DERLEME REVIEW

DOI: 10.5336/biostatic.2017-55474

Gen Ağlarının Matematiksel Modellenmesi

Mathematical Modeling of Gene Networks: Review

Vilda PURUTÇUOĞLU,[a]
Ezgi AYYILDIZ[a]

[a]İstatistik Bölümü,
ODTÜ,
Ankara

Geliş Tarihi/*Received:* 28.02.2017
Kabul Tarihi/*Accepted:* 07.04.2017

Yazışma Adresi/*Correspondence:*
Vilda PURUTÇUOĞLU
ODTÜ,
İstatistik Bölümü, Ankara,
TÜRKİYE/TURKEY
vpurutcu@metu.edu.tr

ÖZET Sistem ve hesaplamalı biyoloji alanları, deneysel teknolojinin ilerlemesiyle gelişen ve biyolojik/kimyasal sistem olaylarının anlaşılmasını sağlayacak matematiksel metotları ve modelleri kapsamaktadır. Bu modeller sistem elemanlarının davranışlarının daha iyi anlaşılmasına ve gerektiğinde, bu sistemlerin simülasyonlar yardımıyla yaratılmasına ve birbirleriyle karşılaştırılmasına yardımcı olmaktadır. Ayrıca, matematiksel modellemeler, sistem hakkındaki mevcut bilgilerin doğruluğunun da test edilmesine ve laboratuvar ortamında yapılması maliyetli olabilecek deneyleri yorumlayabilmemize olanak sağlamaktadır. Bu çalışmada, farklı varsayımlara ve veri çeşitlerine göre karmaşık gen ağlarının nasıl modellenebileceği matematiksel ifadelerle açıklanacak ve önerilen alternatif modellerden çıkabilecek ağlar, grafiksel şekillerle anlatılacaktır. Burada sunulan temel yaklaşımların gerçek veri setleri üzerinde uygulanması durumunda, sistemlerden biyolojik olarak önemli olabilecek sonuçların elde edilmesinin mümkün olabileceği düşünülmektedir. Ayrıca, tanıtılan bu modellerin sadece gen ağlarında değil, yüksek boyutlu başka veri analizlerinde de kullanılması mümkündür. Nitekim sistem biyolojinin alt dallarından olan sistem modellemesi ve gen ağlarının modellenmesi, temelde, yüksek boyutlu ve çok değişkenli verilerin analizine imkan veren, karmaşık modellemeler üzerine yoğunlaşmıştır, ve bu çalışma da bunun bir örneğidir.

Anahtar Kelimeler: Gen düzenleyici şebekeler; modeller, teorik; sistem biyolojisi; sistem modellemesi; gen ağlarının modellenmesi

ABSTRACT System and computational biology areas cover mathematical methods and models that improve with the advancement of experimental technology and enable us to understand biological/chemical systems. These models help to better understand the behaviors of system components, and when necessary, can be able to generate and compare these systems via simulations. Furthermore, mathematical models can be able to validate the current knowledge about the systems and enables us to interpret the experiments that can be expensive to perform in the wet-lab. In this study, we explain how complex gene networks can be modelled by using mathematical expressions under various data types and assumptions, and the networks which are constructed by the suggested alternative models are represented by graphical figures. It is expected to obtain biologically valuable findings from these underlying systems when these fundamental approaches presented here are applied to real data set. Moreover, it is possible to implement these models not only in the field of gene networks, but also, in the application of other high dimensional data analyses. As a matter of fact, systems modeling and gene network modeling are the branches of the systems biology and focus on complex models that enable us to model high-dimensional and multivariate data, and this study is an example of this.

Keywords: Gene regulatory networks; models, theoretical; systems biology; system modelling; gene network modelling

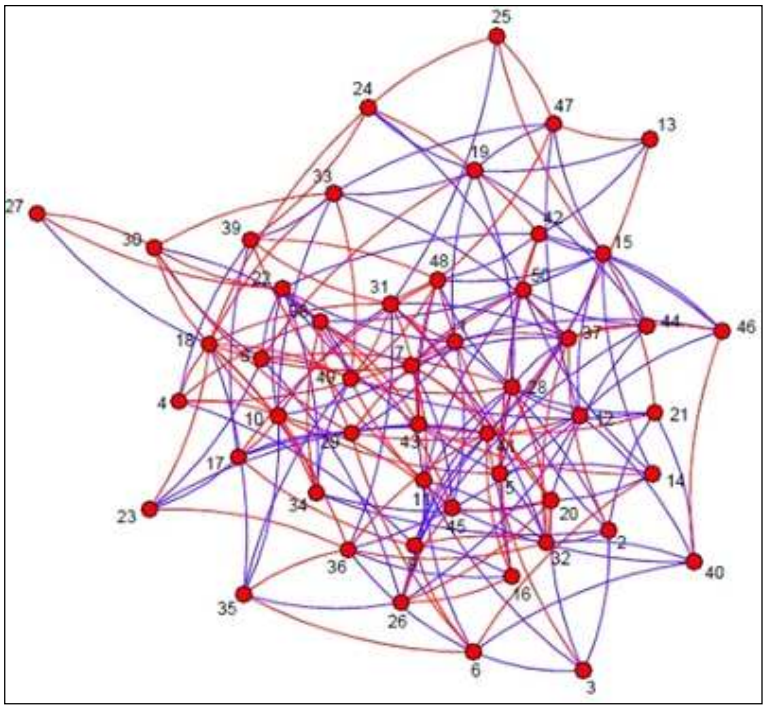
Herhangi bir sistemin elemanlarının etkileşimini gösteren yapıya *ağ* denir. Beklenildiği gibi, farklı alanlarda farklı ağ yapıları kullanılmaktadır. Bunlara, sosyal, biyolojik ve bilgisayar ağları örnek verilebilir. Ağ bileşenleri *düğümler* ve *kenarlar*dır. *Düğümler* sistemin elemanlarını, *kenarlar* da düğümler arası etkileşimi, yani ilişkiyi, temsil eder. *Çizge teorisinde* (graph theory) düğümler çemberle, kenarlar ise çizgilerle gösterilir. Biz de aynı yolu izleyerek, bu çalışmada, gen ağlarıyla ve bunların matematiksel modellenmesiyle ilgileneceğiz.[1,2]

Biyolojik ağlar, kullandıkları sistem elemanı bakımından, temelde moleküler ve proteomik ağlar olmak üzere ikiye ayrılır. Metabolik ağlar moleküler ağlara; protein-protein etkileşim ağları ve gen düzenleyici ağlar ise proteomik ağlara örnek olarak verilebilir. Bu çalışmada sadece *gen düzenleyici ağların*, kısaca *gen ağlarının*, tanımı ve matematiksel modellenmesi açıklanacaktır.

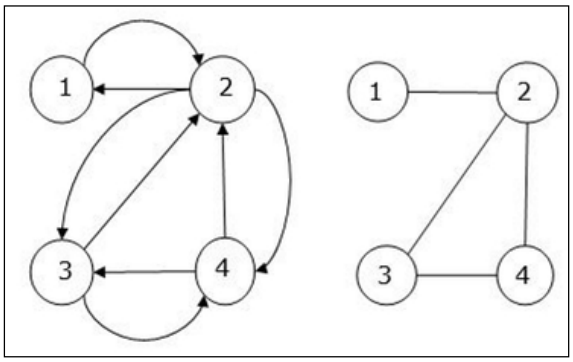
Gen ağlarında düğümler genleri, kenarlar da gen etkileşimlerini gösterir. 50 düğümlü karmaşık bir ağ yapısı örnek olarak verilmektedir (Şekil 1). Ağlar ayrıca *yönlü* ve *yönsüz* olmak üzere de sınıflandırılabilir (Şekil 2). *Yönlü ağlar*, kenarların düğümler arası ilişkinin yönünün oklarla belirtildiği ağ çeşitidir. *Yönsüz ağlar* ise sadece ilişkilerin varlığını ifade eder; fakat yönleriyle ilgili bilgi içermez. Dolayısıyla bu bölümde öncelikle ağ modellerinin ne olduğu açıklanacak ve farklı varsayımlar altında nasıl modellenebilecekleri anlatılacaktır.

TEPKİME

Kimyasal tepkimeler biyolojik sistemlerin temelini oluşturur. Tepkimeler, karmaşık kimyasal süreçlerin ortak bir gösterimle tanımlanmasına olanak sağlar. Kimyasal tepkimelerin tanımlanması modellemede önemli bir yer tutar. Çünkü aynı kimyasal tepkime kümesi, farklı detaylandırmalar sebebiyle farklı mo-



ŞEKİL 1: 50 düğümlü ağ örneği.



ŞEKİL 2: (a) Yönlü ve **(b)** yönsüz ağ yapısı.

dellerin oluşmasına sebep olabilir. Süreçlerin kimyasal denklemlerle tanımlanması, matematiksel modellerin uygulanabilmesini sağlar. Örneğin;

$$n_x X + n_y Y \xrightarrow{k} n_z Z + n_w W,$$

şeklinde genel bir kimyasal tepkimeyi ele alalım. Burada X ve Y molekülleri tepkimeye girerek Z ve W moleküllerini oluşturur. Bu gösterimde okun sol tarafındaki moleküllere *tepken*, sağ tarafındakilere ise *ürün* adı verilir. Bir kimyasal tepkimede sonlu sayıda tepken ve ürün bulunabilir ve bunların sayılarının eşit olması gerekmez. Denklemde tepken ve ürünlerin önlerindeki katsayılar (n_x, n_y)molekül sayılarını gösterir ve *stokiyometrik katsayısı* (stoichiometric coefficient) olarak adlandırılır. Tepkime okunun üzerindeki k değerine ise *hız sabiti* denir. Bu değer, tepkimenin gerçekleşme süresinin belirlemesinde kullanılır.[3,4]

MODELLEME

Sistemin elemanlarını, onların durumlarını ve birbiriyle olan etkileşimlerini keşfetmek için modellemeden yararlanılır. Dolayısıyla modelleme, bir sistemi anlayabilmenin ve analiz edebilmenin en temel yoludur.

Fizik, kimya, biyoloji, ekonomi, psikoloji gibi içinde sistem barındıran bütün bilimlerde modelleme kullanılır. Moleküler biyolojide, özellikle gen düzenlemesi, translasyonu, metabolik yollar, hücre döngüsü, hücresel sinyaller, protein etkileşimleri gibi mekanizmaları anlamak için modelleme yapılmaktadır. Biyolojik sistemler oldukça büyük ve karmaşık yapılara sahiptir. Dolayısıyla böyle bir sistemi matematiksel olarak ifade edebilmek, onu anlayabilmemize, hipotezleri test edebilmemize ve sistemin davranışıyla ilgili tahminde bulunabilmemize imkan sağlar. İyi bir model, sistemin önemli özelliklerini kapsamalı ve belli seviyelerdeki detaylardan arındırılmış olmalıdır.

Matematiksel modeller, sistem elemanlarını ve onların ilişkilerini matematiksel ifadelerle temsil ettiği için sistem analizlerini ve grafiksel gösterimlerini mümkün kılar. Gen ağları için matematiksel modeller temelde *deterministik* ve *stokastik* olmak üzere ikiye ayrılır. *Deterministik modeller*, sistemleri, temel mekanik kurallarını rasgelelik kullanmadan tanımlar. Bu modellerde sistemin şu anki durumu ve bir sonraki davranışı tamamen belirlenebilir. Çünkü aynı sistem elemanları, aynı etkileşimle, her zaman aynı sonuçları verir. *Stokastik modeller* ise rastgeleliği göz önünde bulundurur ve bu sayede sistemin dinamik yapısını temsil edebilir. Bu ifadelerde sistem elemanlarının davranışları olasılık dağılımlarının yardımıyla açıklanır.[2,4]

GEN AĞLARININ MATEMATİKSEL MODELLEMESİ

Gen ağları, çok sayıda değişken içeren büyük ve karmaşık sistemlerdir. Biyolojik verilerin değişken sayılarına kıyasla, örnek sayıları da oldukça düşüktür. Bu durum, incelenmek istenen sistemlerin analizini de güçleştirir.

Biz burada özellikle gen etkileşim sistemlerinin ağ yapılarının modellenmesini, basitten karmaşığa doğru tek tek inceleyeceğiz.

BOOLEAN AĞ MODELİ

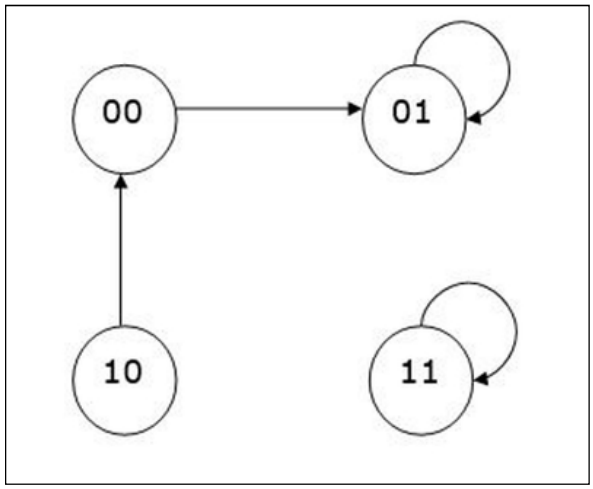
Boolean yaklaşımı gen ağlarının modellenmesinde kullanılan en basit modeldir.[1,2,5] Bu yaklaşım, biyokimyasal modelleri açıklarken sistem durumlarını, genlerin tamamen aktif olup olmadığı bilgisiyle temsil eder. Bunu yaparken de ikili (yani 0-1) mantığı kullanır. Yani aktif olan geni 1, aktif olmayan geni 0 ile gösterir. Boolean modeli, sonlu sayıda durum ve bir sonraki durumu belirlemek için kullanılan Boolean fonksiyonundan oluşur. Bu yapısı Boolean modelinin deterministik bir model olmasını sağlar. Ayrıca kullanılan ikili mantık sistemin 2^n kadar durum içermesini sağlayıp, diğer çoklu (multinominal) durum modellerine göre, model parametre sayısının azalmasına yardımcı olur. Burada n, sistemdeki gen sayısını belirtmektedir.

TABLO 1: VE, VEYA, DEĞİL operatörleri için doğruluk tablosu. D ve Y, sırasıyla, doğru ve yanlış ifadelerini göstermektedir.

p	q	p VE q	p VEYA q	p DEĞİL	q DEĞİL
D	D	D	D	Y	Y
D	Y	Y	D	Y	D
Y	D	Y	D	D	Y
Y	Y	Y	Y	D	D

TABLO 2: İki genli bir sistem için örnek bir durum geçiş tablosu. 0 ve 1, sırasıyla, genlerin aktif ve aktif olmama durumlarını gösterir. Burada A(sonraki) = A(mevcut) VE B(mevcut), B(sonraki)= A DEĞİL (mevcut) VEYA B(mevcut) fonksiyonları kullanılmıştır.

Mevcut Durum	Sonraki Durum
00	01
01	01
10	00
11	11



ŞEKİL 3: Tablo 2'de gösterilen sistem ve durum geçiş tablosu için sonlu durum makinesi.

Sistemleri Boolean modeliyle ifade edebilmenin iki temel yöntemi vardır. Bunlar, durum geçiş tablosu ve sonlu durum makinesidir. Durum geçiş tablosu, sistemdeki genlerin mevcut durumunu ve bir sonraki durumunu gösteren iki sütundan oluşur. Bir sonraki durum Boolean fonksiyonuyla belirlenir. Bu fonksiyon VE, VEYA, DEĞİL şeklindeki 3 mantık operatörünü kullanır. (p, q) ikili değişkeni için bu 3 operatörün doğruluk tablosu verilmiştir (Tablo 1).

Örnek olarak 2 gen (A, B) ve 2 Boolean fonksiyonundan oluşan küçük bir ağ ele alalım. Her genin bir sonraki durumunu belirleyen bu fonksiyonlar,

A(sonraki) = A(mevcut) VE B(mevcut)

B(sonraki)= A DEĞİL(mevcut) VEYA B(mevcut)

şeklinde olsun. A ve B genlerinden oluşan bu sistem için $2^2 = 4$ durumlu örnek bir durum geçiş tablosu yaratılabilir (Tablo 2).

Bu örnekte, durum geçiş tablosunun ilk satırı, mevcut durumda A ve B genlerinin aktif olmadığını ve B geninin sonraki durumda aktif olacağı bilgisini vermektedir. Aynı sistemin sonlu durum makinesi de ayrıca çizilebilir (Şekil 3). Sonlu durum makinesi, sistemi diyagramla gösteren bir yöntemdir. Bu diyagramda, durumlar, çember ve durumlar arası geçiş, oklarla temsil edilir. Buna göre Şekil 3'de verilen ör-

nekte 11 durumunun diğer durumlarla okları bulunmadığı için, bağımsız, diğer durumların ise bağımlı olduğu söylenebilir.

Boolean modeli, gen durumlarını kesikli değişkenlerle ve durumlar arası geçişi de basit mantık fonksiyonlarıyla belirler. Fakat biyolojik sistemler çok daha karmaşık bir yapıya sahiptir.

Bu karmaşık sistemleri tanımlayabilmek için daha fazla bilgi içeren modellere ihtiyaç duyulmaktadır. Kinetik mantık ve sürekli mantıksal ağ modeli Boolean'dan daha gelişmiş modellere örnek olarak verilebilir.

TABLO 3: İki gen ve üç seviyeli bir sistem için örnek bir durum geçiş tablosu. 0, 1 ve 2, sırasıyla, genlerin aktif olmama, düşük seviyede aktif olma ve yüksek seviyede aktif olma durumlarını gösterir.

Mevcut	İstenen sonraki	Olası sonraki
00	11	10, 01
01	00	00, 11, 02
02	00	01, 12
10	22	00, 11, 20
11	10	01, 10, 12, 21
12	10	11, 02, 22
20	22	10, 21
21	21	11, 20, 22
22	20	12, 21

KİNETİK MANTIK VE SÜREKLİ MANTIKSAL AĞ

Kinetik mantık modelinde, Boolean modelinde olduğu gibi durumlar kesikli değişkenle ifade edilir.[2,5] Fakat burada Boolean modelinden farklı olarak iki sonuçlu (binary) değişken yerine çoklu değişken (multinominal) kullanılır (0,1,2,3, ... gibi). Diğer bir deyişle gen durumlarını aktif ve aktif olmayan diye ikiye ayırmak yerine aktif, düşük seviyede aktif, orta seviyede aktif ve yüksek seviyede aktif gibi detaylar verilir. Ayrıca durumlar arası geçişi belirlemek için de eşik fonksiyonu kullanılır. Eşik fonksiyonu, değişkenin değerini, doğrusal bir denklemin sonucunun önceden belirlenen eşik değerden büyük ya da küçük olmasına bakarak belirler.

Doğruluk tablosu ve sonlu durum diyagramı kinetik mantık modelinde de sistemi tanımlamak için kullanılır. İki yöntemde de küçük yapısal farklılıklar mevcuttur. Bu tablolar mevcut durumu, istenen sonraki durumu ve olası sonraki durumu gösterir (Tablo 3).

Kinetik mantık modelinde durumlar arası geçiş adım adım gerçekleşir. Yani mevcut durumda düşük seviyede aktif olan bir gen sonraki durumda yüksek seviyede aktif olamaz. Önce orta seviyede aktif durumuna geçmesi gerekir. Ayrıca bu modelde iki gen aynı zamanda durum değiştiremez. Dolayısıyla olası sonraki durum, birden fazla olabilir. Örnek olarak 21 durumundaki A ve B genleri sonraki durumda 11 veya 20 değerlerini alabilirler. Yani A geni 2 seviyesinden 1 seviyesine ya da B geni 1 seviyesinden 0'a geçebilir. Dolayısıyla yapı Boolean modeline göre stokastik model yapısındadır. Nitekim sonraki durumları ifade etmek için, Boolean'dan farklı olarak, olası birden fazla durum yaratılabilir.

Bu olası durumlar, sonlu durum diyagramında çoklu oklarla temsil edilir. Örnek olarak 2 gen (A, B) ve 3 olası durumdan (0, 1, 2) oluşan bir ağ ele alalım. Bu ağ için bir durum geçiş tablosunun yaratıldığını düşünelim (Tablo 3). Burada mevcut durum kullanılarak, istenilen sonraki durumu belirleyen fonksiyonlar,

$$A(sonraki) = \begin{cases} 0, & 2 * A(mevcut) - B(mevcut) < 0 \\ 1, & 0 \leq 2 * A(mevcut) - B(mevcut) \leq 1 \\ 2, & 2 * A(mevcut) - B(mevcut) > 1 \end{cases}$$

$$B(sonraki) = \begin{cases} 0, & A(mevcut) - 3 * B(mevcut) < -1 \\ 1, & -1 \leq A(mevcut) - 3 * B(mevcut) \leq 0 \\ 2, & A(mevcut) - 3 * B(mevcut) > 0 \end{cases}$$ şeklinde olsun.

Bu durumda sonlu durum diyagramı da çizilebilir (Şekil 4). Bu örnekte, Boolean ağ modelinden farklı olarak, olası sonraki durumun birden fazla olabildiğini görüyoruz. Ayrıca bu farklılık, sonlu durum diyagramında bir düğümden birden fazla ok çıkmasıyla da temsil edilebilir.

Sürekli mantıksal ağ modeli ise durumları daha ileriki bölümlerde anlatılacak diferansiyel denklemlerdeki gibi sürekli yoğunluklarla temsil eder. Sürekli mantıksal ağ modelinde yoğunluk değişim katsayıları sabittir ve sistem basit doğrusal diferansiyel denklemlerle çözülebilir.

OLASILIKSAL BOOLEAN MODELİ

Olasılıksal Boolean modeli deterministik olan Boolean modelinin stokastik yapıdaki halidir.[6,7] Bu yapı bize gen ağlarının dinamik davranışını anlayabilme şansını verir. Olasılıksal Boolean modeli sistemlerin dinamik yapısını tarif ederken Markov zinciri teorisini kullanır. Birinci derece homojen Markov zincirinde $(t+1)$ zamanındaki durum sadece t zamanındaki duruma bağlıdır. Markov zincirinin bu özelliği, genlerin $(t+1)$zamanındaki durumunun sadece t zamanına bağlı olduğu bilgisiyle örtüşür.

Olasılıksal Boolean modelindeki "Boolean" terimi ikili nicelleştirme anlamına gelmemektedir. Bu modelde sonlu sayıda nicelleştirme mümkündür. Olasılıksal Boolean modelinin bileşenleri, düğüm dizisi ve vektör değerli fonksiyon (vector-valued function) dizisidir. Bunlar, sırasıyla, $V = \{X_i\}^n$ ve $\{f_l\}^m$ ile temsil edilir. Burada X_i genlerin ekspresyon değerini ifade eder ve $\{0,1,\dots,d-1\}$ kümesinin elemanıdır. n sistemdeki düğüm, m fonksiyon, d ise nicelleştirme seviyesinin sayısını gösterir.

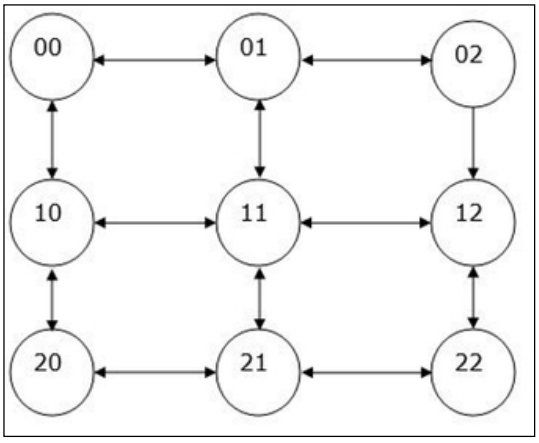
Vektör değerli fonksiyon yardımıyla durum geçiş olasılıkları hesaplanabilir. Bütün durum geçiş olasılıkları hesaplanırsa geçiş matrisi oluşturulmuş olur ve bu sayede modelin kararlı durum dağılımı (steady-state distribution) belirlenebilir. Sonuç olarak bu da bize modelin uzun dönem davranışını (long run behavior) açıklar. Başka bir deyişle, bu modeller sayesinde, "Uzun dönemde A geninin ekspresyonunun olasılığı nedir?" ya da "Uzun dönemde B ve C genlerinin birlikte ekspresyonunun olasılığı nedir?" gibi soruları cevaplayabiliriz.

Olasılıksal Boolean modelinde sistem, geçiş matrisinden yararlanılarak çizilen durum geçiş diyagramıyla gösterilir.

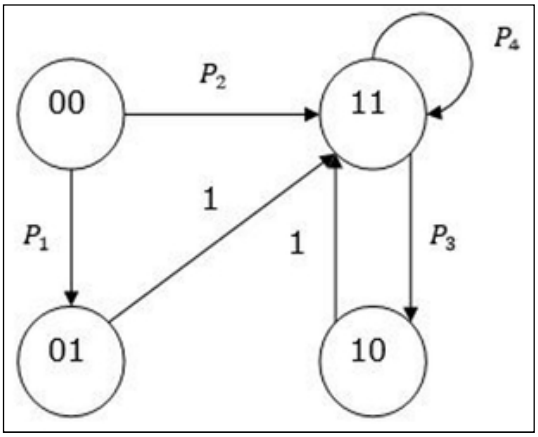
Bu diyagramın Boolean modelinde kullanılan sonlu durum diyagramından farkı, geçişleri gösteren okların geçiş olasılıklarını da belirtmesidir. Örnek olarak A ve B genlerinden oluşan 4 durumlu (sırasıyla, 00, 01, 10, 11) küçük bir sistem düşünelim. Bu sisteme ait geçiş matrisi aşağıda verilen G matrisi olsun.

$$G = \begin{bmatrix} 0 & P_1 & 0 & P_2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & P_3 & P_4 \end{bmatrix}. \quad (1)$$

Matrisin sütunları, sırasıyla, 00, 01, 10, 11 durumlarını göstermektedir. Buna göre geçiş matrisi G'ye ait durum geçiş diyagramı da çizilebilir (Şekil 5). Bu diyagramda da görülebileceği gibi sistemin, (10, 11) durumlarına geçiş yaptıktan sonra tekrar 00 veya 01 durumlarına geçiş yapması mümkün değildir. Burada (10, 11) durumları yutan durumlar (absorbing state) olarak adlandırılır.



ŞEKİL 4: Tablo 3'de gösterilen sistem ve durum geçiş tablosu için sonlu durum makinesi.



ŞEKİL 5: Eşitlik (1)'de gösterilen sistem için durum geçiş diyagramı.

Bir Markov zinciri olasılık dağılımı $\pi = (\pi_1, \pi_2, \dots, \pi_r)$aşağıdaki eşitliği sağlıyorsa durağan dağılıma sahiptir.

$$\pi_j = \sum_{i=1}^{r} \pi_i P^m.$$

Burada P^mm-adımda geçiş olasılığını göstermektedir. Ayrıca $\lim_{m\to\infty} P^m = \pi_i$ eşitliği sağlanıyorsa da Markov zinciri kararlı durum dağılımına sahip demektir. Böylece başlangıç durumu ne olursa olsun, uzun dönemde Markov zincirinin *j* inci durumda olma olasılığı π_i'e eşittir.

Ağın boyutları büyüdükçe, yani sistemdeki gen sayısı arttıkça, Markov zincirinin geçiş matrisini tamamen hesaplamak zorlaşacaktır. Çünkü durum sayısı üstel olarak büyümektedir. Bu durumda geçiş matrisinin her elemanını tek tek hesaplamak yerine başka yöntemler kullanılır. Bunlardan biri, biyolojik sistemlerin geçiş matrisinin seyrek yapısından yararlanmadır. Seyrek bir matrisin birçok elemanı sıfırdır ve geçiş matrisini oluşturmak için sıfırdan farklı elemanları hesaplamak yeterli olacaktır. Bu da hesaplamadaki zorluğu kolaylaştırmaktadır.[1,7]

DİFERANSİYEL DENKLEMLER MODELİ

Diferansiyel denklemler, türevler yardımıyla değişimi tanımlamanın matematiksel yoludur. Bu denklemlere dayalı modellerde gen durumlarının tanımlanması için kimyasalların yoğunlukları kullanılır. Bu yoğunluklar sürekli ve zamana bağlıdır. Yani zamanla değişimleri tepkime hız sabitine göre gerçekleşir ve bu tepkime hızları, adi diferansiyel denklemler kullanılarak hesaplanır. Diferansiyel denklemler deterministiktir. Dolayısıyla sistemlerin deterministik yapısının tanımlanmasını sağlar.[2,5,8] En genel haliyle aşağıda verilen kimyasal tepkimeyi düşünelim.

$$A \xrightarrow{k} A^*.$$

Bu tepkimenin hızı k[A] ile gösterilir ve burada [A], A maddesinin yoğunluğunu temsil eder. Bu tepkimede A'nın yoğunluğunun artması A^* üretiminin de doğrusal şekilde artmasına sebep olur. Bu durumda A ve A^* yoğunluğunun zamana bağlı değişimi diferansiyel denklemlerle aşağıdaki gibi yazılabilir.

$$\frac{d[A]}{dt} = -k[A] \quad \text{ve} \quad \frac{d[A^*]}{dt} = k[A].$$

Model parametrelerinin tespiti için bu diferansiyel denklemler eş zamanlı çözülmelidir.

Kimyasal tepkime sayısı ve tepkimelerin stokiyometrik katsayısı arttıkça, diferansiyel denklemler doğrusal olmayan denklemlere dönüşür. Bu da denklem sisteminin çözümünü zorlaştırır. Böyle durumlarda farklı yaklaşımsal yöntemler kullanılabilir. Fakat çözüm kümesinin tek olmaması, bu modelin en büyük dezavantajıdır. Ayrıca diferansiyel denklemlerin deterministik yapısı sebebiyle bu model, sonraki durumun birden fazla ihtimali olduğu sistemler için uygun değildir.[2]

GAUSSIAN GRAFİKSEL MODELİ

Daha önce de belirttiğimiz gibi grafiksel modeller yönlü ve yönsüz olmak üzere ikiye ayrılır. Yönsüz grafiklerde kenarlar sadece düğümler arası ilişkinin varlığını bildirir ve bu ilişkinin yönü hakkında bilgi içermez. Yönsüz grafiklerin en yaygın kullanıldığı model *Gaussian modelidir*.[7,9] Bu model, düğümlerin çok değişkenli Gauss (Normal) dağılımına sahip olduğu varsayımını kullanır. Düğümleri Y değişkeniyle temsil edersek bu varsayım aşağıdaki gibi gösterilebilir.

$Y \sim N(\mu, \sum)$.

Buradan genden oluşan bir sistem için $\mu = (\mu_1, \ldots, \mu_n)$ ortalamayı ve $(n \times n)$ boyutlu $\sum$ varyans-kovaryans matrisini verir. Gaussian grafiksel modeli düğümler arası ilişkinin yönünü göstermese de, düğümler arası koşullu bağımsızlık bilgisini içermektedir ve bu durum matematiksel olarak K, L ve M düğümleri için $K \perp L \mid \mathrm{M}$ şeklinde ifade edilebilir. Bunu, grafikte, iki düğüm arasındaki (K ve L), eldeki veriye (M) dayalı olarak, kenarın yokluğu ile gösterir (Şekil 6).

Bu yapılar biraz daha karmaşık ağ yapısı biçiminde de verilebilir. Örneğin beş düğümlü bir ağda koşullu bağımsızlıklar $1 \perp 5 \mid 4, 1 \perp 3 \mid 2,\ 2 \perp 5 \mid 4, 3 \perp 4 \mid 2 \text{ve} 3 \perp 5 \mid (2,4)$ şeklinde listelenebilir (Şekil 7).

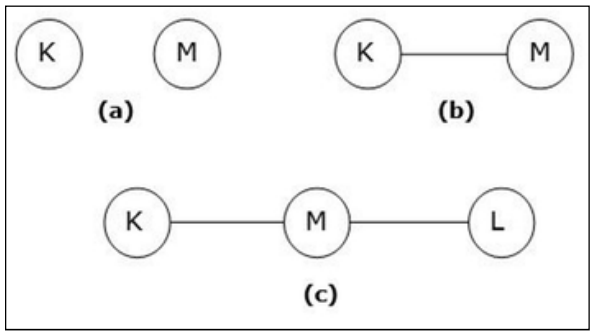
Grafiksel gösterime ek olarak *koşullu bağımsızlık,* kesinlik matrisindeki sıfır elemanlarıyla da ifade edilebilir. *Yoğunluk matrisi* olarak da adlandırılan *kesinlik matrisi*, varyans-kovaryans matrisinin tersidir ve Denklem (2)'deki gibi Θ ile gösterilir:

$$\Theta = \sum^{-1}. \qquad (2)$$

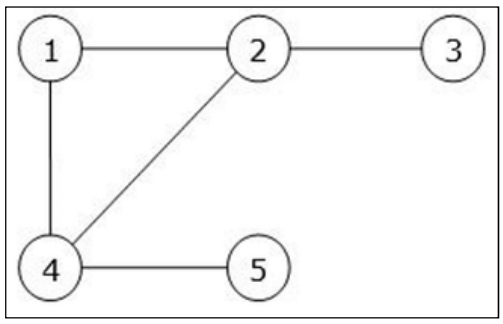
Kesinlik matrisi ile kısmi korelasyon matrisi arasında da matematiksel bir ilişki vardır. Bu ilişki Denklem (3)' de verilmektedir. Burada Θ'da, i ve j genleri arasındaki kesinlik değerini Denklem (3)' deki gibi θ_{ij} ile ifade edebiliriz. Bu durumda i ve j değerleri toplam n gene sahip bir model için n'ye kadar gider.

$$\pi_{ij} = \frac{-\theta_{ij}}{\sqrt{(\theta_{ii}\theta_{jj})}}. \qquad (3)$$

Diğer yandan Denklem (3)' teki π_{ij}, diğer bütün değişkenlerin verildiği durumda$Y^{(i)}$ve $Y^{(j)}$ değişkenleri arasındaki kısmi korelasyonu temsil eder.



ŞEKİL 6: (a) Bağımsız, **(b)** bağımlı ve **(c)** koşullu bağımsız düğümler.



ŞEKİL 7: Beş düğümlü ağ örneği.

Biyolojik veriler yapıları itibariyle çok fazla sayıda değişken ve buna oranla çok daha az sayıda örnekten oluştukları için standart kovaryans ve korelasyon matrislerinin kullanımı uygun olmaz. Çünkü örnek kovaryans tahmin edicisi içerdiği çok sayıdaki sıfır özdeğerinden ötürü tekil matrise dönüşür. Bu durum da kovaryans matrisinin tersinin hesaplanmasını imkansız hale getirir. Dolayısıyla pozitif tanımlı bir kovaryans matrisinin tahmin edilebilmesi önemli bir sorundur. Kısmi korelasyon matrisini hesaplamanın, kesinlik matrisini kullanmak dışında yolları da vardır. Bu yön-

temlerden biri regresyon analizidir. Bu analizde kısaca her düğümün kalan bütün düğümlerle regresyon analizi yapılır ve ortaya çıkan regresyon katsayıları (β_i) ağın koşullu bağımsızlık yapısını belirler. Yani $\beta_{ij} = 0$ ise $Y^{(i)}$ve $Y^{(j)}$ koşullu bağımsızdır.

Aşağıdaki bölümlerde belirtilen bu kısmi korelasyon matrisinin çözülmesi ile ilgili alternatif yöntemler anlatılmaktadır.

Lasso Tabanlı Yöntemler

Lasso tabanlı yöntemler, genellikle seyrek yapılı ağların tahmin edilmesinde kullanılırlar. Seyrek ağlar, az sayıda kenar içerir ve kesinlik matrisinde de çok sayıda sıfır bulunur. Seyreklik gen ağların genel özelliğidir. Regresyon tabanlı yöntemlerin, hesaplamadaki etkinliği ve değişkenlerin bileşik dağılımını doğru yakınsamak gibi önemli avantajlarının yanında varyans kovaryans matrisinin, dolayısıyla, kesinlik matrisinin simetrikliğini garantilemez. Seyrek ve simetrik bir kesinlik matrisi tahmin edebilmek için regresyon katsayıları yerine kesinlik matrisi elemanlarına L_1-cezalandırması (L_1-penalty) uygulanmalıdır.

Bu durumda Lagrangian ikili formuna göre cezalandırılmış en çok olabilirlik optimizasyonu aşağıdaki gibi yazılabilir.

$$\max_{\Theta}[\log|\Theta| - Trace(S\Theta) - \lambda|\Theta|_1]. \quad (4)$$

Burada λ, negatif olmayan Lagrange çarpanıdır. A vektörü için $|.|_1$ ifadesi $|A|_1 = \sum_i |A_i|$ göstermektedir ve $Trace(.)$ seçilen matris için iz değerini verir. Denklem (4)'de λ değeri arttıkça modellenen ağın seyrekliği de artar. Farklı λ değerleri için örnek bir ağın değişimi gösterilmiştir (Şekil 8).

Bu yöntemle elde edilen tahmin edici, kesinlik matrisinin simetrik olma şartını sağlar. Bu optimizasyon problemi farklı şekillerde çözülebilir. *Koordinat iniş yöntemi* (coordinate descent) de bunlardan biridir.[9,10]

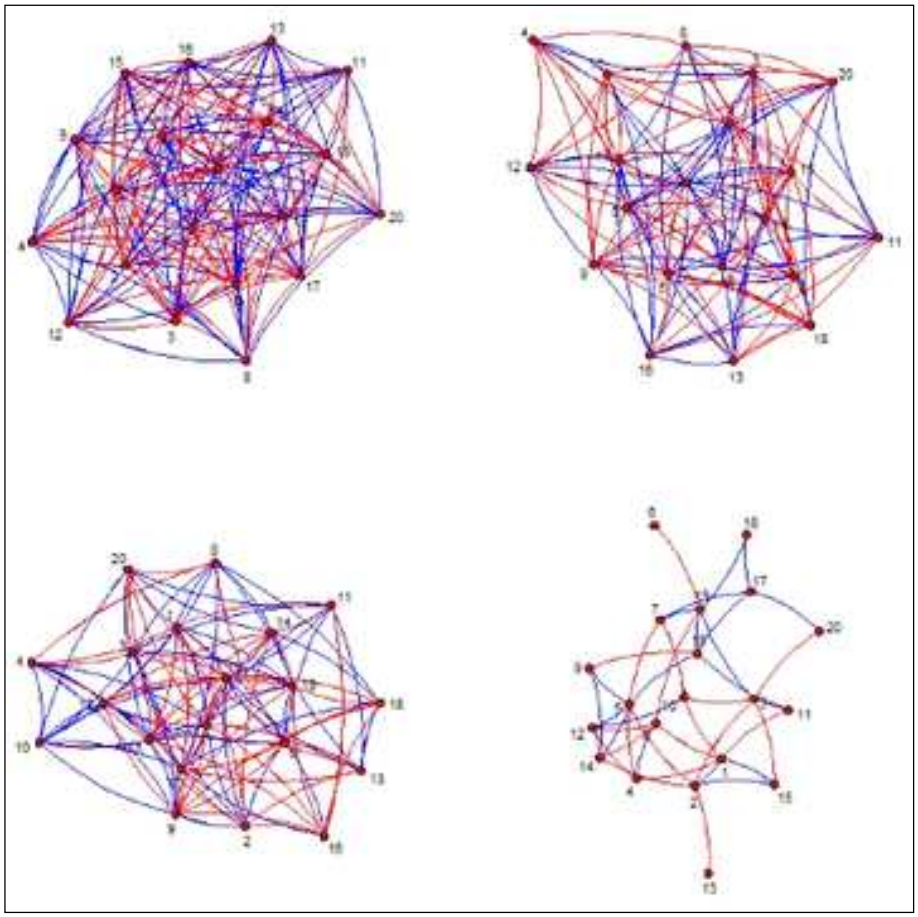
Küçültme (Shrinkage) Yöntemi

Küçültme yöntemi, kovaryans matrisini tahmin etmenin alternatif bir yoludur. Bu yöntemde hem örnek kovaryans matrisi hem de hedef matrisi (target matrix) kullanılır ve kitle kovaryans matrisi aşağıdaki formülle hesaplanır.

$$S^* = \lambda H + (1-\lambda)S_y.$$

Burada $S_y$yansız örnek kovaryansını $\left(S_y = \frac{n}{n-1}S\right)$, H hedef matrisini ve λ küçültme yoğunluğunu temsil eder. S_y yansız bir tahmin edici olmasına rağmen yüksek varyansa sahiptir. Hedef matrisi ise kovaryans matrisinden çok daha az parametre içeren bir matristir ve daha düşük varyanslıdır. Fakat yüksek yanlılık içerir. Son olarak küçültme yoğunluğu her zaman $[0, 1]$ aralığında değer alan bir oranı verir ve bu değer S_y ve H matrislerini uygun bir katsayı kullanarak birleştirir. Dolayısıyla küçültme yöntemiyle ortaya çıkan yeni tahmin edici S^*, örnek kovaryansından daha düşük bir ortalama karesel hataya sahip olur.

Küçültme yönteminde hedef matrisinin ve küçültme yoğunluğunun seçimi, iyi bir tahmin edicinin bulunmasında önemli rol oynar. Literatürde birim matrisinin ve onun skaler çarpımlarının hedef matrisi olarak kullanılması önerilir. Biyolojik problemlerde genel olarak farklı varyanslardan oluşan köşegensel matrisler hedef matrisi olarak kullanılır. Çünkü bu matris kullanılarak oluşturulan küçültme tahmin edi-



ŞEKİL 8: Farklı λ değerleri için örnek ağlar (soldan sağa doğru birinci ve ikinci satırda kullanılan λ değerleri, sırasıyla, λ=0.01, 0.07, 0.11 ve 0.15).

cisi otomatik olarak pozitif tanımlı yapıdadır. Ortalama karesel hatayı en aza indirgeyen λ değeri de küçültme yoğunluğu olarak seçilir.[11]

STOKASTİK MODEL

Stokastik model, kullandığı bilginin detayı sebebiyle, Boolean ve diferansiyel denklem modellere göre daha kapsamlı bir modeldir. Bu model, sistemi açıklarken genin aktif olup olmadığı ya da kimyasalın yoğunluğu yerine, molekül sayılarını modelin girdileri olarak kullanır. Bu modelde, durumlardaki değişim, kesikli niceliklerle açıklanırken, hangi değişimin ne zaman gerçekleşeceği olasılıksal olarak ifade edilir. Birim zamanda gerçekleşecek kesikli olayın olasılığı ise *tepkime hız sabiti* ile belirlenir.[2,4]

En genel haliyle

$$A \xrightarrow{k} A^*$$

şeklinde bir kimyasal tepkime düşünelim. Burada k tepkime hız sabitini temsil eder. Bu tepkime sözel olarak bir molekül A'nın A^*'a dönüştüğü şeklinde yorumlanır. Verilen t zamanda bu olayın gerçekleşme olasılığı kt'ye eşittir. Bir miktar A molekülünün belirli t zamanında A^*'a dönüşmesi ise $k\{\#A\}t$ ile hesap-

lanır ve {#A} , A'nın içerdiği molekül sayısını gösterir. Stokastik modelleme, bir sonraki bölümde anlatılacak olan kimyasal ana denkleme dayanmaktadır.

Kimyasal Ana Denklem

Durumlar molekül sayılarıyla temsil edildiğinde, bu durumlarda bulunma olasılığı kimyasal ana denklem yaklaşımında değişken olarak kullanılabilir. Bu değişkenler, zamanın birer fonksiyonu olduğu için olasılıklardaki değişimin hesaplanmasında diferansiyel denklemlerden yararlanılabilir.

Molekül sayısı Y ile temsil edilirse, t zamanında Y durumunda bulunma olasılığı, bileşik olasılık dağılımı $p(Y,t)$ ile aşağıdaki gibi yazılabilir.

$$p(Y,t+\Delta t) = p(Y,t)\left(1-\sum_{j=1}^{r}\alpha_j \Delta t\right)+\sum_{j=1}^{r}\beta_j \Delta t.$$

Bu denklemde, r sistemdeki toplam tepkime sayısını gösterir. t zamanında Y durumunda bulunan sistemde, $[t,t+\Delta t]$ zaman aralığında j tepkimesinin gerçekleşme olasılığı $\alpha_j \Delta t$ ile temsil edilir. Ayrıca $\beta_j \Delta t$, tepkime j'nin $[t,t+\Delta t]$ zaman aralığında sistemi başka bir durumdan Y durumuna getirme olasılığını belirtir. Buna göre, *kimyasal ana denklemler* (chemical master equations) Y'nin davranışını diferansiyel denklemleri kullanarak aşağıdaki gibi tanımlar.[4,12]

$$\frac{\partial}{\partial t}p(Y,t)=\sum_{j=1}^{r}\left(\beta_j-\alpha_j\, p(Y,t)\right).$$

Kimyasal ana denklemler, sadece olası durum sayısının az olduğu sistemler için kullanışlı bir yöntemdir. Çünkü sistemdeki durum sayısı arttıkça diferansiyel denklemler de hızlı bir şekilde karmaşık bir yapıya ulaşır. Sonuçta da kimyasal ana denklemlerin model parametrelerinin tahmininde kullanılması imkansız hale gelir. Bu durumda kimyasal ana denklemin stokastik diferansiyel denklemce yaklaşımı olan Langevien denklem modeli kullanılabilir. Bu model bir sonraki bölümde anlatılacaktır.

LANGEVİEN VE DİFÜZYON MODELİ

Langevien modeli, diferansiyel denklem modelinin gürültü terimi eklenerek genişletilmiş şeklidir ve aşağıdaki gibi yazılabilir:[4,13]

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathrm{Y}(t)=\mu(Y,\Theta)+W(t). \qquad (5)$$

Burada t zamanı, $W(t)$ zamana bağlı stokastik süreci ve $\mu(Y,\Theta)$ durumlar arası değişimin ortalamasını temsil eder. Modele eklenen stokastik süreç, $W(t)$, Y durumunun deterministik değerine stokastik dalgalanma katmak için kullanılmaktadır.

Modeldeki rastgelelik sebebiyle, Denklem (5) klasik diferansiyel denklem tekniğiyle çözülememektedir. Bunun yerine bazı varsayımlar altında Y durumunun beklenen değeri hesaplanmaktadır. Bu varsayımlar, $W(t)$'nin ortalamasının ve iki farklı zaman için korelasyon değerinin sıfır olması şeklindedir.

Difüzyon modeli ise aşağıdaki gibi yazılabilir.

$$\mathrm{dY}(t)=\mu(Y,\Theta)dt+\beta^{\frac{1}{2}}(Y,\Theta)dW(t).$$

Burada $\mu(Y,\Theta)$ ortalamayı, $\beta(Y,\Theta)$ varyansı, dt zamandaki değişimi ve $dW(t)$ Brown hareketindeki t'ye bağlı değişimi gösterir. Tüm bu ifadeler durum Y ve model parametresi olan Θ'nın fonksiyonlarıdır.

Difüzyon sürecinde, gözlenen son noktadaki durumların eksikliği sebebiyle *kesikli difüzyon modelinin* oluşturulması ihtiyacı doğmaktadır. *Euler-Maruyama yaklaşımı* olarak adlandırılan bu model, örneklem için eksiksiz veriyle en çok olabilirlik fonksiyonunun hesaplanabilmesine imkan sağlar.[14,15]

Bu durumda kesikli difüzyon modeli aşağıdaki gibi yazılabilir:

$$\Delta Y(t) = \mu(Y(t),\Theta)\Delta t + \beta^{\frac{1}{2}}(Y_t,\Theta)\Delta W(t).$$

Bu denklemde diğer ifadelerde olduğu gibi $Y(t) = (Y_1(t), \dots, Y_n(t))$ sistemin t zamandaki durumunu, $\Theta = (c_1, \dots, c_r)$ parametre vektörünü, n ve r, sırasıyla, sistemdeki toplam tür veya gen ve toplam tepkime sayısını temsil eder. $\Delta W(t)$ ise sıfır ortalamaya ve kesikli zaman aralığı Δt 'ye eşit varyansa sahip normal dağılımdan $(\Delta W_t \sim N(0, I\Delta t))$ rastgele üretilen Brown vektörünü gösterir. Son olarak Δt kesikli zaman aralığını temsil eder.

İLERİ MODELLER

Gelişen teknolojinin yardımıyla gen ağlarının modellenmesi, analizi, özellikle son yılların en ilgi çekici konularından biri olmuştur. Dolayısıyla yukarıda bahsedilen modeller dışında, özellikle yüksek boyutlu gen ağlarının modellenmesinde birçok yeni model de önerilmiştir. Bunların bir kısmı Gaussian grafiksel denkleme alternatif olarak önerilen *parametrik olmayan olasılıksal grafik modelleri* ve *parametrik olmayan, genelleştirilmiş toplamsal modeller* altında bulunan modellerdir. MARS bunlara örnek olarak verilebilir.[16,17] Parametrik model alternatifleri içinde ise en öne çıkanları Kopula GGM modelidir.[18,19] Bu model, kısaca, GGM ifadesindeki çok değişkenli normal dağılım fonksiyonunun, *Gaussian kopula* yardımıyla yazılmasından türetilmiştir. Diğer yandan verinin özellikle zaman bağlı olması durumunda *durum uzay modelleri* ve *zaman serisi GGM* ifadeleri geliştirilmiştir.[20,21] Geliştirilen bu modeller için farklı ve özellikle hesaplama hızı yüksek model parametre tahmin metotları ise alanda modelleme ile beraber çalışılan bir diğer konudur. Bahsedilen modellerin tahmininde *optimizasyona dayalı yöntemlerin* yanı sıra, *cezalandırılmış olabilirlik*, *ters atlamalı markov zinciri Monte Carlo* ve *doğum-ve-ölüm algoritması* gibi parametrik tahmin yöntemleri ve *eşik meyil iniş algoritması* gibi parametrik olmayan yöntemler sıklıkla kullanılmaktadır.[10,18,19,22] Fakat modelleme ve parametre tahmin yöntemleri halen bu konudaki araştırma başlıklarıdır. Bu nedenle her geçen gün yeni ifadelerin tanımlanması ve uygun algoritmaların geliştirilmesi konularında, çalışmalar yapılmaya devam edilmektedir.

SONUÇ

Bu çalışmada protein etkileşim sistemlerinin farklı veri türleriyle, dolayısıyla farklı varsayımlar altında nasıl modellenebileceğini tanıttık. Bu modellerin her biri bir proteinin aktif olup olmama durumunu, çalışılan sistemin rassal davranıp davranmama haline göre değerlendirmektedir. Bu tür bilgiler özellikle kanser, kalp hastalıkları gibi hayati etkileri olan sağlık problemlerine karşı tedavi yöntemlerinin geliştirilmesinde büyük önem taşımaktadır. Bahsedilen sistemlerin çözümlenmesinde en az modelleme kadar önemli olan bir diğer konu ise model parametrelerinin tahminidir. Tahmin edilen parametre sayısının fazla olması, ya Bayesci metodlar gibi hesaplama süresi uzun, fakat doğruluğu yüksek olan metodlarla ya da yaklaşık metodlar gibi hesaplama süresi kısa, ancak nispeten doğruluğu düşük olan yöntemlerle çö-

zümlerin bulunmasını zorunlu kılmaktadır. Diğer yandan karmaşık sistemler için model parametrelerinin tahmini devam eden araştırma alanlarından biridir.

Çıkar Çatışması

Yazarlar herhangi bir çıkar çatışması veya finansal destek bildirmemiştir.

Yazar Katkıları

Bu çalışmanın fikir ve denetlemesi Vilda Purutçuoğlu tarafından yapılırken, analiz, yazım ve yorumlama Vilda Purutçuoğlu ve Ezgi Ayyıldız tarafından yapılmıştır.

KAYNAKLAR

1. Bolouri H. Computational Modeling of Gene Regulatory Networks: A Primer. 1st ed. London: Imperial College Press; 2008. p.326.
2. Bower J, Bolouri H. Computational Modeling of Genetic and Biochemical Networks. 1st ed. London: MIT Press; 2001. p.336.
3. Purutçuoğlu V, Ayyıldız E. [Statistics in the field of bioinformatics]. Biyoinformatik Alanında İstatistik. 1. Baskı. Ankara: Nobel Akademik Yayıncılık; 2014. p.258.
4. Wilkinson DJ. Stochastic Modelling for Systems Biology. 2nd ed. Boca Raton, Florida: Chapman & Hall/CRC Press; 2011. p.363.
5. de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol 2002;9(1):67-103.
6. Shmulevich I, Dougherty ER. Probabilistic Boolean Networks: The Modeling and Control of Gene Regulatory Networks. 1st ed. Philadelphia: Society for Industrial and Applied Mathematics; 2010. p.267.
7. Whittaker J. Graphical Models in Applied Multivariate Statistics. 1st ed. New York: Wiley; 1990. p.448.
8. Defterli Ö, Purutçuoğlu V, Weber GW. Advanced mathematical and statistical tools in the dynamic modeling and simulation of gene-enviroment networks. In: Pinto AA, Zilberman D, eds. Modeling, Dynamics, Optimization and Bioeconomics. 1st ed. Heidelberg: Springer; 2014. p.237-57.
9. Li H. Statistical methods for inference of genetic networks and regulatory modules. In: Dehmer M, Streib FE, eds. Analysis of Microarray Data: A Network-Base Approach. 1st ed. Weinheim: Wiley-VCH; 2008. p.143-67.
10. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 2008;9(3):432- 41.
11. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 2005;4(1):Article 32:1-30.
12. Kampen NGV. Stochastic Process in Physics and Chemistry. 3rd ed. Amsterdam: North Holland; 2007. p.464.
13. Lawrence ND, Girolami M, Rattray M, Sanguinetti G. Learning and Inference in Computational Systems Biology. 1st ed. Cambridge: MIT Press; 2010. p.362.
14. Purutçuoğlu V, Wit E. Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters. Bayesian Analysis 2008;3(4):851-86.
15. Purutçuoğlu V. Inference of the stochastic MAPK pathway by modified diffusion bridge method. Cent Eur J Oper Res 2013;21(2):415-29.
16. Friedman N. Inferring cellular networks using probabilistic graphical models. Science 2004;303(5659):799-805.
17. Ayyıldız E, Ağraz M, Purutçuoğlu V. MARS as the alternative approach of Gaussian graphical model for biochemical networks. J Appl Stat 2016;1-9.
18. Dobra A, Lenkoski A. Copula Gaussian graphical models and their application to modeling functional disability data. Ann Appl Stat 2011;5(2A):969-93.
19. Mohammadi A, Wit EC. Bayesian structure learning in sparse Gaussian graphical models. Bayesian Analysis 2015;10(1):109-38.
20. Haavisto O, Hyötyniemi H, Roos C. State space modeling of yeast gene expression dynamics. J Bioinform Comput Biol 2007;5(1):31-46.
21. Abegaz F, Wit E. Sparse time series chain graphical models for reconstructing genetic networks. Biostatistics 2013;14(3):586-99.
22. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. Ann Stat 2006;34(3):1436-62.