Cilt: 07 Sayı: 01 Kategori: 02 Sayfa: 70-82 ISSN No: 1303-6319 Volume: 07 Number: 01 Category: 02 Page: 70-82

BİYOKİMYASAL REAKSİYONLAR İÇİN STOKASTİK SİMÜLASYON ALGORİTMALARINA GENEL BİR BAKIŞ

Vilda PURUTÇUOĞLU*

ÖZET

Biyolojik bir sistemi anlayabilmek için hangi genlerin/proteinlerin organizmanın neresinde, ne zaman ve nasıl reaksiyonda olduğunu bilmemiz gerekmektedir. Bu kadar detaylı, karmaşık ve metabolik seviyede rassal olan reaksiyonları içeren biyokimyasal bir mekanizmada, hücresel aktivitelerin deneysel olarak ispatlanması, teknolojik imkanların sınırlı olması sebebiyle çoğu kez mümkün olmamakta veya yüksek deney maliyetine sebep olmaktadır. Biyokimyasal modelleme; bir sistemin elemanlarının farklı zaman ve şartlar altındaki durumunu, sistemi oluşturan proteinler ve moleküller arasındaki etkileşimi sistemin bilinen özellikleri yardımıyla ifade etmenin matematiksel yoludur. Bu çalışmada; reaksiyonların nasıl formülize edildiği ve bu reaksiyonlardan oluşan sistemin stokastik modellemelerinin biyoinformatik ve matematiksel biyoloji alanlarında hangi simülasyon algoritmalarıyla yapıldığı tanıtılmaktadır.

Anahtar Kelimeler: Matematiksel modelleme, Simülasyon, Stokastik algoritmalar.

1. GİRİŞ

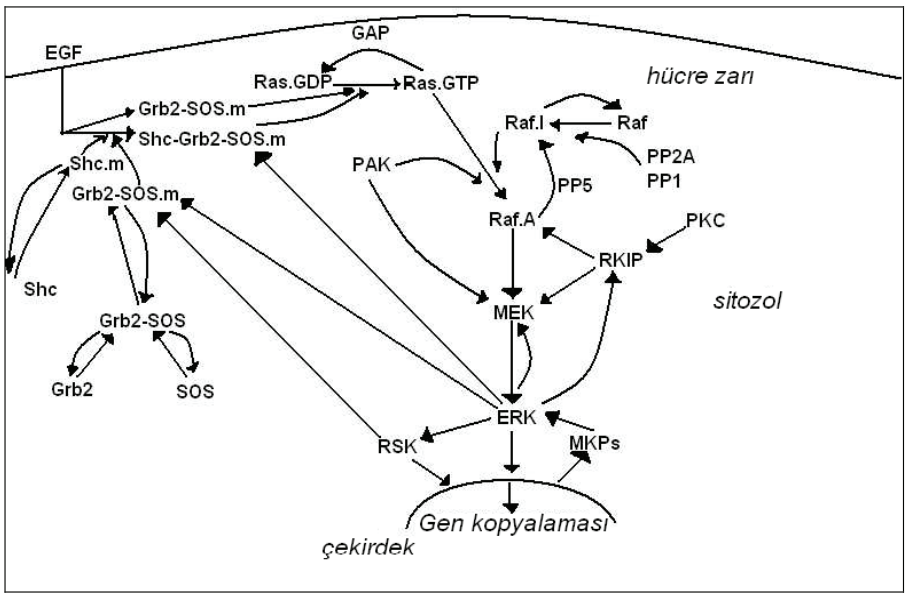
Biyokimyasal reaksiyon, bir biyokimyasal işlemin nitel veya nicel olarak tanımlanmasıdır. Basit bir biyokimyasal reaksiyon (1) nolu ifadede verilen biçimde gösterilebilir:

$$m_1 R_1 + m_2 R_2 + \ldots + m_r R_r \longrightarrow s_1 P_1 + s_2 P_2 + \ldots + s_p P_p \qquad (1)$$

Burada R ile gösterilen soldaki terimlere "reaksiyona girenler" ve P ile gösterilen sağdaki terimlere "reaksiyondan üretilenler" denir. $m_i \ (i = 1,\ldots,r)$ ve $s_j \ (j = 1,\ldots,p)$ katsayıları, sırasıyla $i.$ reaksiyona girenin "stokometrik katsayısını" (stoichiometric coefficients) ve bu reaksiyondan üretilenin "stokometrik katsayısını" göstermektedir. r, reaksiyona girenlerin sayısı ve p bunun sonucunda üretilenlerin sayısını gösterir. Dolayısıyla bu eşitliğin kimyasal yorumu; moleküller Brownian hareketiyle rassal olarak hareket ederken, R_i cinsinden m_i kadar molekül birbiriyle çarpışır ve P_j cinsinden s_j kadar molekül üretir şeklindedir. Bir başka ifade ile bir kimyasal reaksiyon, sıcaklık dengesi altında ve sabit bir hacimde hangi moleküllerin, hangi oranda birbiriyle reaksiyona girdiklerini ve sonucunda ne üretildiğini ifade eder (Wilkinson, 2006).

* Öğretim Görevlisi Dr., Orta Doğu Teknik Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, e-posta: vpurutcu@metu.edu.tr

İfade (1) aynı zamanda $MY \rightarrow SY$ şeklinde vektör formu ile de tanımlanabilir. Burada $Y = (Y_1,...,Y_n)$ ve n, sırasıyla sistemin o andaki "durum vektörünü" (state vectors) ve molekül cinslerinin sayısını verir. $M = (m_1,...,m_n)$ ve $S = (s_1,...,s_n)$ ise, sırasıyla reaksiyona girenlerin ve reaksiyondan üretilenlerin stokometri vektörleridir. Burada da önceki açıklamaya benzer şekilde, bir reaksiyon gerçekleştiği zaman Y_g $(g = 1,...,n)$'nin molekül sayısı m_g kadar azalır ve s_g kadar artar. Sonuç olarak, molekül transferi sistemde $V = S - M$ kadarlık net değişime sebep olur. Bu ifadede v, n boyutlu "net etki vektörünü" (net effect vector) anlatır (Bower ve Bolouri, 2001).

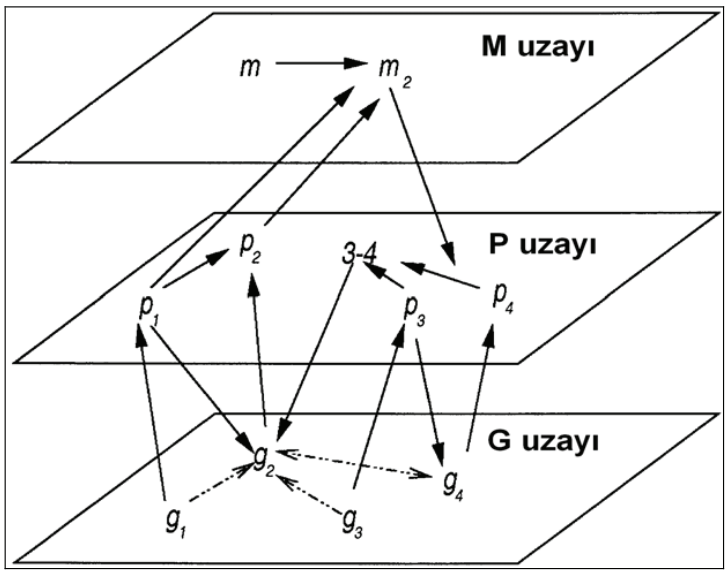


Şekil 1. Mitojen Aktivleyici Protein Kinezi (MAPK) Sisteminin Şemasal Olarak Basit Gösterimi

Eğer bir kimyasal aktiviteyi anlatan r tane eşitlik varsa, bu reaksiyon seti Şekil 1'de gösterildiği gibi bir biyolojik sistemi oluşturur (Purutçuoğlu ve Wit, 2008). Gelişen teknoloji, biyolojik sistemlerin daha yakından tanınmasında büyük kolaylıklar getirmiştir. Bu alanda özellikle son yıllarda yapılan araştırmalar bu sistemlerdeki karmaşık ve çok seviyeli yapının ortaya çıkmasını sağlamıştır. Şekil 2 farklı seviyelerdeki biyolojik sistem yapısını basitçe göstermektedir (Khanin ve Wit, 2006). Şekildeki her seviye birbiriyle ilişkili ve yakın biçimde bağımlı olmalarına rağmen araştırmalarda birbirinden bağımsız seviyeler olarak ele alınır. Burada G uzayı gen (g_i) etkileşim uzayını, P uzayı protein (p_i) etkileşim uzayını ve M uzayı molekül (m_i) etkileşim uzayını temsil etmektedir. Koyu renkli oklar direkt etkileşim bağlantılarını, buna karşın noktalı oklar genlerin dolaylı etkileşimlerini sembolize etmektedir. Karmaşık biyolojik yapıların çözülmesi, bir çok biyolojik aktivitenin daha iyi anlaşılmasında ve özellikle kanser, kalp rahatsızlığı gibi ciddi hastalıklarda kilit rol oynayan proteinlere yönelik yeni tedavi yöntemlerinin geliştirilmesinde çok büyük

öneme sahiptir. Matematiksel modelleme yöntemleri bu yapıların henüz deneysel olarak gün ışığına çıkmamış özelliklerini bulmada ve biyolojik açıdan yeni soruların ortaya çıkmasında ilgili alanlarda çalışan araştırmacılara yeni kapılar açmaktadır (Endy ve Brent, 2001; Brent, 2004).

Farklı matematiksel modeller altında bir sistemi meydana getiren reaksiyon seti farklı şekillerde tahmin edilebilir. Bir biyokimyasal sistemi modellemek için üç temel teknik vardır: Bunlar Boolean, diferansiyel eşitlikler ve stokastik metodlardan meydana gelen saf veya katıksız teknikler; kinetik mantık ve sürekli mantık sistemlerini (continuous logical networks) içeren ara teknikler (intermediate methods) ve son olarak Langevin ve Fokker-Planck yaklaşımlarını ifade eden, melez tekniklerdir (Bower ve Bolouri, 2001; Jong, 2002). Bu çalışmada, saf/katıksız tekniklerden biri olan, stokastik metodlar ve onların simülasyon algoritmalarının üzerinde durulacaktır.



Şekil 2. Farklı Seviyelerdeki Biyolojik Sistem Yapısının Genel Anlamda Gösterimi

2. STOKASTİK METODLAR

Stokastik metodlar, reaksiyonda moleküllerin sayısının tam olarak bilindiği durumlarda kullanılır. Bu modelde durumlar (states), olasılıksal olarak sistemin sonraki duruma geçmesine neden olan her bir molekül cinsinin o anki sayısını gösterir. Diğer bir deyişle hangi değişimin olduğu ve bunun ne zaman olduğu olasılık yaklaşımıyla ifade edilir. Örneğin;

$$A \xrightarrow{c} A'$$

şeklindeki kimyasal ifade, verilen bir tane A molekülü için bir tane A' molekülü oluşturmasının, dt zamanı içindeki olasılığını $c \times dt$ olarak tanımlar. c, bir biyokimyasal olayın bir birim zaman başına olma olasılığını gösteren "reaksiyon oran sabitidir" (reaction rate constant). Buna bağlı olarak kısa bir zaman aralığı içinde A'dan A''a dönüştürülmüş molekül olasılığı $c[A]dt$ ile gösterilir. $[A]$, A'nın o anki molekül sayısıdır (Bower ve Bolouri, 2001; Wilkinson, 2006).

Stokastik modeller, durum değişikliğini olasılıksal olarak gösterdiği için aynı şartlar altındaki reaksiyonlardan farklı cevaplar alınabilir. Bu rassallık altında işlemin bazı istatistiklerini bulmak için, rassal sayıların üretilmesine dayanan Monte Carlo teknikleriyle sistemin simülasyonu yapılabilir (Wilkinson, 2006).

Stokastik sistemlerde her bir molekül cinsinin molekül sayısı, denge halindeki sistemin sıradan diferansiyel eşitliklerdeki (Ordinary Differential Equations – ODE) gibi çözülmesiyle bulunamaz. Çünkü her oluşan reaksiyon, molekül sayısını değiştirir. Buna karşın sistemin mümkün olan bütün molekül sayısı üzerinden olasılık dağılımı hesaplanabilir (Bower ve Bolouri, 2001). Bu olasılığı hesaplamak için de "verilen molekülün, sabit hacim içerisinde her hangi bir yerde bulunması eşit olasılığa sahiptir" varsayımı yapılabilir.

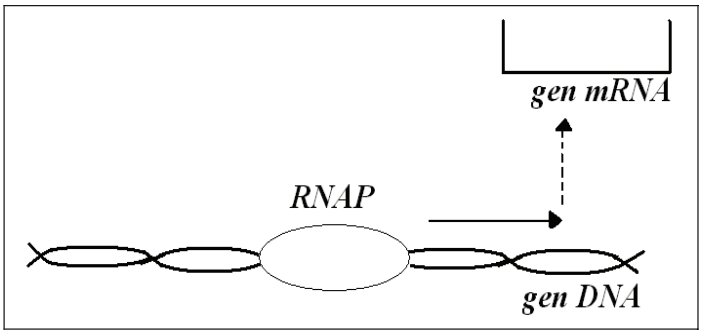
3. STOKASTİK SİMÜLASYON ALGORİTMALARI

Gen düzenleri (gene regulation) genellikle ODE ile modellenir. ODE modelleri kütle hareket kanununa (mass actian laws) ve her bir kimyasal proteinin sürekli konsantrasyonuna dayanır (Lok, 2002; Lok ve Brent 2005). Her ne kadar bu metodlar lineer üretim ve parçalanma reaksiyonları (reaction of degradation) gibi bazı reaksiyonları ifade etmede başarılı olsalar da, gerçek reaksiyonların küçük sistem çeşitliliğini açıklayamazlar. Biyokimyasal sistemleri modellemek için stokastik işlemler bu sebeple doğal bir tercih olmuştur (Fedoroff ve Fontana, 2002; Turner vd., 2004). Nitekim bu şekildeki dinamik modelleme, biyolojik zamanda az sıklıkla oluşan, İfade (2)'de basitçe anlatılan ve Şekil 2'de de gösterilen protein kopyalanması gibi farklı biyolojik reaksiyonların olasılıksal durumlarını ele alabilir (Hume, 2000).

$$\begin{aligned} DNA + RNAP &\longrightarrow DNA \times RNAP \\ DNA \times RNAP &\longrightarrow DNA + RNAP + mRNA \end{aligned} \qquad (2)$$

İfade (2)'de ilk reaksiyon, ilgilenilen gene RNA polimerazının (*RNAP*) bağlanmasını, dolayısıyla DNA-RNAP kompleksinin oluşmasını (*DNA.RNAP*), ikinci reaksiyon ise DNA'sı okunan genin kopya RNA'sının (*mRNA*) oluşturulmasını göstermektedir.

Kimyasal sistemlerin stokastik algoritmaları "master denklemlerine" (master equations), diğer bir isimle "Chapman-Kolmogorov denklemlerine" dayanır (Kampen, 1981; Jong, 2002). Master denklemleri sistemin stokastik davranışını aşağıdaki eşitlikle gösterir:



Şekil 2. DNA Kopyalanmasının Basit Bir Gösterimi

$$\frac{\partial P(Y,t)}{\partial t} = \sum_{j=1}^{r} \left\{ h_j(Y - v_j) P(Y - v_j, t) dt - h_j(Y) P(Y,t) dt \right\} \qquad (3)$$

Bu eşitlikte r, sistemdeki reaksiyon sayısını verir. v_j, $(r \times n)$ boyutlu net etki matrisi V'nin $j.$ satırını ifade eder. Buna bağlı olarak $h_j(Y)$ "risk fonksiyonunu" (hazard function), diğer bir adla "reaksiyonun oran kanununu" (reaction rate laws) anlatır. $h_j(Y)$, $j.$ reaksiyon için stokastik oran sabiti c_j ile Y durumunda elde bulunan ve reaksiyona giren moleküllerin farklı kombinasyonlarının çarpımıdır. Matematiksel ifadeyle ve Eşitlik (1)'deki terimlerle $h_j(Y) = a_j(Y)c_j$, $(j = 1,...,r)$ ve $a_j = \binom{R_1}{m_1} \times \binom{R_2}{m_2} ... \times \binom{R_r}{m_r}$'dir. Buradan Eşitlik (2)'deki $h_j(Y - v_j)P(Y - v_j, t)dt$ ifadesinin $[t, t + dt]$ zaman aralığında sistemin $Y - v_j$ durumundan Y durumuna geçerken $j.$ reaksiyonunun olma olasılığını verdiği sonucu çıkar.

Master denklemleri iki aşamalı olarak kurulur:

i) Tüm olası reaksiyonlar olma olasılıkları ile sıralanır,

ii) Sistem birim zaman ve birim molekül başına sabit reaksiyon riskleri ile birlikte lineer diferansiyel denklemler biçiminde ifade edilir.

Master denklemleri, olası durum sayısının küçük olduğu sistemlerde çözülebilir denklemlerdir. Buna karşın büyük sistemlerde, bu denklemlerin çözülmeleri, değişkenlerin sayısı, dolayısıyla her bir durumda bulunma olasılıkları hızlı bir biçimde arttığı için farklı Monte Carlo algoritmalarıyla yapılır. Bu algoritmalar temelde, gerçek olasılık dağılımının yaklaşık sonucunu, yaklaşık dağılımlardan tekrarlı örneklem seçerek bulur. Bu örnekler bilgisayarların rassal sayı yaratan üreteçleri sayesinde oluşturulur (Wilkinson, 2006; Turner vd., 2004).

Stokastik simülasyon yapan dört temel metod vardır. Bu dört metod da Eşitlik (3)'de verilen master denklemini sağladıkları için üretilen değerler yaratılmak istenen biyolojik sistemi tam olarak yansıtır. Dolayısıyla aşağıda detaylarıyla tanıtılacak metodlar arasında değerlerin doğrulukları (accuracy) ve sapmaları (deviation) açısından homojen bir sistem için fark yoktur. Ancak özellikle hesaplama süreleri, kullanıcı kolaylığı ve işlemler için kullanılan zaman aralıklarının sürekli veya kesikli olmaları bakımından

kendi içinde farklılıklar göstermektedir. Bu farklılıklar algoritmaların tanıtılması sırasında belirtilmekle beraber Tablo 1'de de özet halinde sunulmaktadır.

3.1 Direkt Metot (Gillespie Algoritması)

Gillespie algoritma olarak da bilinen direkt metot (Gillespie, 1977), Eşitlik (3)'te verilen kimyasal master denklemlerine dayanan en yaygın ve fazla karmaşık olmayan sistemlerde hesaplama süresi açısından genellikle en hızlı simülatördür.

Algoritma, reaksiyon olasılık dağılım fonksiyonunu, zaman rassal değişkeni τ $(0 \le \tau < \infty)$ uzayında ve kesikli reaksiyon belirleyici (discrete reaction indicator) değişkeni j $(j = 1,...,r)$ ile aşağıda verilen fonksiyon yarımıyla tanımlar.

$$P(\tau, j) = h_j(Y)\exp\{-h_0(Y)\tau\};\ (0 \le \tau < \infty). \qquad (4)$$

Bu ifadede $h_0(Y) = \sum_{j=1}^{r} h_j(Y) = \sum_{j=1}^{r} a_j(Y)c_j$ 'dir. Gillespie algoritması özet olarak şu adımları takip eder.

(1) Sisteme başlangıç zamanı olan $t = 0$ durumu için, $c_1,...,c_r$ reaksiyon oran sabitleri ve her bir türün (proteinin, molekülün veya genin) başlangıç zamanındaki $Y_1, Y_2,..., Y_n$ olan türleri için molekül popülasyon sayıları verilir. r ve n, sırasıyla sistemdeki toplam reaksiyon ve tür sayısını göstermektedir.
(2) Türlerin reaksiyon riskleri, her reaksiyon için $h_j(Y) = a_j(Y)c_j$ $(j = 1,...,r)$ ile hesaplanır.
(3) Rassal olarak oluşacak bir sonraki reaksiyonun ne kadar sürelik zaman aralığında olduğunu bulmak için, olası zaman aralığını gösteren τ, $h_0(Y)$ oranlı üstel dağılımdan $(\tau \sim Exp\{h_0(Y)\})$ oluşturulur ve bu zaman aralığında hangi reaksiyonun olacağı $h_j(Y)/h_0(Y)$ $(j = 1,...,r)$ olasılığı ile, her olay birbirinden bağımsız kabul edilerek seçilir.
(4) Zaman değişkenini güncellemek için son bulunulan zaman birimi t, τ kadar arttırılır $(t := t + \tau)$ ve tür popülasyonunun seviyesi, reaksiyonların stokometri katsayıları yardımıyla güncellenir. Eğer güncellenen t başta verilen zaman aralığı T 'den küçükse $(t < T)$, ikinci adıma geri dönülür ve algoritma $t \ge T$ olana kadar tekrarlanır. T simülasyon için başlangıçta belirlenen toplam süredir.

Gillespie algoritması her ne kadar yüksek doğruluğa sahip sonuçlar verse de ve sistemi tam olarak oluştursa da uzaysal heterojenlik veya lokalizasyonluk gerektiren durumlarda uygulanması zordur (Gillespie, 1992; Kitano, 2001). Ayrıca her ne kadar küçük sistemleri yaratmada oldukça başarılı olsa da karmaşık ve büyük sistemlerde, yarattığı zaman aralıklarının sadece tek bir reaksiyonun olmasını sağlayacak kadar küçük olması sebebiyle hesaplama süresi açısından avantajlı değildir (Gibson ve Bruck, 2000; Bower ve Bolouri, 2001; Turner vd., 2004).

3.2 İlk Reaksiyon Metodu

İlk reaksiyon metodu, Gillespie metodunun karmaşık sistemlerdeki hesaplama zamanını kısaltmak için önerilen ve temelde aynı hesaplama mantığını kullanan alternatif bir saf simülasyon metodudur. Fakat Gillespie, reaksiyon belirleyici j ve zaman değişkeni τ'u direkt olarak üretirken, ilk reaksiyon metodu, oluşabilecek ilk reaksiyon j için tahmin edilen zamanı τ_j'yi üretmesi yönünden farklıdır (Gillespie, 1992; Gibson ve Bruck, 2000; Wilkinson, 2006). Ayrıca her ne kadar her iki metot da j ve τ'yu seçmek için aynı olasılık dağılımını kullanıyor olsalar da, ilk reaksiyon metodu her tekrarlamada (iteration) bir yerine, r tane reaksiyon yaratır.

Bu algoritmanın aşamaları aşağıdaki gibi sıralanabilir:

(1) Her bir tür (protein, molekül veya gen) için başlangıç değeri olan türlerin popülasyon sayıları verilir ve zaman göstergesi t'nin başlangıç değeri sıfıra eşitlenir.
(2) $h_j(Y) = a_j(Y)c_j$ risk fonksiyonu tüm j'ler için hesaplanır.
(3) $h_j(Y)$ parametreli üstel fonksiyondan her bir j için tahmin edilen zaman τ_j yaratılır ($\tau_j \sim Exp\{h_j(Y)\}$).
(4) En küçük τ_j sonraki reaksiyon için zaman adımı olarak seçilir ($\tau = \tau_j$).
(5) j ve τ'ya göre türlerin popülasyon sayıları güncellenir.
(6) t, τ kadar arttırılır ($t := t + \tau$). Eğer $t < T$ ise algoritma ikinci adımdan itibaren tekrarlanır.

Genel olarak bu metod, Gillespie metod gibi heterojenlik veya lokalizasyonluk gerektiren durumlar için bir çözüm üretememektedir. Ancak hesaplama süresi açısından karmaşık sistemlerde Gillespie'ye göre biraz daha hızlıdır. Kullanım yaygınlığı açısından ise Gillespie'den daha yaygın değildir. Çünkü bu metodun üstünlüğü ancak sistem karmaşıklaştıkça gözükmekte, avantajlı olduğu durumlarda ise Gillespie'nin algoritmadaki basitliği sebebiyle yine de göreceli olarak daha az tercih edilmektedir.

3.3 Sonraki Reaksiyon Metodu (Gibson-Bruck Algoritması)

"Sonraki reaksiyon metodu" olarak da bilinen Gibson-Bruck algoritması özellikle karmaşık sistemler için Gillespie ve ilk reaksiyon algoritmalarından daha hızlı ve etkili olan bir metoddur (Gibson ve Bruck, 2000; Lok ve Brent, 2005; Cao vd., 2006). Algoritma her bir reaksiyon için sadece risk fonksiyonu $h_j(Y)$'yi kullanmak yerine, hem zaman adımı olan τ_j, hem de $h_j(Y)$ $(j = 1,...,r)$'yi birlikte kullanan etkili bir hesaplama yöntemi geliştirmiştir. Temel olarak bu metod sistem için oluşturulan bir bağlantı grafiği ς (dependency graph) yardımıyla $h_j(Y)$'yi hesaplar. Bahsedilen bağlantı grafiği j. reaksiyonunun olması halinde $h_j(Y)$ değerinden etkilenen türleri (protein, molekül veya gen) gösterir. Bu işlem, algoritma tarafından j türünün, j'nin değişmesi halinde j'ye bağlı olan ve kendi risk fonksiyonlarının da güncellenmesi gereken diğer tüm proteinleri birbirine bağlayan bir yapının kurulmasıyla sağlanır. Bu

amaçla bağlantı grafiği yerine Petri net gösterimi de alternatif yöntem olarak kullanılabilir (Wilkinson, 2006). Bu şekilde sistem lokal olarak güncellenebildiği için simülasyon hızı artar. Diğer yandan bu lokal alt gruplar arasındaki bağlantılar "endeksli öncelik sırası" (indexed priority queue) adı verilen bir grafik yapısıyla sağlanır. Bu grafik iki temel eleman içerir:

i) (j, τ_j) çifti şeklinde sıralı ağaç yapısı. Burada j reaksiyon sayısını, τ_j ise, j reaksiyonunun oluşması durumunda tahmin edilen zaman aralığını ifade eder.

ii) (j, τ_j) içeren ağaç yapısında j. elemanın yerini gösteren endeks yapısı.

Sırayı oluşturan bu ağaç yapısı içinde her reaksiyon kendisine bağlı olan alt grup reaksiyonlardan daha küçük τ_j'ye sahiptir.

Gibson-Bruck algoritması şu şekilde çalışmaktadır:

(1) Reaksiyon sabitlerine c_j $(j = 1,...,r)$ ve türlerin (protein, molekül veya gen) popülasyon sayılarını gösteren Y'ye başlangıç değerleri verilir. Bunlara bağlı olarak risk fonksiyonu $h_j(Y)$ hesaplanır. Bu risk değerleri $h_j(Y)$ parametresiyle üstel dağılımdan yaratılan ve ilk yaratılan zaman aralıkları olan τ_j'leri hesaplamada kullanılır ($\tau_j \sim Exp\{h_j(Y)\}$).

(2) Endeksli öncelik sırası içinde en küçük τ_j, k. endeks sırasını alır ve t, τ_k'ya eşitlenir ($t = \tau_k$).

(3) k. reaksiyonun oluşmasıyla durum vektörü Y güncellenir.

(4) Yeni durum Y'ye göre $h_k(Y)$ güncellenir. Yeni tahmin edilen zaman aralığı üstel $Exp\{h_k(Y)\}$'dan üretilir ve $\tau_k := t + Exp\{h_k(Y)\}$ olarak hesaplanır.

(5) Reaksiyon k ve risk fonksiyonu, değişen her j reaksiyonu için ($j \neq k$)

a. $h_j'(Y) = h_j(Y)$ olarak güncellenir ve eski $h_j(Y)$ geçici süreyle tutulur.

b. $\tau_j := t + (h_j(Y) / h_j'(Y))(\tau_j - t)$ olarak hesaplanır.

c. Eski $h_j(Y)$ sistemden silinir.

(6) Her j reaksiyonu için ($j = k$) dördüncü adım tekrarlanır.

(7) Eğer t istenilen toplam T `den küçükse, $t < T$, algoritma ikinci adımdan itibaren tekrarlanır.

Bu algoritma saf stokastik simülasyon metodları içinde, büyük ve oluşan her bir reaksiyonun sistemdeki diğer reaksiyonların çoğunluğunu etkilemediği (loosely-coupled) biyolojik sistemlerde hesaplama süresi açısından en etkin metottur (Cao vd., 2006). Ancak, algoritmanın yazılım dilinin C olması sebebiyle kullanıcı kolaylığı açısından daha teknik donanım bilgisine gereksinim duymaktadır. Dolayısıyla Gillespie ve ilk reaksiyon metodlarına göre daha az yaygınlıkta kullanılmaktadır. Buna rağmen heterojenlik, lokalizasyonluk gibi sistem problemlerini çözebilmesi ve oldukça büyük sistemlerin "loosely-coupled" şartını sağladığı durumlarda, özel koşullar için tercih edilen bir algoritmadır.

3.4 Stokastik Simülatör (StochSim)

Stokastik Simülatör, kısaca StochSim (Morton-Firth ve Bray, 1998), Gillespie, ilk reaksiyon veya Gibson-Bruck algoritmalarında uygulanan "reaksiyonu simülasyon etme" mantığı yerine, kesikli ve belli bir zaman aralığı içinde look-up tablosu denilen hazır tablolar ve sahte türler (pseudo proteins, molecules or genes) yardımıyla her bir türü (protein, molekül veya gen) ayrı ayrı yaratan bir yöntemdir.

Algoritma, Vol hacminde n tane proteine/moleküle sahip bir sistem için her adımda iki tür seçerek, tekli ve ikili tür içeren reaksiyonların olasılığını hesaplar. İlk seçim her zaman gerçek bir türdür ve sistemdeki n tanecik içinden seçilir. İkinci seçimden önce sisteme n_0 sahte tür ilave edilir ve ikinci seçim ($n+n_0$) tür arasından yapılır. Bu ikinci seçimde eğer ikinci seçilen bir sahte tür ise algoritma tek proteinli/moleküllü bir reaksiyon, şayet gerçek bir tür ise iki proteinli/moleküllü bir reaksiyon yaratır. Örnek olarak A türünden sistemde n_A adet protein/molekül varsa t zamanı için tek proteinli/moleküllü bir reaksiyon olması

$$\frac{d[A]}{dt} = -c_1[A]$$

ile ifade edilir. Bu eşitlikte c_1, tek tür içeren reaksiyonun reaksiyon oran sabiti ve $[A]$, A'nın o andaki popülasyonudur. Bu durumda çok kısa bir zaman aralığı içinde A'nın popülasyonundaki n_A değişimi $\Delta n_A = -c_1 n_A \tau$ olarak bulunur ve p_1 tek tür içeren reaksiyon oranı olmak üzere

$$-\Delta n_A = \frac{n_A}{n} \times \frac{n_0}{n+n_0} \times p_1 \qquad (5)$$

olarak hesaplanır.

Eğer sistem A ve B türleri arasında iki proteinli/moleküllü bir reaksiyon yaratıyorsa, bu reaksiyon

$$\frac{d[A]}{dt} = -c_2[A][B]$$

olarak gösterilir. c_2, iki tür içeren reaksiyonun reaksiyon oran sabitidir. $[A]$ ve $[B]$ ise, sırasıyla, A ve B türlerinin popülasyonlarıdır. Tekli reaksiyona benzer şekilde burada da çok kısa bir zaman aralığı τ içinde A'nın popülasyonundaki n_A değişimi $\Delta n_A = -\frac{c_2 n_A n_B \tau}{N_A Vol}$ olarak bulunur. Bu eşitlikte N_A Avogadro sabitidir ve n_B, B'nin popülasyonudur. p_2, iki proteinli/moleküllü reaksiyon olasılığını göstermek üzere

$$-\Delta n_A = 2 \times \frac{n_A}{n} \times \frac{n_B}{n+n_0} \times p_2 \qquad (6)$$

şeklinde formüle edilir. Eşitlik (5) ve (6)'nın birlikte çözülmesiyle, sırasıyla,

$$p_1 = \frac{c_1 n(n+n_0)\tau}{n_0} \quad \text{ve} \quad p_2 = \frac{c_2 n(n+n_0)\tau}{2N_A Vol} \tag{7}$$

biçiminde bulunur.

Eşitlik (7)'deki matematiksel ifadedeki n_0 sayısı, en hızlı tekli ve ikili protein/molekül reaksiyonlarının olasılıklarının yaklaşık olarak eşit olabileceği şekilde ayarlanır. Bu da

$$n_0 = INT\left(2N_A \frac{c_{1,mak}}{c_{2,mak}}\right)$$

ifadesiyle hesaplanır. Burada $INT(x)$, x'e en yakın pozitif tamsayıyı verir. $c_{1,mak}$ ve $c_{2,mak}$ ise, sırasıyla, maksimum tekli ve maksimum ikili protein/molekül reaksiyonlarının, reaksiyon oran sabitleridir.

StochSim simülatörü hazır tablo elemanlarını Eşitlik (7)'deki olasılık değerleriyle oluşturur. Eğer standart uniform $(U \sim (0,1))$ 'dan üretilen rassal sayı, bulunan p_1 veya p_2'den küçük ise o türler arasındaki reaksiyon oluşmuş kabul edilir. Aksi takdirde reaksiyonun gerçekleşmediği sonucuna varılır. Son aşamada simülatör, türlerin bağlanıp bağlanamamasına göre sistemi günceller. Simülasyon, sonraki zaman aralığının belirlenmesi ve yeni bir protein/molekül çiftinin seçilmesiyle tekrar başlatılır.

Bu özelliği ile StochSim enzim sistemlerinin modellenmesine daha uygundur ve özellikle küçük bir hacim içerisinde çok sayıda reaksiyonun olması durumunda hesaplama zamanı açısından daha verimlidir (Bower ve Bolouri, 2001).

Table 1. Biyokimyasal reaksiyonlar için önerilen saf simülasyon algoritmalarının karşılaştırılması

Karşılaştırma kriterleri	Direkt metod (Gillespie)	İlk reaksiyon metodu	Sonraki reaksiyon metodu (Gibson-Bruck)	Stokastik simülatör (StochSim)
Doğruluk	Tam	Tam	Tam	Tam
Hesaplama süresi	Uzun	Kısa	Çok kısa	Çok kısa
Kullanıcı kolaylığı	Çok yaygın	Yaygın	Az yaygın	Çok yaygın
İşlemsel zaman aralığı	Sürekli zaman	Sürekli zaman	Sürekli zaman	Kesikli zaman
Uygulandığı reaksiyon türleri	Protomik/moleküler reaksiyonlar	Protomik/moleküler reaksiyonlar	Protomik/moleküler reaksiyonlar	Enzim reaksiyonları
Heterojen ve lokalizasyonlu reaksiyon performansı	Etkili değil	Etkili değil	Etkili	Etkili

4. SONUÇ VE TARTIŞMA

Bu çalışmada stokastik simülasyon teknikleriyle biyolojik organizmalardaki karmaşık sistemlerin bilgisayar yardımıyla oluşturulması, kullanılan algoritmalar tanıtılarak ve birbirleriyle kıyaslanarak (Tablo 1) açıklanmıştır. Bahsedilen dört yöntem, verdiği cevapların doğruluğu açısından oldukça etkili yöntemlerdir. Buna karşın ilgilenilen sistemlerdeki yapının büyüklüğü, sistem biyoloji ve hesaplamalı biyoloji alanlarındaki araştırmaların yerine, yaklaşık cevaba yönelik olan ancak, işlemsel olarak çok daha hızlı olan algoritmaların geliştirilmesi ihtiyacını ortaya çıkarmaktadır. Bu konuda da yine son yıllarda farklı metodlar önerilmiş ve kıyaslamalı çalışmalar yapılmıştır (Turner vd., 2004, Cao vd., 2005; Auger vd., 2006; Purutçuoğlu ve Wit, 2006; Purutçuoğlu, 2010). Önerilen yaklaşık simülasyon metodlarında ise performans kriteri, bahsedilen bu dört metodun, özellikle kullanıcı yaygınlığı ve kolaylığı açısından avantajlı olan Gillespie algoritmasının verdiği sonuçlara göre, yüksek doğruluk ve düşük standard sapmaya sahip olması olarak değerlendirilir (Turner vd., 2004, Cao vd., 2005; Auger vd., 2006). Analizler, hala farklı boyutlardaki sistemleri modellemede kabul edilebilir hesaplama süresi önerebilen yeni tam/katıksız veya yaklaşık stokastik simülasyon tekniklerine ihtiyaç olduğunu göstermektedir.

5. KAYNAKLAR

Auger, A., Chatelain, P., Koumoutsakos, P., 2006. R-leaping: Accelerating the stochastic simulation algorithm by reaction leaps.

Bower, J. M., Bolouri, H., 2001. Computational modelling of genetic and biochemical networks. Massachusetts Institute of Technology.

Brent, R., 2004. A partnership between biology and engineering. Nature Biotechnology, 22:469–482.

Cao, Y., Gillespie, D. T., Petzold, L. R., 2005. Avoiding negative populations in explicit poisson Tau-Leaping. Journal of Chemical Physics, 123:054104.1-054104.8.

Cao, Y., Li, H., Petzold, L., 2006. Efficient formulation of the stochastic simultion algorithm for chemically reacting system. Journal of Chemical Physics, 121 (9): 4059-4067.

Endy, D., Brent, R., 2001. Modelling cellular behaviour. Nature, 409:391–395.

Fedoroff, N., Fontana, W., 2002. Genetic networks: Small numbers of big molecules. Science, 297:1129–1131.

Gibson, M. A., Bruck, J., 2000. Efficient exact stochastic simulation of chemical systems with many species and many channels. Journal of Physical Chemistry, A(104):1876–1889.

Gillespie, D. T., 1977. Exact stochastic simulation of coupled chemical reactions. Journal of Physical Chemistry, 81(25):2340–2361.

Gillespie, D. T., 1992. A rigorous derivation of the chemical master equation. Physica A, 188:404–425.

Hume, D. A., 2000. Probability in transcriptional regulation and its implications for leukocyte differention and inducible gene expression. Blood, 96:2323–2328.

Jong, H. D., 2002. Modeling and simulation of genetic regulatory systems: A literature review. Journal of Computational Biology, 9 (1), 67-103.

Kampen, N. G. V., 1981. Stochastic Processes in physics and chemistry. Elsevier.

Khanin, R., Wit, E., 2006. How scale-free are biological networks. Journal of Computational Biology, 13 (3), 810-818.

Kitano, H., 2001. Foundations of systems biology. Massachusetts Institute of Technology.

Lok, L., 2002. Pathfinder and other tools for analyzing signal transduction networks. Ann. N. Y. Acad. Sci., 971:589–594.

Lok, L., Brent, R., 2005. Automatic generation of cellular reaction networks with moleculizer 1.0. Nature Biotechnology, 23(1):131–136.

Morton-Firth, C., Bray, D., 1998. Predicting temporal fluctuations in an signalling pathway. Journal of Theoretical Biology, 192:117–128.

Purutçuoğlu, V., Wit, E., 2006. Exact and approximate stochastic simulations of the MAPK pathway and comparisons of simulations' results. Journal of Integrative Bioinformatics, 3, 231-243.

Purutçuoğlu, V., Wit, E., 2008. Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters. Bayesian Analysis, 3 (4), 851-886.

Purutçuoğlu, V., 2010. Stochastic simulation of large biochemical systems by approximate Gillespie algorithm. Proceeding of the 5rd International Symposium on Health, Informatics and Bioinformatics, IEEE Xplore, 181-186.

Turner, T. E., Schnell, S., Burrage, K., 2004. Stochastic approaches for modelling in vivo reactions. Computational Biology and Chemistry, 28:165–178.

Wilkinson, D. J., 2006. Stochastic modelling for systems biology. Chapman and Hall/CRC.

AN OVERVIEW TO STOCHASTIC SIMULATION ALGORITHMS FOR BIOCHEMICAL SYSTEMS

ABSTRACT

In order to understand a biological system, we should know which genes/proteins react together, where, when, and how they react in the organisms. In such a biochemical mechanism which is detailed, complex, and stochastic in metabolic level, the experimental validations of cellular activations cannot be typically applicable due to the current technological limitations or the high expenses of the possible experiments. The biochemical modelling is a mathematical way to describe the elements of a system, their proteomic and metabolic interactions, their states under different time points and various conditions by using the known theories about that system. In this study we review how formalize the biochemical reactions and which simulation algorithms can be performed to stochastically model a system whose components are described by these biochemical reactions in the frameworks of bioinformatics and mathematical biology.

Keywords: Mathematical modelling, Simulation, Stochastic algorithms.