itüdergisi/*c*
fen bilimleri
Cilt:5, Sayı:1, 41-48
Kasım 2007

İçerik-temelli ağlar üzerinde analitik hesaplar

Duygu BALCAN[*]**, Ayşe ERZAN**

İTÜ Fen Bilimleri Enstitüsü, Fizik Mühendisliği Programı, 34469, Ayazağa, İstanbul

Özet

Bu makalede içerik-temelli ağlar üzerinde, ağın topolojik özelliklerini belirlemek için, ortalama-alan yaklaşımlarıyla yapılan analitik hesapların güvenilirliği tartışılacaktır. İçerik-temelli ağları, "tanıma ve bağlanma" mekanizmalarının belirleyici olduğu kontrol çizgelerinin topolojik özelliklerini tasvir etmek için önermiştik. Birçok karmaşık ağ yapısının bu tür enformasyon paylaşımına dayalı bir prensibe göre inşa edildiğini söyleyebiliriz. Örneğin gen ifadesinin düzenlemesinde, anahtar/kilit olarak niteleyebileceğimiz elemanların özelleşmiş etkileşimleri söz konusudur. Bu sebeple modelimizin biyolojik çizgeler de dahil olmak üzere, birçok gerçek ağ yapısının tasviri için uygun olduğunu düşünüyoruz. İçerik-temelli ağımızda, ağın düğümlerini bir ya da birden fazla rastgele dizi ile eşleştirip, düğümler arasındaki etkileşimleri onlara atanan dizilerin birbirleri içinde tekrarlanma koşulu altında inşa ediyoruz. Böylece, bu dizilerin uzunlukları ve içerikleri, ortaya çıkacak olan çizgenin tüm topolojik özelliklerini belirlemektedir. Düğüm çiftleri arasındaki bağlanma olasılıklarının hesabında yapılan ortalama-alan yaklaşımlarının ise, dizilerin uzunluk dağılımlarına bağlı olarak, varılan sonuçlarda ağın gerçek özelliklerinden önemli farklılaşmalara yol açabileceği görülüyor. Bu yaklaşımlarda, dizilerin farklı enformasyon içerikleri ihmal edilmekte ve olasılıklar sadece dizilerin uzunlukları cinsinden elde edilmektedir. Halbuki her sonlu dizi için, dizinin içerdiği farklı sembol sayısı ek bir enformasyon içermektedir. Burada sergilemeye çalışacağımız, kabalaştırılmış ortalama-alan türünden yaklaşımların, belli ekstrem durumlarda, tasvir etmeyi amaçladıkları ağın özelliklerinden uzak sonuçlar verebileceğidir. Ancak gerçek biyolojik ağ yapılarının modellenmesinde karşımıza çıkan uzunluk dağılımlarında ortaya çıkan hatalar hiçbir zaman burada sergileyeceğimiz örneklerde olduğu kadar büyük olmamış, bilakis ortalama-alan yaklaşımı simülasyon sonuçlarına oldukça yakın sonuçlar vermiştir.

Anahtar Kelimeler: *Karmaşık ağ yapıları, içerik-temelli ağlar, ortalama-alan yaklaşımı.*

[*]Yazışmaların yapılacağı yazar: Duygu BALCAN. balcand@itu.edu.tr; Tel: (212) 285 7243.
Bu makale, birinci yazar tarafından İTÜ Fen Bilimleri Enstitüsü, Fizik Mühendisliği Programında tamamlanmış olan "İçerik-temelli ağların özellikleri" adlı doktora tezinden hazırlanmıştır. Makale metni 02.03.2007 tarihinde dergiye ulaşmış, 08.03.2007 tarihinde basım kararı alınmıştır. Makale ile ilgili tartışmalar 31.05.2008 tarihine kadar dergiye gönderilmelidir.

Analytical calculations on content-based networks

Extended abstract

Content-based networks have been proposed (Balcan and Erzan, 2004; Mungan et al., 2005) to model the topological properties of complex networks built on the principle of information sharing, where the interactions between system components assume the simultaneous fulfillment of a series of constraints (Mezard et al., 2002). In content-based networks, the constraint-satisfaction problem is realized by means of a sequence-matching rule between sequences associated with the nodes of a network.

In the case of transcriptional gene regulation, the transcription factors recognize special subsequences of DNA and bind them. This is one instance of constraint-satisfaction, which can be realized with a sequence-matching rule between two different classes of sequences (Balcan et al., 2006). Another example is the so called the RNA interference (Balcan and Erzan, 2004), where sequence-specific gene silencing occurs at the level of post-transcriptional gene regulation.

In our content-based networks, n *linear codes are associated with each node of the network. For* n=2*, one of the sequences associated with the node represents the key-sequence through which the node recognizes other nodes, whereas the second sequence represents the lock-sequence through which the same node is recognized. An interaction between a pair of nodes is established if the key-sequence associated with the first node is repeated as an uninterrupted subsequence in the lock-sequence associated with the second node. Thus, the length distributions of these sequences are the most important parameters determining the topological properties of the content-based networks.*

In this article we will discuss the validity of analytical calculations performed on the topological properties of content-based networks in the mean-field approximation (Balcan and Erzan, 2007), by means of two examples. In this mean field approach (Mungan et al., 2005) the pair-wise connectivity probabilities are only functions of the respective lengths of the sequences which must satisfy an inclusion requirement, and of the size r *of the alphabet from which the symbols are drawn. This approximation ignores the correlations between the overlapping subsequences within a sequence. Moreover the fluctuations in the information content of finite sequences are neglected. In Balcan and Erzan (2007), the correlations between the edges co-incident on the same node were also ignored.*

In the first example, the key- sequences of unit length (thus, they consist of single letters) are searched in lock-sequences of an arbitrary fixed length. Via this simple example it is possible to show that the probability that lock-sequences will be recognized by a key-sequence depends not only on the length of the lock-sequence but also on the number of distinct subsequences embedded in it. At this point the coarse grained approximation neglecting the fluctuations in the information content of the finite lock sequences about their mean information content, misses the behavior of the in-degree distribution. This error is in fact identical to neglecting the correlations between edges incident upon a given node.

In the second example, the lengths of the key sequences are fixed at an arbitrary value l*, and the lock-sequences are chosen to be of length* k=l+1*, one character longer than the key-sequences. In this example, it is clear that the correlations between the two subsequences of length* l *cannot be neglected. It has already been shown (Guibas and Odlyzko, 1981; Mungan et al., 2005; Mungan, 2007; Bilge et al., 2004) that the connection probability of a key-sequence depends on the "shift-match number" which measures the auto-correlations within a sequence, in other words, the degree to which successive subsequences are correlated with each other. We show here by an explicit and rather transparent calculation that, neglecting this correlation yields out- and in-degree distributions that are totally in error.*

The mean-field approximations used in the calculation of the topological properties of the double-string model (Balcan and Erzan, 2007) yield results that are in good agreement with the simulations, since i) the lengths k *of the lock sequences far exceed* r*, ii) the number of distinct substrings contained in any given lock string is large (* $k-l >> r^l$ *) and iii) the fine structure of the topological properties are determined by the fact that there is a distribution of lock- and key-string lengths.*

Keywords: *Complex networks, content-based networks, mean-field approach.*

Giriş

İçerik-temelli ağlar (Balcan ve Erzan, 2004; Mungan vd., 2005; Balcan vd., 2006) tanınan/tanıyan veya anahtar/kilit olarak niteleyebileceğimiz türden elemanların birbirleriyle etkileşimlerini ve bu etkileşimler sonunda ortaya çıkan ağların topolojik özelliklerini tasvir etmek amacıyla önerilmiştir. Bu tür sistemler biyolojide yaygın olarak yer almaktadır. RNA girişimi (RNA interference) -RNAi (Hannon, 2002) olarak isimlendirilen transkripsiyon sonrası gen susturulması bunun en çarpıcı örneklerinden biridir. Bir başka örnek gen ifadesinin ilk kontrol noktası olan transkripsiyonel gen düzenlemesinden (Alberts vd., 2002; Wray vd., 2003) gelmektedir. Burada transkripsiyon faktörü olarak isimlendirilen proteinler DNA üzerinde regülasyon (promoter) bölgesi diye bilinen özel dizileri tanıyıp bağlanmakta ve gen ifadesinin düzenlenmesine katkıda bulunmaktadırlar.

Bugüne dek en ayrıntılı biçimde çalışılmış ve hakkında en çok veri bulunan Maya'nın (*Saccharomyces cerevisiae*) gen düzenleme ağını modellemek için önerdiğimiz (Balcan vd., 2006) çift-dizili içerik-temelli ağ modelinde, çizgedeki her düğüm, biri anahtar diğeri kilit vazifesi gören, ortak bir alfabeden gelişigüzel türetilen iki dizi ile eşleştirilmektedir. Eğer bir anahtar dizi bir kilit dizinin içinde en az bir kere tekrarlanırsa, anahtar dizinin eşleştirildiği düğümden kilit dizinin eşleştirildiği düğüme doğru bir kenar yerleştirilerek ağ oluşturulur. Böylece, bu dizilerin birbirleri içinde tekrarlanma olasılıkları iki düğüm arasındaki bağın (kenarın) oluşma olasılığına karşılık gelmektedir.

İçerik-temelli model ağlar üzerinde yapılacak analitik hesaplar düğümler arasındaki bağlanma olasılıklarının, yani anahtar-dizilerin kilit-diziler içinde tekrarlanma olasılıklarının, hesabıyla başlamak zorundadır. Daha önce yayınlanmak üzere sunulan bir çalışmada (Balcan ve Erzan, 2007) içerik-temelli ağların (Balcan ve Erzan, 2004; Mungan vd., 2005; Balcan vd., 2006) topolojik özelliklerini hesaplayabilmek için bir ortalama-alan teorisi yaklaştırımı önerilmiştir. Bu yaklaştırımda, dizilerin içerikleri tamamen ihmal edilmiş ve birbirleriyle örtüşen alt-dizilerin bağımsız olduğu varsayılmış olup, eşit uzunluklu tüm dizilerin birbirlerine özdeşliği varsayılmıştır; bunların başka diziler içinde bulunma olasılıkları ya da başka dizileri içlerinde bulundurma olasılıklarının birbiriyle aynı olduğu ve bunların sadece karşılaştırılan dizilerin uzunluklarına bağlı olarak ifade edilebileceği varsayılmıştır. Buna göre uzunluğu l olan rastgele seçilmiş bir dizinin uzunluğu k olan yine gelişigüzel seçilen bir dizide en az bir kere tekrarlanma olasılığı $p(l,k)$,

$$p(l,k) = 1 - \left(1 - \frac{1}{r^l}\right)^{k-l+1} \qquad (1)$$

ile ifade edilebilmektedir. Bu ifadede r, dizilerin türetildiği ortak alfabedeki harflerin sayısını göstermekte olup tüm harflerin eşit olasılıkla gerçeklendiği varsayılmaktadır. Halbuki, başka yazarlar (Guibas ve Odlyzko, 1981; Bilge vd., 2004; Mungan, 2007) tarafından gösterildiği üzere, dizilerin başka diziler içinde tekrarlanma yada başka dizileri barındırma olasılıkları, içlerindeki alt-diziler arasındaki korelasyonlara bağlı olarak değişiklik göstermektedir. Yukarıda bahsi geçen çalışmalarda, bu ortalama-alan yaklaştırımına ek olarak, bütün topolojik özelliklerin hesabında, düğüm çiftlerinin bağlanma olasılıkları birbirlerinden bağımsız olarak ele alınmıştır; bir düğümün komşuları arasındaki ilintililikler (korelasyonlar) ihmal edilmiştir.

Burada iki tane, çok basit fakat yol gösterici, kesin çözülebilir örnek üzerinde, dizilerin içeriklerinden kaynaklanan, ince-yapı olarak tabir edilebilecek farklılıkların ihmal edilemeyeceği durumlar sergilenecektir. Örneklerde vurgulanan noktalar şöyle özetlenebilir: *i)* Kilit dizilerin sonlu uzunlukta olduğu durumda enformasyon içeriklerinin ortalama enformasyon içeriğinden farklı olmaları. *ii)* Alt-dizilerin arasındaki korelasyonların, k-$l >> r^l$ olmadığı durumda ihmal edilemeyeceği. Burada belirtmek isteriz ki, ortalama-alan yaklaştırımının içerik-temelli ağın özelliklerini ne kadar yakından izleyeceği anahtar ve kilit dizilerin uzunluk dağılımlarıyla, özel olarak da, kilit dizilerin uzunluklarının anahtar dizilere kısayla ne kadar daha büyük olduğu ile belirlenmektedir.

Sabit uzunluklu diziler

Örneklerde, sistem büyüklüğü N olan içerik-temelli ağlar üzerinde, anahtar diziler sabit bir l uzunluğunda, kilit diziler ise yine sabit bir $k > l$ uzunluğunda tutulmuştur. Dizilerin içerikleri r tane, eşit ağırlıklandırılmış, harf içeren ortak bir alfabeden türetilmiştir. Dizilerin birbirleri içinde tekrarlanma olasılıklarını içeriklerine bağlı olarak hesaplayarak, bunlar ışığında, ağın giriş- ve çıkış-derece dağılımlarını analitik olarak elde edip sonuçlarımızı bilgisayar deneyleriyle yani simulasyonlarla karşılaştıracağız. Giriş ve çıkış "dereceleri", bir düğüme giren ve çıkan yönelimli kenar sayısıdır. Düğümün derecesi, giren veya çıkan kenarlarla o düğüme bağlı olan düğümlerin sayısıdır. İçerik-temelli bir ağda, bir düğümün çıkış-derecesi o düğümle eşleşmiş olan anahtar-dizinin kaç kilit dizi içinde tekrarlandığı, giriş-derecesi ise onunla eşleşen kilit-dizinin kaç anahtar diziyi barındırdığıyla belirlenir.

Birim uzunluklu anahtar diziler

İçerik-temelli modelde, her düğümün l=1 uzunluğunda bir anahtar ve gelişigüzel sabit bir k uzunluğunda bir kilit dizi ile eşlendiği bir ağ topluluğu düşünelim (Calcott vd., 2005). Burada, anahtar diziler 1 uzunluklu olduğu için (yani tek tek harflerden oluştukları için) verili bir anahtarın k uzunluklu rastgele seçilmiş bir kilit dizide bulunma olasılığı, hiç bir yaklaşıklık gerektirmeden, (1)'de verildiği gibidir, $p(1,k) = 1-\left(1-r^{-1}\right)^k$. Göstermek mümkün olacaktır ki, bir kilit diziye rastgele seçilen bir anahtarın (bu durumda, bir harfin) oturma olasılığı $p_I(1,k) = I/r$ ile verilip, burada I kilit dizinin içinde bulunan *farklı* haflerin (yani anahtarların) sayısıdır. Ağdaki düğüm sayısının çok büyük olduğu limitte (böylece, k uzunluklu tüm kilit diziler gerçeklenecek), giriş- ve çıkış-dereceleri binom dağılımlarını takip edecektir. Çıkış-derece dağılımı $P_{\text{ç}}(d)$'yi hemen yazmak mümkündür,

$$P_{\text{ç}}(d) = C(N,d)\left[p(1,k)\right]^d\left[1-p(1,k)\right]^{N-d} \qquad (2)$$

burada $C(N,d) = N!/\left[d!(N-d)!\right]$, kombinatorik faktör olup, çıkış-derece dağılımı $P_{\text{ç}}(d)$, rastgele seçilen bir düğümün derecesinin d olma olasığını vermektedir. Giriş-derece dağılımının hesabında ise daha dikkatli olmak gerekmektedir. Uzunluğu k olan kilit dizilerin içinde bulunabilecekleri farklı konfigürasyonların toplam sayısını $\omega_k = r^k$ ile içinde I farklı harf bulunduran kilit dizilerin konfigürasyonlarının sayısını ise $\omega_k(I)$ ile gösterelim. Hemen görüleceği üzere, bu iki büyüklük birbirine;

$$\omega_k = \sum_{I=1}^{\min(k,r)} \omega_k(I) \qquad (3)$$

biçiminde bağlıdır. Eğer bir dizinin içinde I farklı harf bulunduğu biliniyorsa, gelişigüzel seçilen bir harfin bunlardan biri olma olasılığı I/r'dir. Böylece bu tür kilit dizilerin giriş-derece dağılımı $P_I^{\text{g}}(d)$,

$$P_I^{\text{g}}(d) = C(N,d)\left(\frac{I}{r}\right)^d\left(1-\frac{I}{r}\right)^{N-d} \qquad (4)$$

ile verilecektir. Toplam giriş-derece dağılımı ise;

$$P_{\text{g}}(d) = \sum_{I=1}^{\min(k,r)} \frac{\omega_k(I)}{\omega_k} P_I^{\text{g}}(d) \qquad (5)$$

şeklindedir. Burada $\omega_k(I)/\omega_k$ rastgele seçilen bir kilit dizinin içinde I farklı harf bulunma olasılığıdır. Görüldüğü üzere, kilit diziler aynı uzunlukta olsalar dahi içlerinde bulundurdukları farklı alt dizilerin sayısına göre, genel durumda, $\max(I_l) = \min(k-l+1, r^l)$ tane alt gruba ayrılmaktadır. Şimdi yapmamız gereken, $\omega_k(I)$'yı, yani içinde I farklı harf bulunan k uzunluklu dizilerin sayısını hesaplamaktır. Uzunluğu k olan bir dizi içinde, "a_i" harfinin tekrarlanma sayısını n_{a_i} ile gösterelim. İçinde I tane farklı harf bulunduran dizilerin sayısını, bu harfler ve tekrarlanma sayıları $\{n_{a_i}\}$ verildiği durumda $\omega_k(I \mid \{n_{a_i}\})$ ile belirtelim. Bu harfler k uzunluklu bir dizi üzerinde;

$$\omega(I \mid \{n_{a_i}\}) = M(k, n_{a_1}, n_{a_2}, ..., n_{a_I}) \qquad (6)$$

farklı konfigürasyonda sıralanabilecek olup, burada $M(k, n_{a_1}, n_{a_2}, ..., n_{a_I}) = k!\left(n_{a_1}!n_{a_2}!...n_{a_I}!\right)^{-1}$ multinom faktörüdür. Kilit dizilerin uzunluğu k ve içlerindeki farklı harflerin sayısı I verildiğinde, iki koşulu sağlamamız gerektiğini hatırlayalım,

$$k = \sum_{i=1}^{I} n_{a_i} \quad , \quad 1 \le n_{a_i} \le k - I + 1 \quad . \qquad (7)$$

Şimdi buradaki ikinci koşulu da kullanarak, tüm mümkün harf $\{a_i\}$ ve tekrarlanma sayıları $\{n_{a_i}\}$ uzayını tarayarak $\omega_k(I)$'yı hesaplayacağız,

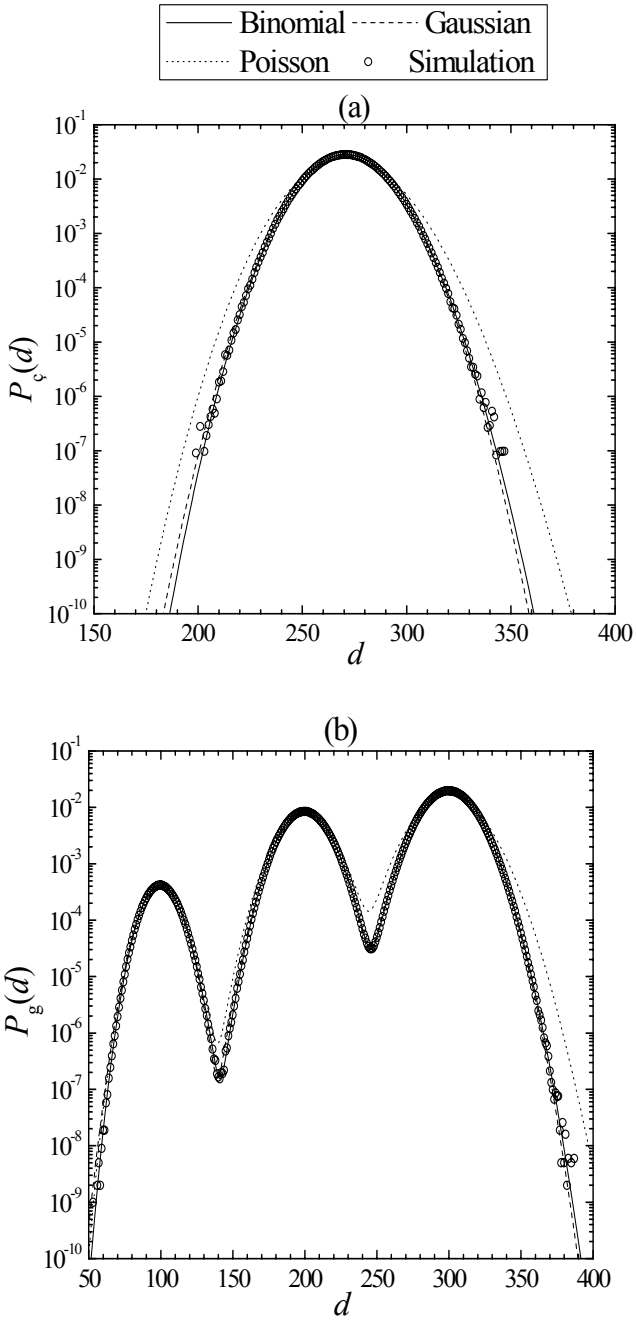
$$\begin{aligned} \omega_k(I) &= C(r,I) \sum_{\{n_{a_i}\}} \omega_k(I \mid \{n_{a_i}\}) \quad , \\ &= C(r,I) \sum_{n_{a_1}=1}^{k-I+1} ... \sum_{n_{a_{I-1}}=1}^{k-(n_{a_1}+...+n_{a_{I-2}})-1} \omega_k(I \mid \{n_{a_i}\}) \quad , \end{aligned} \qquad (8)$$

burada $C(r,I)$, r harften I tanesini kaç farklı şekilde seçebileceğimizi sayan kombinatorik faktördür. Toplamları yapmaya en sondaki elemandan başlarsak,

$$\omega_k(I) = C(r,I) \sum_{n=0}^{I-1} C(I,n)\left(I-n\right)^k \left(-1\right)^n \qquad (9)$$

elde ederiz.

Şekil 1'de, 10^6 gerçekleşim üzerinden ortalama alınarak elde edilen simülasyon sonuçları, çıkış ve giriş-derece dağılımları için elde ettiğimiz analitik hesaplarla karşılaştırılmaktadır. Burada sunulan sonuçlar $N = 10^3$ düğümden oluşan ağ üzerinde, k=3 ve r=10 alınarak elde edilmiştir. Aynı şekil üzerinde, binom dağılımları yerine yaklaşık Gauss ve Poisson dağılımları kullanılarak elde edilen sonuçlar da gösterilmektedir. Denklem (2) ve (5)'in grafiği denklem (9) kullanılarak hesaplanmış ve binom eğrisi olarak gösterilmiş olup, simülasyon sonuçları ile tamamen örtüşmektedir. Binom dağılımına Gauss yaklaşımı oldukça yakın kalmakta buna karşın Poisson yaklaşımı daha az isabet sağlamaktadır. Görüldüğü üzere giriş ve çıkış-derece dağılımları tamamen birbirlerinden farklıdırlar. Eğer giriş-derece dağılımının hesabında dizilerin içeriklerini ihmal etseydik (ortalama-alan yaklaşımında olduğu gibi), giriş-derece dağılımı için elde edilecek olan analitik ifade çıkış-derece dağılımınkiyle aynı olacaktı ve ağın yapısında tamamen yanılmış olacaktık.



Şekil 1. Çıkış- (a) ve giriş- (b) derece dağılımları

Alt-diziler arasındaki oto-korelasyonlar

İçerik-temelli ağın her düğümünün gelişigüzel sabit l uzunluklu bir anahtar diziye ve k=l+1 uzunluklu bir kilit diziye sahip olduğunu düşünelim. Bir önceki örnekten öğrenilenleri kullanarak ve genelleştirerek söylemek mümkün ki, içinde I_l farklı alt dizi bulunduran k uzunluklu dizilerin giriş-derece dağılımı;

$$P^{\mathrm{g}}_{I_l}(d) = C(N,d)\left(\frac{I_l}{r^l}\right)^d \left(1-\frac{I_l}{r^l}\right)^{N-d} \qquad (10)$$

olacaktır. Toplam giriş dağılımı ise;

$$P_{\mathrm{g}}(d) = \sum_{I=1}^{\min(k-l+1,r^l)} \frac{\omega_k(I_l)}{\omega_k} P^{\mathrm{g}}_{I_l}(d) \qquad (11)$$

şeklinde verilecektir. Bu örnek için, k-l=1 olduğundan I_l, 1 ve 2 değerlerini alabilecektir. Hemen görüleceği gibi içinde l uzunluklu tek örnek anahtar dizi bulunan k uzunluklu kilit dizilerin sayısı $\omega_k(I_l=1)=r$ olup geriye kalan tüm kilit dizilerin içinde en az iki tane birbirinden farklı l uzunluklu anahtar dizi bulunmaktadır, $\omega_k(I_l=2)=r^k-r$. Göstermek mümkün ki, anahtar diziler içlerinde bulunan alt diziler arasındaki korelasyonları ölçen otokorelasyon sayıları (Mungan vd., 2005; Mungan, 2007), başka bir deyişle "bit-vector"ler (Guibas ve Odlyzko, 1981) ya da kaydırma-eşleme sayılarına (Bilge vd., 2004) ("Shift-Match Number" SMN olarak kısaltılacaktır) göre özdeşlik sınıflarına ayrılmaktadırlar. Buna göre SMN'si aynı olan tüm anahtar dizilerin kendilerinden uzun diziler içinde bulunma olasılıkları aynıdır. Kaynak (Bilge vd., 2004)'deki notasyonu takip ederek, uzunluğu l olan bir anahtar diziyi $a=a_1a_2...a_l$ ile ve onun SMN'sini $s=s_1s_2...s_l$ ile gösterelim, s'nin i'inci elemanı $s_i=\prod_{j=i}^{l}\delta_{a_{1-i+j},a_j}$ ile verilirecektir. Buradan görüleceği üzere, bir dizinin SMN'si onunla eşit uzunlukta ve ilk elemanı 1 olan ikilik-tabandaki sayıdır. Örneğin, $a=$xxy ise $s=100$ olacaktır. Buna göre, SMN'si s olan anahtar dizilerin çıkış-derece dağılımı $P^{\mathrm{ç}}_s(d)$;

$$P^{\mathrm{ç}}_s(d) = C(N,d)\left[p_s(l,k)\right]^d\left[1-p_s(l,k)\right]^{N-d} \qquad (12)$$

ile verilip buradaki $p_s(l,k)$ uzunluğu l ve SMN'si s olan bir anahtar dizinin rastgele seçilen k uzunluklu bir kilit dizide bulunma olasılığıdır. Toplam çıkış-derece dağılımı ise;

$$P_{\mathrm{ç}}(d) = \sum_s \frac{\tilde{\omega}_l(s)}{\omega_l} P^{\mathrm{ç}}_s(d) \qquad (13)$$

şeklinde elde edilir. Burada $\tilde{\omega}_l(s)$ SMN'si s olan l uzunluklu dizilerin konfigürasyonlarının sayısı olup, $\tilde{\omega}_l(s)/\omega_l$ rastgele seçilen bir dizinin SMN'sinin s olma olasılığını vermektedir. Şimdi $p_s(l,k)$'yı hesaplamak için şöyle bir yol izlenecektir: Uzunluğu l ve SMN'si s olan bir anahtar dizi verilmiş olsun. Bu dizinin sağına ve soluna bir harf ekleyerek ondan türetilebilecek k=l+1 uzunluklu kilit dizileri sayalım. *i)* Eğer bu verili dizinin SMN'si $s^*=s=1...1...1$ ise, yani anahtar dizi $a=$x...x...x gibi tek bir harf (x) içeriyorsa, bu dizinin sağına ya da soluna yerleştireceğimiz her $\mathrm{y}\neq\mathrm{x}$ için yeni bir k=l+1 uzunluklu kilit dizi elde ederiz. Böylece, bu verili diziyi içeren kilit dizilerin sayısı $n_{s^*}(l,k)=2(r-1)+1$ ve rastgele seçilmiş bir kilit dizide bu verili diziyi bulma olasılığımız $p_{s^*}(l,k)$,

$$p_{s^*}(l,k) = \frac{n_{s^*}(l,k)}{\omega_k} = \frac{2r-1}{r^k} \qquad (14)$$

olur. Böyle anahtar dizilerin sayısı $\omega_l(s^*)=r$'dir. *ii)* Şimdi SMN'si $s^*\neq s$ olan bir anahtar dizi verildiğini varsayalım. Böyle bir dizinin sağına ya da soluna yerleştireceğimiz her harf için yeni bir k=l+1 uzunluklu kilit dizi elde ederiz. Böylece, bu verili diziyi içeren kilit dizilerin sayısı $n_{s\neq s^*}(l,k)=2r$ ve rastgele seçilmiş bir kilit dizide bu verili diziyi bulma olasılığımız $p_{s\neq s^*}(l,k)$,

$$p_{s\neq s^*}(l,k) = \frac{n_{s\neq s^*}(l,k)}{\omega_k} = \frac{2r}{r^k} \qquad (15)$$

olacaktır. Böyle anahtar dizilerin sayısı $\omega_l(s \neq s^*) = r^l - r$ 'dir. Burada dikkat edilecek husus, bu sonuca s'nin değerini değil yanlızca onun s^*'dan farklı olduğunu bilerek ulaştık. Bu her durumda geçerli değildir. Denklem (14) ve (15)'deki sonuçlar k-l=1 durumu için geçerlidir. Burada elde ettiğimiz sonuçları kullanırsak;

$$P_{s^*}^{ç}(d) = C(N,d)\left(\frac{2r-1}{r^k}\right)^d \left(1-\frac{2r-1}{r^k}\right)^{N-d} \qquad (16)$$

ve

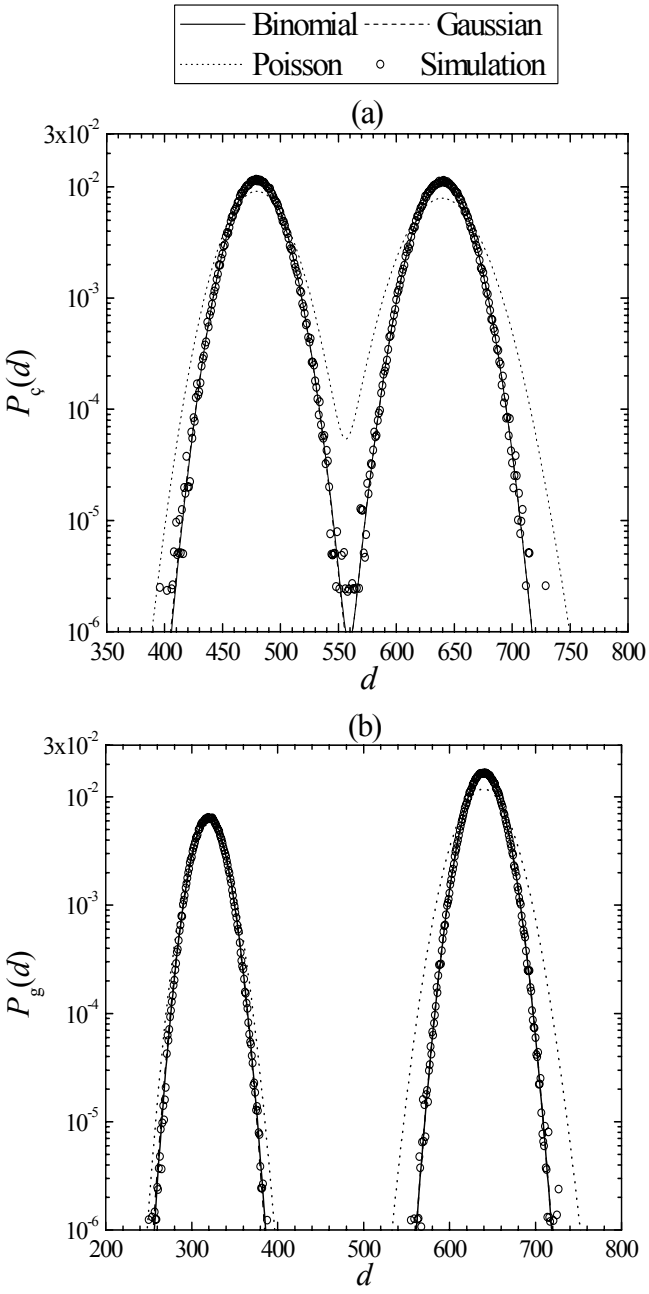
$$P_{s\neq s^*}^{ç}(d) = C(N,d)\left(\frac{2r}{r^k}\right)^d \left(1-\frac{2r}{r^k}\right)^{N-d} \qquad (17)$$

olur. Sonunda, toplam çıkış-derece dağılımı,

$$P_{ç}(d) = \frac{\tilde{\omega}_l(s^*)}{\omega_l} P_{s^*}^{ç}(d) + \frac{\tilde{\omega}_l(s\neq s^*)}{\omega_l} P_{s\neq s^*}^{ç}(d) \qquad (18)$$

olarak elde edilir. Burada gördüğümüz yozluk, yani SMN'si $s \neq s^*$ olan anahtar dizilerin aynı giriş-derece dağılıma tabi olmaları, bu örneğe hasdır, k-l arttıkça bu yozluk ortadan kalkacaktır (Bilge vd., 2004).

Şekil 2'de, 10^5 gerçekleşim üzerinden ortalama alınarak elde edilen simülasyon sonuçları, çıkış- ve giriş-derece dağılımları (denklem (11) ve (18)) için elde ettiğimiz analitik hesaplarla karşılaştırılmaktadır. Burada sunulan sonuçlar $N = 1280$ düğümden oluşan ağ üzerinde, l=2 ve r=2 alınarak elde edilmiştir. Aynı şekil üzerinde, binom dağılımlarının Gaussian ve Poisson yaklaşıklıkları da gösterilmektedir. Görüldüğü üzere, teorik eğriler (sürekli çizgi) simülasyonlarla tamamen örtüşmektedir. Eğer bu dağılımların hesabında dizilerin içeriklerini ihmal etseydik, giriş-derece dağılımı için elde edilecek olan analitik ifade çıkış-derece dağılımınkiyle aynı, (1)'deki olasılık ifadesinde r=2 alınırsa, maksimumu $Np(2,3)$'de olan bir binom dağılımı elde edecektik.



Şekil 2. Çıkış- (a) ve giriş- (b) derece dağılımları

Sonuçlar

Elde edilen sonuçlar şöyle özetlenebilir:

- Dizi içeriklerini ihmal ederek yapılan hesaplar, içerik-temelli ağların topolojik özelliklerini tasvir etmekten uzaklaşabilmektir. Burada dikkat edilecek husus, bunun dizilerin uzunluk dağılımlarına bağlı olduğudur.

- Giriş-derece dağılımının hesabında, kilit diziler içlerinde bulunan farklı alt dizilerin sayılarına göre gruplandırılabilirler. Aynı gruba dahil olan tüm kilit diziler aynı derece dağılımına tabidir.
- Çıkış-derece dağılımının hesabında, anahtar diziler SMN'lerine göre gruplandırılabilirler. Aynı SMN'ye sahip tüm anahtar diziler aynı derece dağılımına tabidir.

Kaynaklar

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. ve Walter, P., (2002). *Molecular biology of the cell*, Garland Science, New York.

Balcan, D. ve Erzan, A., (2007). Content-based networks: a pedagogical overview, submitted to *Chaos*.

Balcan, D., Kabakçıoğlu, A., Mungan, M. ve Erzan, A., (2006). A content-based approach to modeling the topological properties of the transcriptional regulation network of yeast, submitted to *PLoSONE*; q-bio.MN/0605045.

Balcan, D. ve Erzan, A., (2004). Random model for RNA interference yields scale free network, *European Physical Journal B,* **38**, 253-260.

Bilge, A.H., Erzan, A. ve Balcan, D., (2004). The shift-match number and string matching probabilities for binary sequences, unpublished; q-bio.GN/0409023.

Calcott, B., Balcan, D. ve Hohenlohe, P., (2005). *Modeling the evolution of development* in *Complex Systems Summer School Final Project Papers*, Santa Fe Institute, Santa Fe, NM.

Guibas, L.J. ve Odlyzko, A.M., (1981). Periods in strings, *Journal of Combinatorial Theory A*, **30**, 19-42.

Hannon, G.J., (2002). RNA interference, *Nature*, **418**, 244-251.

Mézard, M., Parisi, G. ve Zecchina, R., (2002). Analytic and algorithmic solution of random satisfiability problems, *Science*, **297**, 812-815.

Mungan, M., (2007). String matching and 1d lattice gases, *Journal of Statistical Physics*, **126**, 207-242.

Mungan, M., Kabakçıoğlu, A., Balcan, D. ve Erzan, A., (2005). Analytical solution of a stochastic content-based network model, *Journal of Physics A*, **38**, 9599-9620.

Wray, G.A., Hahn, M.W., Abouheif, E., Balho., J.P., vd., (2003). The evolution of transcriptional regulation in eukaryotes, *Molecular Biology and Evolution*, 20, 1377-1419.