Matematiksel Evrim Yaz Okulu, 9-11 Eylül 2013

Ders Notları

Biyoloji ve Evrimde Ağ yapıları

Ayşe Erzan

Fizik Bölümü, Fen-Edebiyat Fakültesi,

İstanbul Teknik Üniversitesi, İstanbul 34 469

Özet

Sistem biyolojisi, karmaşık biyolojik sistemleri etkileşen alt-sistemler olarak tasvir eder. Bu etkileşim ağlarının topolojileri, bu ağlar üzerinde yer alan bir dizi olguyu anlamak için kilit önemi haiz ipuçlar sunar. Bu notlarda ağ yapılarının özellikleri, bazı biyolojik ağ yapıları, evrilme yöntemleri ve karmaşıklık kavramı irdelenecektir.

I. BİYOLOJİK AĞ YAPILARINA ÖRNEKLER

Biyolojik sistemlerde ağ yapıları, fizyolojik ağ yapıları (sinir ağları, dolaşım ağları gibi) ya da daha soyut etkileşim ağları biçiminde (gen regülasyonu ağları, protein-protein etkileşim ağları, metabolik ağlar gibi), ekolojik ağ yapıları (beslenme şebekeleri) veya ağ yapılarının birbirleri ile olan ilişkilerinden doğan "meta-ağlar" biçiminde karşımıza çıkmaktadırlar.

Meta-ağlara bir örnek, birbirlerinden tek bir değişinim ile farklılaşan gen regülasyonu ağları olabilir. Birbirinden tek değişinimla ayrılan bu ağların birinden diğerine geçildiğinde eğer fenotipte bir değişiklik olmuyorsa, tüm bu meta-ağyapısına "nötr ağ yapısı" tabir edilmektedir. Bu nötr-ağlar, evrimleşebilmenin temelinde yatan bir oldugurlar. Zira bu sayede organizma herhangi yeti kaybına uğramadan çeşitli genotipleri denemekte ve sonunda, küçük bir sıçrayışla başlangıç noktasından tamamen farklı bir fenotipe ulaşıp morfolojik ya da işlevsel olarak tamamen farklı bir yapıya kavuşabilmektedir.

II. AĞ YAPILARI İÇİN TANIMLAR

Bir "Ağ yapısı" (*network*) ya da "çizge" (*graph*) şu ögelerden oluşuyor: düğüm noktaları (*nodes, vertices*) ve kenarlar ya da bağlar (*edges*). Genellikle düğüm noktaları birbirleri ile bir biçimde ilişkilendirilen nesneleri gösteriyorlar. Eğer bir çift düğüm noktası arasında bir bağ ya da "kenar" varsa, bu aralarında bir ilişki olduğunu gösteriyor. Kenarlar sadece mevcut ya da namevcut olabilirler; ya da ağırlıklara sahip olabilirler.

Bir çizgenin yapısını özet olarak verebilmek için ya birbirlerine bağlı düğüm noktalarının listesini verebiliriz (veri tabanlarında yapılan genellikle bu oluyor), ya da "komşuluk matrisi"ni yazabiliriz.

Eğer bir çizgenin tüm düğüm noktalarını $i = 1, \ldots, N$ ile etiketlersek, o vakit aşağıdaki gibi bir liste hangi düğüm noktasının hangisine bağlı olduğunu bize gösterir. örneğin, toplam düğüm sayısı $N = 5$ olan bir çizge için, eğer bağlar simetrik bir ilişkiyi gösteriyorlarsa (yani yönlendirilmiş değillerse), ağ yapısını anlayabilmemiz için şöyle bir liste yeterlidir:

$$(1,2); (1,3); (2,3); (2,5); (3,5); (4,5) \ . \tag{1}$$

Bu çizgenin üzerinde kimin kime bağlı olduğunu aynı zamanda aşağıdaki gibi bir komşuluk matrisi $\mathbf{A}$ ile de göstermemiz mümkündür. Burada sütünlar 1'den 5'e kadar, satırla da

yine 1'den 5'e kadar sıralanmışlardır. çizgenin her bir düğüm noktasını 1'den beşe kadar etiketleyebileceğimiz gibi, her bir kenarı da, hangi iki düğüm noktasını birleştirdiğine göre, iki etiketle (bir satır ve bir sütün etiketi ile) etiketleyebiliriz. Böylece aşağıdaki matrisi elde ederiz.

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} & a_{1,5} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} & a_{2,5} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & a_{3,5} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} & a_{4,5} \\ a_{5,1} & a_{5,2} & a_{5,3} & a_{5,4} & a_{5,5} \end{pmatrix} \tag{2}$$

Eğer (i, j) çifti yukarıdaki listemizde (Denklem 1) mevcutsa, $a_{i,j} = 1$, eğer mevcut değilse $a_{i,j} = 0$ diyeceğiz. Hemen anlaşılabileceği gibi, simetik etkileşimler için, $a_{i,j} = a_{j,i}$ olacaktır.

Şimdi komşuluk listesi Denklem (2)'de verilmiş olan bir çizgenin, komşuluk matrisini yazalım.

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \tag{3}$$

Sıfır olan elemanlar, o elemanın satır ve sütün etiketini taşıyan düğüm noktalarının birbirlerine bağlı olmadıklarını, tersine, 1 değerini taşıyan elemanlar ise, birbirlerine bağlı olduklarını göstermektedir.

Etkileşimler yönelimli olduğu takdirde, komşuluk matrisi simetrik olmayacaktır. Örneğin buradaki düğüm noktaları genleri gösterdiği, ve listedeki ilk genin ikinci geni düzenlediği bir durumda, vereceğniz liste şöyle olabilir:

$$(1, 2); (1, 3); (2, 3); (3; 2); (3, 5); (4, 5); (5, 1); (5; 2) \quad . \tag{4}$$

Alıştırma: Bu listeye bakarak, simetrik olmayan etkileşim matrisinizin elemanlarını oluşturun. Listede olmayan çiftlere karşı gelen elemanları sıfır yapmayı unutmayın. Şimdi bir kağıda beş düğüm noktasi çizin ve bu listede verilen çiftler arasında etkileşimin yönünü de gösteren oklar koyarak çizgenizi elde edin. Göreceksiniz ki bir çift düğüm noktası arasındaki etkileşim iki yönlü de olabiliyor.

Genelde, etkileşimler farklı ağırlıklar da taşıyabilirler. Bu durumda konşuluk matrisinin

elemanları sadece 1, 0 değerlerini değil, herhangi gerçel "ağırlık" değerini taşıyabilirler $a_{i,j} = w_{i,j}$.

III. AĞ YAPILARININ ÖZELLIKLERI

Ağ yapılarının özellikleri arasında en önemlilerinden biri, her bir düğüm noktasının kaç kenara sahip olduğudur. Buna bir düğüm noktasının sahip olduğu **derece** (*degree*) denir. Eğer etkileşimler simetrik ise, çizge üzerinde i etiketli düğüm noktasının derecesi

$$d_i = \sum_j a_{i,j} = \sum_j a_{j,i} \tag{5}$$

olarak hesaplanır.

Alıştırma: Yukarıdaki simetrik örnekte, dereceler $d_1 = 2$, $d_2 = d_3 = d_5 = 3$ ve $d_4 = 1$ olarak bulunur. Gösteriniz!

Eğer etkileşimler yönelimli ise, kendine yönelimli derece şöyle hesaplanır:

$$d_i^{\text{in}} = \sum_j a_{j,i} \ . \tag{6}$$

Alıştırma: Eğer etkileşimler iki yönlü yani simetrik ise, bu toplam dereceye eşit olacaktır. Gösteriniz!

Dışarı yönelimli derece

$$d_i^{\text{out}} = \sum_j a_{i,j} \ , \tag{7}$$

ve toplam derece

$$d_i^{\text{toplam}} = \sum_{i,j} a_{j,i} \ . \tag{8}$$

olur. **Alıştırma:** Bu dereceleri, Denklem 4'te verilen listeye göre, her bir düğüm noktası için bulunuz.

Bir çizge için **derece dağılımı** (*degree distibution*) onun önemli topolojik özelliklerinin başında gelir. Bundan kasıt, her bir farklı derece değerine sahip düğüm noktalarından kaçar tane olduğudur. Eğer her bir d değeri için gözlenme sıklığını toplam düğüm noktası sayısına, yani N'ye bölersek, faklı derece değerlerinin olasılık dağılımını buluruz. Buna $P(d)$ diyelim. **Alıştırma:** Yukarıda örnek olarak verdiğimiz simetrik çizge için $P(d)$ dağılımını bulun. (Çözüm $P(1) = 0.2$, $P(2) = 0.2$, $P(3) = 0.6$.)

Konumuza ilişkin ağ yapılarının önemli karakteristiklerinden biri de kümelenme katsayısıdır (*clustering coefficient*). Bu büyüklük, herhangi düğüm noktasının komşularının (yani en az bir kenarla bağlı olduğu noktaların) kendi aralarında ne oranda bağlı olduklarını ölçer. Simetrik bir çizge için, herhangi i'nci düğüm noktasının kümelenme katsayısını C_i olarak tanımlarsak,

$$C_i = \frac{2\sum_{j,k} a_{i,j}a_{j,k}a_{k,i}}{d_i(d_i-1)} \tag{9}$$

olur. Bunu şöyle de hesaplayabiliriz : bir noktanın komşu sayısı d_i olduğuna göre, komşuları arasında en fazla $d_i(d_i-1)/2$ tane simetrik kenar olabilir. Bu sayıya, komşular arasında gerçekten var olan kenar sayısını oranlarsak, C_i bulunur. Tüm çizge için kümelenme katsayısı ise, bu büyüklüğün tüm düğüm noktaları üzerinden ortalamasıdır. Bunu sembolik olarak şu biçimde ifade edebiliriz $C = \sum_i C_i/N$.

Farklı derece dağılımları ve dayanıklılık

Rasgele ağlar (ya da Erdös-Renyi ağ yapıları), çizge üzerindeki her bir düğüm noktası çiftini birleştirecek bir kenarın p olasılığı ile atanması (ya da $1-p$ olasılığı ile atanmaması) yoluyla oluşurlar. Bu ağ yapısının yegane parametresi p dir. Her düğüm noktasının ortalama derecesi pN olarak bulunur. (Eğer düğüm noktalarının kendi kendileri ile etkileşimlerine izin yoksa, bu değer $p(N-1)$ olur. Gösteriniz! Eğer $N \gg 1$ ise, N'ye kıyasla 1 ihmal edilebilir.) Bu ağ yapıları için derece dağılımı ise bir binom dağılımı olacaktır:

$$P(d) = \frac{N!}{d!(N-d)!} \quad . \tag{10}$$

Bağlanma olasığı p eğer çok küçükse ($p \ll 1$), bu binom dağılımı bir Poisson dağılımı tarafından temsil edilir. Eğer ortalama dereceye $pN = a$ dersek, Poisson dağılımı

$$P(d) = \frac{e^{-a}a^d}{d!} \tag{11}$$

ile verilir.

Rasgele ağların bir şaşırtıcı özelliği bunların üzerindeki düğüm noktaları arasında her vakit oldukça kısa yolaklar (hep kenarları takibederek gidilebilecek kestirme yollar) bulunabilmesidir. Öyle ki, tipik olarak, bir rasgele ağ üzerindeki düğüm noktaları arasında en kestirme yolaklar seçildiğinde, bu yolakların maksimim uzunluğu N'ye değil de, ancak - çok daha yavaş artan - $\log N$'ye bağlı olarak büyür. Buna "küçük dünya" özelliği tabir edilir.

Rasgele ağlar, eğer kenarları rasgele kopartılırsa, kolayca parçalara ayrılırlar.

IV. KARMAŞIK AĞ YAPILARI

Şimdi farklı bir derece dağılımına sahip olan **ölçeksiz ağ yapıları**na bakalım.

Derece dağılımı $P(d)$, eğer $d^{-\gamma}$'ya orantılı ise, bir "kuvvet yasası"na tabi olduğu söylenir.

$$P(d) \propto d^{-\gamma} \quad , \tag{12}$$

ve ayrıca $\gamma < 3$ ise, bu dağılımın bir ortalama değerini hesapayabilsek de, çok büyük bir ağ için varyansı o kadar büyük olacaktır ki o ortalamanın anlamı olmayacaktır. Bu tür ağlara, onları karakterize eden bir ortalama değer olmadığı için ölçeksiz ağlar denir. Ölçeksiz ağlar, **karmaşık** ağları modellemede önemli bir yere sahiptirler. Bazı biyolojik ağ yapıları ve teknolojik ağların çoğu bu sınıfa girmektedirler. Bu tür ağlarin düğüm sayısı büyüdükçe, daha daha büyük dereceli düğüm noktaları ortaya çıkmaktadır.

Ölçeksiz (*scale free*) ağlar, rasgele saldırılara karşı çok dayanıklıdırlar.(Albert et al. 2000) Bunun nedeni, yeterince büyük ağlarda, neredeyse ağın tüm düğüm noktalarına bağlı olan çok yüksek sayıda kenara sahip düğüm noktalarının bulunmasıdır. Ancak, saldırı hedefli ise, yani bu yüksek kenarlı noktalardan biri kopartılırsa, o vakit ağ büyük zarar görecektir.

Biyolojik ağ yapılarının pek çoğu, iddia edildiği üzere (Barabasi et al. 2004) ölçeksiz değildir. Balcan et al. (2007)'de gösterildiği gibi, örneğin bira mayasının (*S. cereviciae*) gen regülasyonu ağının derece dağılımı ölçeksiz bir ağ yapısından da daha karmaşık özellikler taşımaktadır.

K-katman ayrıştırması (*K-core decomposition*), bir regülasyon ya da iletişim ağında, en önemli yönetici rollere sahip katmanları birbirinden ayırmak için kullanılan bir yöntemdir. Bu yöntemle, ilk önce sadece bir kenarla bağlı düğüm noktaları bir 1. kümeye konulur. Sonra bunların bağları kesilir ve geri kalan çizge üzerinde tek kenarla bağlı noktalara bakılır ve onlar da 1. kümeye dahil edilir. Bu sayma ve koparma işlemi, tek kenarla bağlı hiç bir düğüm noktası kalmayana kadar tekrar edilir. Şimdi sıra iki kenarla bağlı noktaları 2. kümeye koymaya gelir. Bunlar kopartıldıktan sonra, iki kenarla kalmış düğüm noktalarına aynı işlem uygulanır vs.; ortada başka nokta kalmadığında en yüksek "katman" sayısına ulaşılmış olur.

Bira mayasının gen regülasyonu ağı için (Balcan et al. 2007) K-katman ayrıştırması yöntemi ile hem gerçek hem de model ağda dokuz katman bulunmuştur. Benzer düğüm noktası sayısı ve etkileşim (kenar) yoğunluğuna sahip ölçeksiz ağlarda, ancak iki ya da üç katman bulunması, bunların gerçek karmaşık biyolojik ağları modellemede yetersiz olduklarının başka bir göstergesidir.

Ağ yapılarının kopyalama-değişinim yoluya evrimleşmesi

Biyolojik ağların rasgele süreçlerle modellenmesinde hesaba katılması gereken bir husus, genlerin ve o genin aktivasyonunda etkin olan protein-kodlamayan bölgelerin, bir genden diğerine benzerlik gösterebileceği, aralarında ilintililikler (*correlations*) olabileceğidir. (Wagner 1994)

Wagner'in "kopyalanma ve değişinim" (*duplication and divergence*) modelinde, protein-protein etkileşim modelleri ele alınmıştır. Proteomda, bir proteini kodlayan gen kopyalandığı zama, o proteinin tüm eski etkileşimlerini koruyacağı, ancak daha sonra gerçekleşecek değişinimlardan dolayı eski etkileşimlerinden bazılarını kaybedip, yeni etkileşimler edinebileceği kabul edilmektedir. Bu tür bir evrim senaryosu ile gerçekçi protein etkileşim ağları elde edilebilmektedir. (Wagner 1994 ve buna referans veren çok sayıda makale)

V. KARMAŞIKLIK, ENFORMASYON

Genel olarak karmaşıklığın niceliklendirilmesi ilginç bir problemdir. Biyolojik sistemleri genelde karmaşık sistemler olarak algıladığımıza göre, karmaşıklığın "ölçeksizlik"ten farklı bir ölçüsünü tartışmakta yarar vardır. Zira biyolojik yapılar, ölçeğe sıkı sıkıya bağlı özellikler de göstermektedirler.

Karmaşıklığın bir ölçüsü enformasyon içeriği olabilir. Shannon (1948) entropiyi enformasyon içeriğinin bir ölçütü olarak önermiştir. Entropi aslında bir sistemin içinde bulunduğu durumu belirtmek için vermemiz gereken bilgi mıktarını ölçmektedir. Bu anlamda, bir sistemin gösterebileceği çeşitliliğin ölçüsüdür.

Bir sistem M tane farklı durumda bulunabiliyor olsun. Bu durumları m indisi ile gösterelim ve her farklı durumun gözlenme olasılığı p_m olsun. Sistemin entropisi, ya da Shannon enformasyonu,

$$I = -\sum_{m}^{M} p_m \log p_m \tag{13}$$

olacaktır. Sadece bir durumun olasılığı sıfırdan farklı, diğerleri ise sıfırsalar, entropi sıfır olacaktir! Tüm durumların eşit olasılığa sahip olduğu durumda $I = \log M$ bulunur. Bu sistem için entropinin alabileceği en yüksek değer budur.(Gösterin!) En rasgele durumun entropisini I_0 ile gösterelim.

Sistemi biz hazırlamış olmasak bile, onun hangi durumda olduğu konusunda ne kadar bilgi sahibi olduğumuzu ölçmek istiyorsak, o vakit başvuracağımız büyüklük, entropinin, olabileği

en yüksek değerden (en rasgele sisteminkinden) ne kadar farklılaştığıdır. Buna "göreli entropi" ya da göreli enformasyon diyelim. Göreli enformasyonu

$$I_R \equiv I_0 - I \tag{14}$$

olarak tanımlayalım. Birçok durumda bu büyüklüğün daha uygun bir enformasyon ölçüsü olacağı açıktır.

Eğer N uzunluğunda Boolcu (sadece 0 ve 1'lerden oluşan) bir dizinin kompozisyonunu kesin olarak *biliyorsak*, o vakit o dizi hakkında sahip olduğumuz bilgiyi "göreli enormasyon" cinsinden hesaplayalım. Bu dizinin içinde bulunabileceği durumların sayısı $M = 2^N$dir. Yani $I_0 = N \log 2$. Hangi durumda bulunduğunu kesin olarak bildiğimiz sistem için ise durumlardan sadece birinin olasılığının 1, diğer durumların olasılığını 0 olduğunu söylüyoruz - öyleyse bu sistemin entropisi sıfırdır. Göreli enformasyon ise $I_R = N \log 2 - 0$ olacaktır. Bu bizim sistem hakkında bilgimizin ölçüsüdür, ve bu bilginin $I_R/\log 2$ "bit"ten oluştuğunu da söyleyebiliriz!

Tanımlamış oldğumuz göreli enformasyon yerine, "istatistiksel karmaşıklık" kavramı da kullanılabilir. İstatistiksel karmaşıklık da, olabilecek en rasgele dağılımla elimizdeki sistemin olasılık dağılımı arasındaki uzaklığı ölçmeye dayalıdır. Bu kez, bu uzaklığın ölçüsü "Jensen-Shannon diverjansı"dır.

Jensen-Shannon diverjansını tanımlamak için, bir referans dağılımı (elimizdekinden daha rasgele olduğunu bildiğimiz bir dağılım; buna Q diyelim) ile kendi sistemimizin sahip olduğu olasılık dağılımını (P) alalım. Her iki dağılımda da M tane durum olsun. Eğer $P \equiv \{p_1^{(P)}, \ldots, p_M^{(P)}\}$ ve $Q \equiv \{p_1^{(Q)}, \ldots, p_M^{(Q)}\}$ ise, bu iki dağılımın ortalamasını aynen şöyle tanımlayabiliriz:

$$(Q + P)/2 \equiv \sum_m^M [p_m^{(P)} + p_m^{(Q)}]/2 \ . \tag{15}$$

Jensen-Shannon diverjansı (farklılaşması)

$$J(P, Q) = I[(Q + P)/2] - [I(Q) + I(P)]/2 \tag{16}$$

olarak tanımlanır. Hemen görüleceği üzere, $P = Q$ ise bu diverjans - yani fark - sıfır olacaktır. Seçtiğimiz referans dağılımı Q ile kendi sistemimizin dağılımı P aracılığı ile tanımlayacağımız "istatistiksel karmaşıklık ise,

$$C(P, Q) \equiv J(P, Q)I(P)/I(P)_{\max} J_{\max} \tag{17}$$

olarak tanımlanır. Bu karmaşıklık tanımının avantajı, hem sistemin sadece tek bir durumda bulunduğu, hem de olabileceği kadar rasgele olduğu durumlarda değerinin sıfır olmasıdır. İstatistiksel karmaşıklık, sistemde hem çeşitlilik hem de yüksek derecede ilintililik - yani bir kısmından diğerinin bir ölçüde kestirilebilmesi imkanı olduğu durumda, $I(P)$'nin fonksiyonu olarak en büyük değerini alacaktır.

Bibliografya

Albert R, Jeong H and Barabasi AL (2000) Nature 406: 378-382

Albert R, Barabasi, AL (2001) Statistical mechanics of complex networks. Rev Mod Phys 74: 47-97.

Dorogovstsev SN, Mendes JFF (2002) Evolution of networks. Adv Phys 51: 1079-1187.

Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nature Reviews-Genetics 5: 101-113

Balcan D, Kabakçıoğlu A, Mungan M and Erzan A 2007 The Information Coded in the Yeast Response Elements Accounts for Most of the Topological Properties of Its Transcriptional Regulation Network; PLoS ONE 2 e501

Colizza V, Flammini A, Maritan A, Vespignani A (2005) Characterization and modeling of protein-protein interaction networks. Physica A 352: 1-27.

Shannon, C.E. (1948), A Mathematical Theory of Communication; Bell System Technical Journal 27 379423 and 623656

Sole RV, Pastor-Satorras R (2002) Complex Networks in Genomics and Proteomics. In: Bornholdt S, Schuster HG, eds (2002) Handbook of Graphs and Networks. Berlin: Wiley-VCH Verla

Wagner A 1994 Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization; Proc. Natl. Acad. Sci. USA 91 4387-4391