-Altmann law: Statistical mechanical interpretation as applied to a linguistic organization”, Journal of Statistical Physics, 157(2), 392-405.
- Eroglu, S. 2014b. “Language-like behavior of protein length distribution in proteomes”, Complexity, 20(2), 12-21.
- Eroglu, S. 2014c. “Self-organization of genic and intergenic sequence lengths in genomes: Statistical properties and linguistic coherence”, Complexity, DOI: 10.1002/cplx.21563.
- Eroglu, S. 2017. “Information Content Estimate of Proteomes: A Primary Structure Perspective”, Current Bioinformatics, DOI: 10.2174/1574893612666161215165052 (Yayın aşamasında).
- Gimona, M. 2006. “Protein linguistics—A grammar for modular protein assembly?”, Nature Reviews Molecular Cell Biology, 7, 68-73.
- Grosse, I., Herzel, H., Buldyrev, S. V., Stanley H. E. 2000. “Species independence of mutual information in coding and noncoding DNA”, Physical Review E, 61, 5624-5629.
- Hartley, R. V. L. 1928. “Transmission of Information”, Bell System Technical Journal, 7, 535-563.

- Herzel, H., Ebeling, W., Schmitt, A. O. 1994. “Entropy of biosequences: The role of repeats”, Physical Review E, 50, 5061-5071.
- Ji, S. 1999. “The linguistics of DNA: Words, sentences, grammar, phonetics, and semantics”, Annals of the New York Academy of Sciences, 870, 411–417.
- Ji, S. 2002. “Microsemiotics of DNA”, Semiotica, 138, 15–42.
- Ji, S., Ciobanu, G. 2003. “Conformon-driven biopolymer shape changes in cell modeling” Biosystems, 70, 165-181.
- Machado, J. A. T., Costa, A. C., Quelhas, M. D. 2013. “Can Power Laws Help Us Understand Gene and Proteome Information?”, Advances in Mathematical Physics, Art. ID. 917153.
- Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M., Stanley, H.E. 1994. “Linguistic features of noncoding DNA sequences”, Physical Review Letters, 73, 3169-3172.
- Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M., Stanley, H.E. 1995 “Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics”, Physical Review E, 52, 2939-2950.
- Motomura, K., Fujita, T., Tsutsumi, M., Kikuzato, S., Nakamura, M., Otaki, J.M. 2012. “Word decoding of protein amino acid sequences with availability analysis: A linguistic approach”, PLoS ONE, 7, e50039.
- Nyquist, H. 1928. "Certain topics in telegraph transmission theory", Transactions of the AIEE, 617–644 (Reprinted as Classical Paper: 2002. Proceedings of the IEEE, 90, 280-305).
- Ohno, S. 1972. “So Much ‘Junk DNA’ in Our Genome”. Smith, H.H. (Ed.). Gordon and Breach, New York, 366-370.
- Penisi, E. 2012. “Genomics. ENCODE Project writes eulogy for junk DNA”, Science, 337 (6099), 1150-1161.
- Popov, O., Segal, D.M., Trifonov, E.N. 1996. “Linguistic complexity of protein sequences as compared to texts of human languages”, Biosystems, 38, 65-74.
- Przytycka, T., Srinivasan, R., Rose, G.D. 2002. “Recursive domains in proteins”, Protein Science, 11, 409-417.
- Rose, A.B. 2008. “Intron-mediated regulation of gene expression”, Current Topics in Microbiology and Immunology, 326, 2772-90.
- Sadovsky, M. G. 2006. “Information capacity of nucleotide sequences and its applications”, Bulletin of Mathhematicla Biology, 68, 785-806.
- Schmitt, A. O., Herzel, H. 1997. “Estimating the entropy of DNA sequences”, Journal of Theoretical Biology, 188, 369-377.

• Schneider, T. D. 1997. “Information content of individual genetic sequences”, Journal of Theoretical Biology, 189, 327-441.

• Searls, D.B. 1997. “Linguistic approaches to biological sequences”, Bioinformatics, 13, 333–344.

• Shannon, C. E. 1948. “A mathematical theory of communication” Bell System Technical Journal, 27, 379-473, 623-656.

• Smith, J. M. 2000. “The concept of information in biology”, Philosophy of Science, 67, 177-194.

• Strait, B. J., Dewey, T. G. 1996. “The Shannon information entropy of protein sequences”, Biophysical Journal, 71, 148-155.

• Weiss, O., Jimenez-Montano, M. A., Herzel, H. 2000. “Information content of protein sequences”, Journal of Theoretical Biology, 206, 379-386.

• White, S. H., Jacobs, R. E. 1990. “Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution”, Biophysical Journal, 57, 911-921.

TÜBİTAK
PROJE ÖZET BİLGİ FORMU

Proje Yürütücüsü:	Doç. Dr. SERTAÇ EROĞLU
Proje No:	114F504
Proje Başlığı:	Proteom Ve Genomlarin Bilgi İçeriklerinin Modellenmesi
Proje Türü:	3001 - Başlangıç AR-GE
Proje Süresi:	18
Araştırmacılar:	
Danışmanlar:	
Projenin Yürütüldüğü Kuruluş ve Adresi:	ESKİŞEHİR OSMANGAZİ Ü. FEN EDEBİYAT F. FİZİK B.
Projenin Başlangıç ve Bitiş Tarihleri:	01/05/2015 - 01/08/2017
Onaylanan Bütçe:	72084.0
Harcanan Bütçe:	33635.34
Öz:	Genel olarak bilgi, olayların/olguların elle tutulup gözle görülemeyen bir özelliğidir ve bilgiyi alan alıcının durumunda (halinde) değişiklik yaratmak için gereken şartlara sahiptir. Komünikasyon mühendisliğinde bilgi teorisinin matematiksel temeli 1948?de Claude Shannon tarafından atılmıştır. O zamandan günümüze bilgi teorisi, DNA ve proteinler gibi bilgi taşıyan sistemleri incelenmesinde de kullanılmaktadır. Bilgi teorisine göre, bir sistemin içerdiği yapısal bilgi, o sistemi oluşturan bileşenlerin yapısal organizasyonunda gizlidir. Semboller dizininden oluşan bir mesajın içerdiği bilginin miktarı, mesajda yer alan semboller arasındaki korelasyon/düzen ile ilişkilidir. Benzer şekilde bir mesaj kaynağının içerdiği bilginin miktarı, kaynakta yer alan mesajlar arasındaki korelasyon/düzen ile ilişkilidir. Sürdürmekte olduğumuz araştırma çalışmalarımız, sembollerden oluşmuş bilgi içeren dizinlerin oluşturduğu kaynaklarda ortak bir yapısal düzen özelliğinin varlığı ortaya çıkarmış ve sonuçlar yayınlanmıştır. Tamamlanan bu projede proteom ve genom kaynaklarının içerdikleri bilginin modellenmesinde takip edilecek strateji, literatüre kazandırılmış olan bu ortak yapısal özelliğin modellenmesinde kullanılmıştır. Bu projede, özgün bir yaklaşımla proteom ve genom kaynaklarının içerdikleri bilginin nicel ölçütünün teorik modellenmesi yapılmıştır. Model ile farklı üst âlemlerden seçilmiş model organizmaların proteom ve genomlarının bilgi içerikleri elde edilerek, sonuçlar fiziksel ve biyolojik bakış açılarıyla yorumlanmıştır. 10 proteomdan oluşan bir model proteom setinin sahip olduğu ölçülen, model ile tahmin edilen (dildeki uzunluk dağılım modeli) ve simüle edilen (gelişigüzel uzunluk dağılımı) bilgi içeriklerini analiz ettik ve karşılaştırdık. Proteinlerin ölçülen ve model ile tahmin edilen bilgi içeriklerinin mükemmel bir şekilde uyumlu oldukları gözlenmiştir. Proteomik protein başına ortalama bilgi ölçülen/modellenen durum için 8 ve simüle edilen durum için 7 bit olarak elde edilmiştir. Araştırma, biyolojik etkileşme mekanizmalarının temelde etkileşmeyi başlatan protein dizininin aminoasit düzeninden öteye protein dizinindeki aminoasitlerin sayısı ile daha ilişkili olabileceğini ortaya koyar. Bu çalışmada sunulan yöntem bir organizmada ya da bir grup organizmada gerçekleşen biyolojik süreçlerin karşılaştırılmasında pratik bir yaklaşım olmakla birlikte; birincil ve ikincil yapı gibi farklı hiyerarşik yapılarda proteomik ve genomik sayısız bilgi karakteristiği araştırmalarına olanak sağlama potansiyeline sahiptir. Proje çıktılarının şu ana başlıklarla verilen yaygın etkilerinin olacağı öngörülmektedir: i) Proteom ve genomların bilgi ölçütünü veren özgün bir modelin literatüre kazandırılması, ii) Bu model ile literatürde cevap bekleyen bazı önemli soruların farklı bir açıdan ele alarak cevaplanması ve bu konudaki bazı boşlukların doldurulması, iii) Dolayısıyla canlı organizmalarda gerçekleşen moleküler bilgi depolama/iletimi süreçlerinin daha iyi anlaşılması, iv) Yeni birtakım araştırma sorularının bilimsel gündeme taşınması.
Anahtar Kelimeler:	Proteomik bilgi, bilgi teorisi, protein uzunluk dağılımı, Menzerath-Altmann dağılım modeli
Fikri Ürün Bildirim Formu Sunuldu Mu?:	Hayır

TÜBİTAK