

Moleküler Evrimde Birleşme (Coalescence) Modelleri ve Filocoğrafya

İsmail K. Sağlam

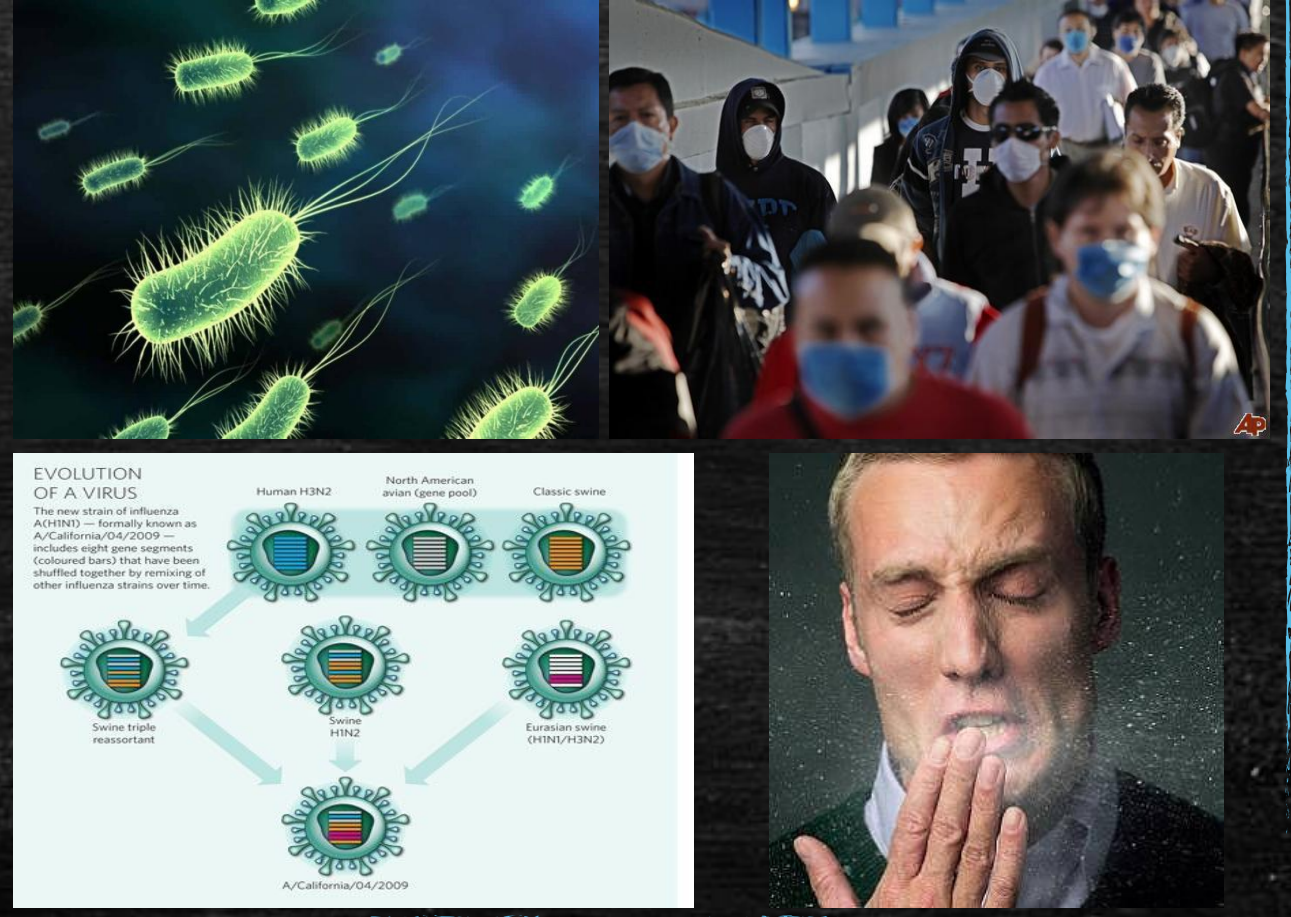
Hacettepe Üniversitesi, Biyoloji Bölümü, Ekolojik Bilimler Araştırma Laboratuvarları

Ders Boyunca Neler Hakkında Konuşacağım

- Hangi problemleri çözmeye çalışıyoruz?
- Populasyon genetiği ve filogenetiğin kesişim noktası : Birleşme
 - Birleşme (Coalescent) nedir?
 - Standard Birleşme Kuramı.
 - Standart Birleşmenin Uzantıları.
- Birleşme kuramı altında çözümleme.
 - Bayes Kuramı ve Olabilirlik (Likelihood).
 - MCMC Algoritması.
 - Ön ve ardıl Olasılık Dağılımları.
- Uygulama.

Çözülmesi Gereken Problemler

- Yeni hastalıkların ortaya çıkış ve yayılış hızları nedir?
- Hastalıkların orijin noktalarından yayılış yolları nelerdir?
- Kaç farklı tür grip virüsü vardır? Birbirleriyle olan evrimsel ilişkileri nedir?
- Mutasyon birikim hızları nedir?



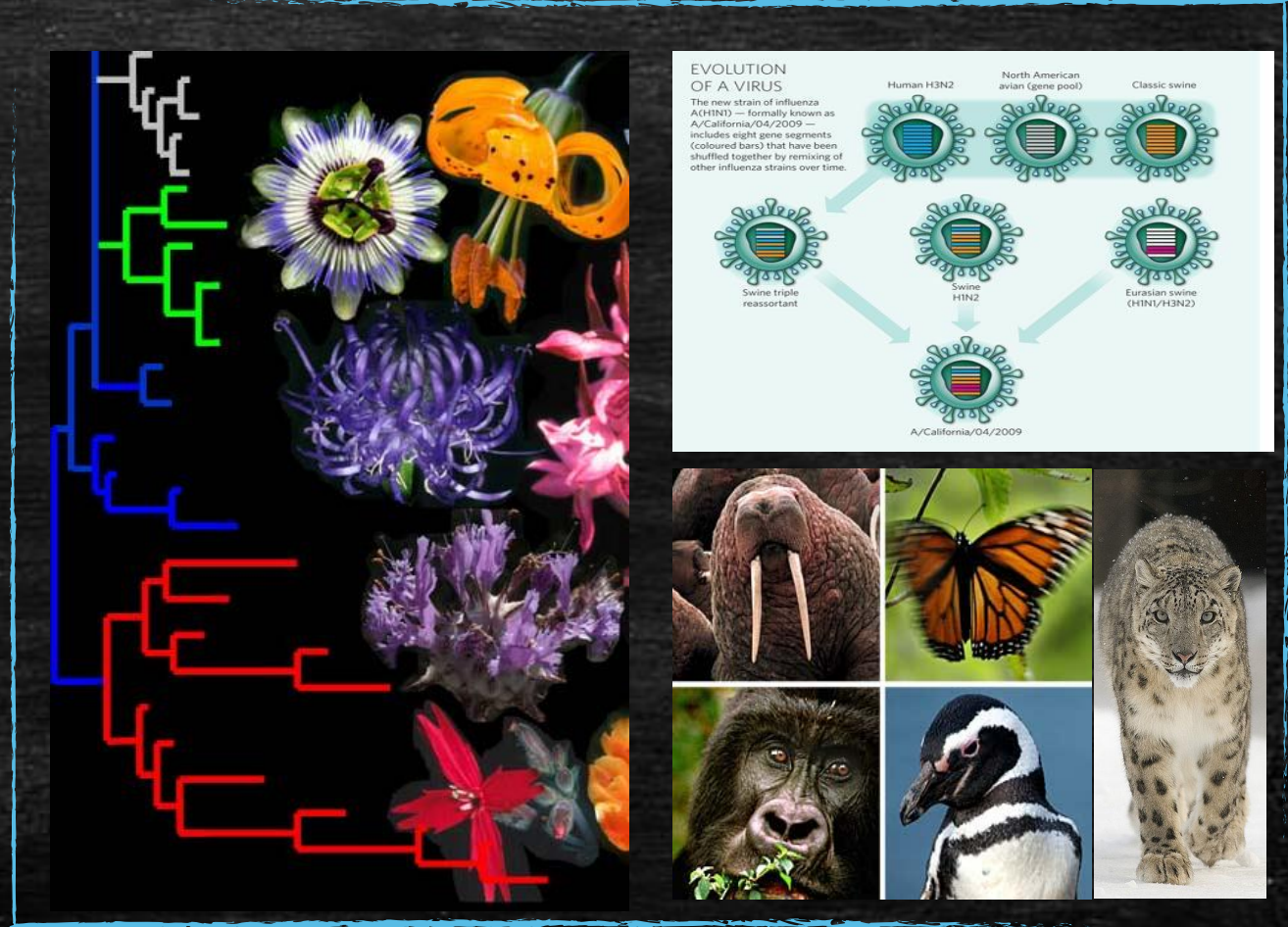
Çözülmesi Gereken Problemler

- Tehlike altındaki türlerin popülasyon büyüklükleri sürdürülebilir olmaları için ne kadar küçük olabilir?
- Popülasyonların genetik çeşitliliğini nasıl yüksek tutabiliriz.
- Popülasyonlar arasındaki bağlantılar nasıldır?
- Günümüzdeki dinamikler ile geçmiş dinamikler arasındaki farklar nelerdir?



Çözülmesi Gereken Problemler

- Türler, populasyonlar, bireyler birbirlerine evrimsel tarihçe ile bağlıdır.
- Sorularımızı evrimsel tarihçeyi göz önüne alarak cevaplandırmamız lazım.



Kesişim Noktası: Birleşme Kuramı

FİLOGENETİK

Ağaçlar

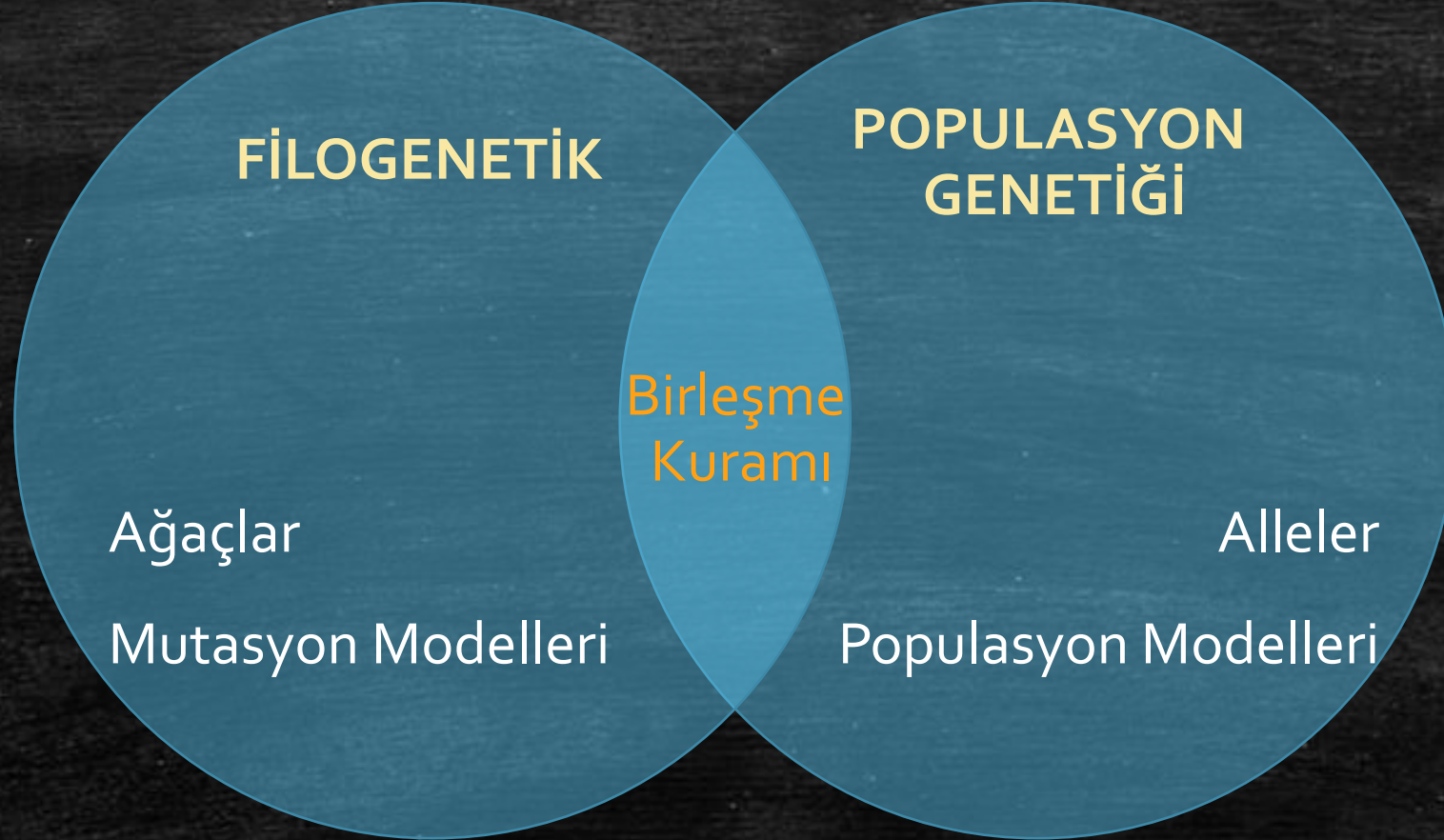
Mutasyon Modelleri

POPULASYON GENETİĞİ

Alleler

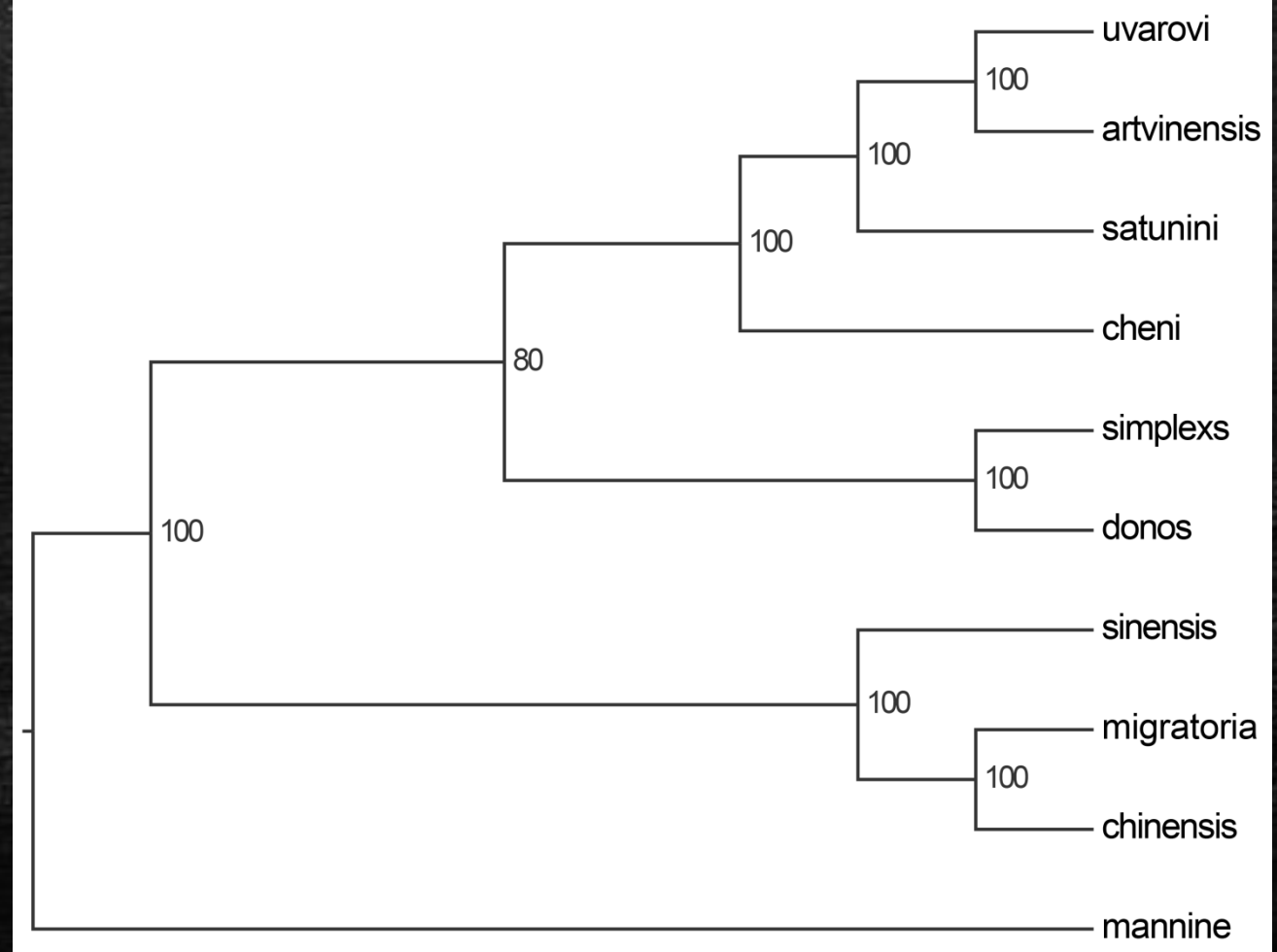
Populasyon Modelleri

Kesişim Noktası: Birleşme Kuramı



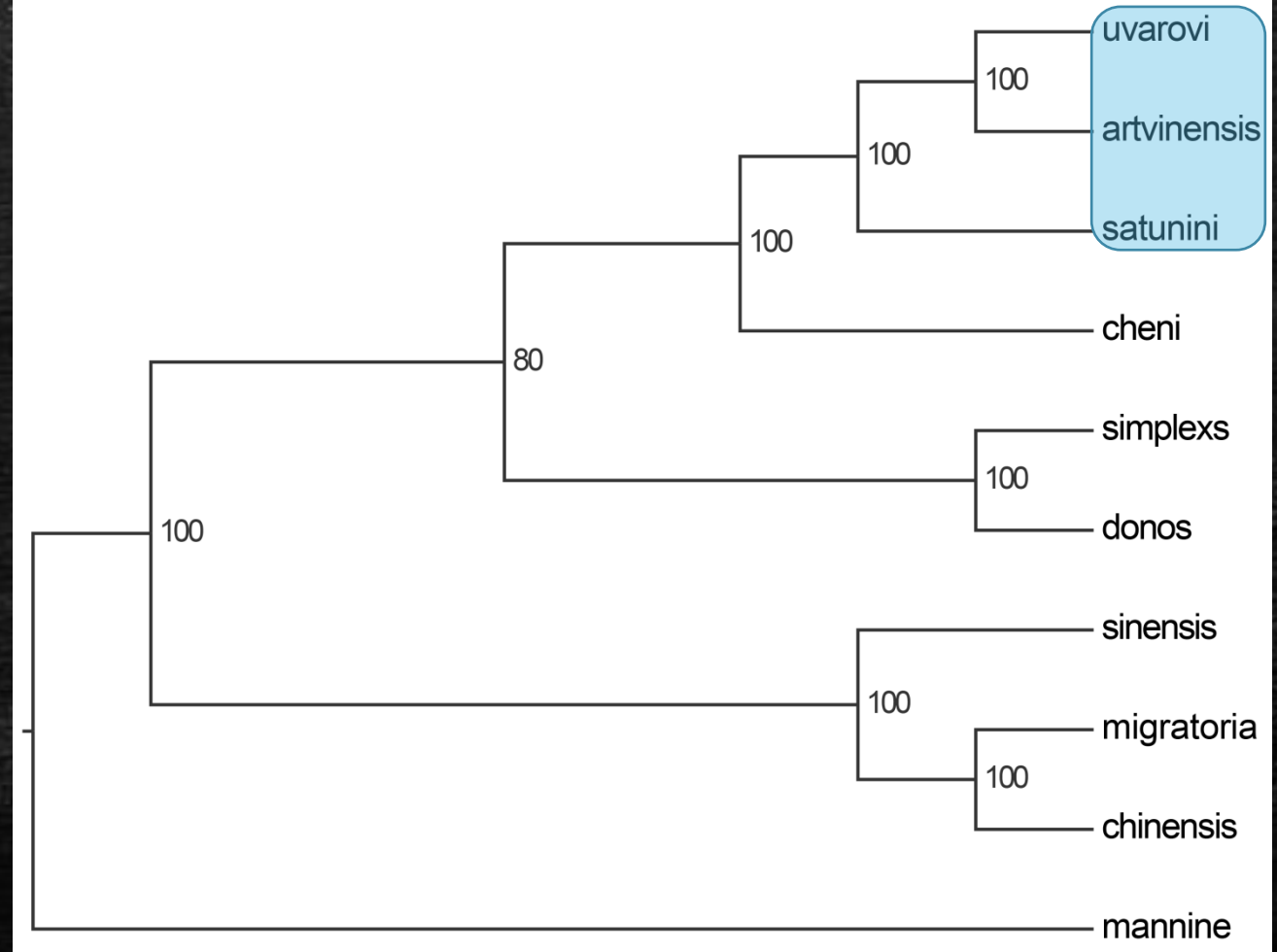
Kesişim Noktası: Birleşme Kuramı

Tür Ağaçları



Kesişim Noktası: Birleşme Kuramı

Tür Ağaçları



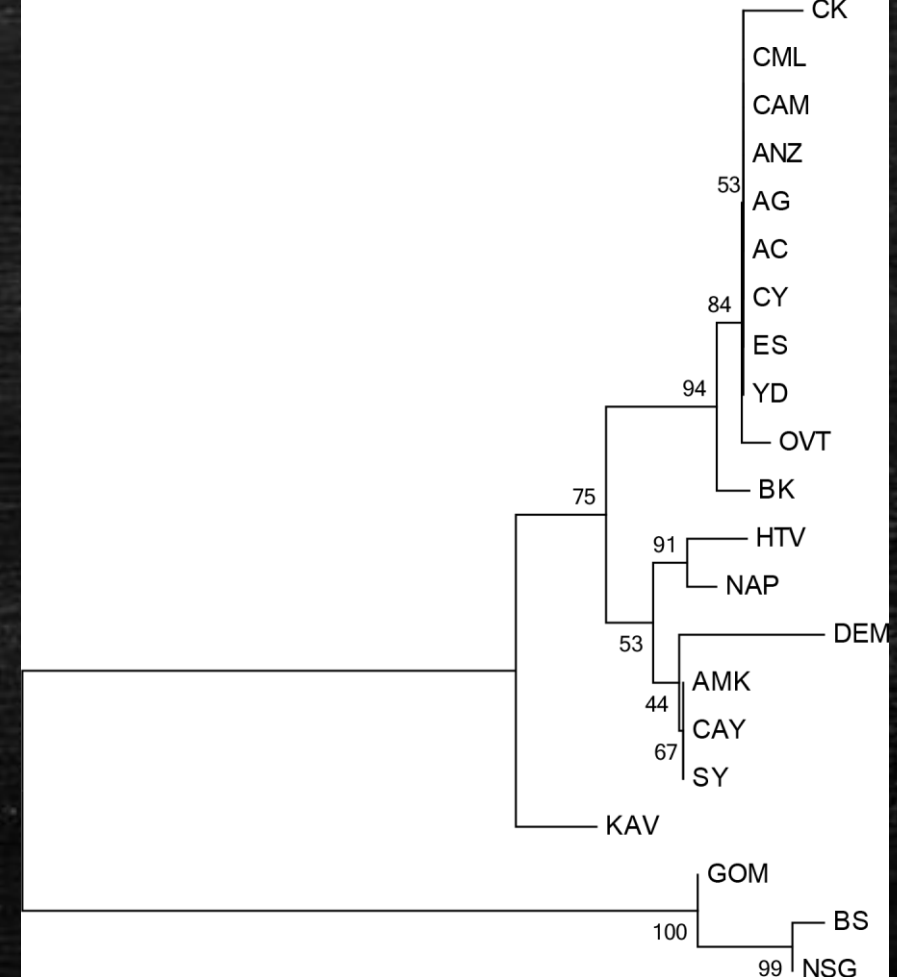
Kesişim Noktası: Birleşme Kuramı

Tür Ağaçları



Çözünürlük

Populasyon Ağaçları



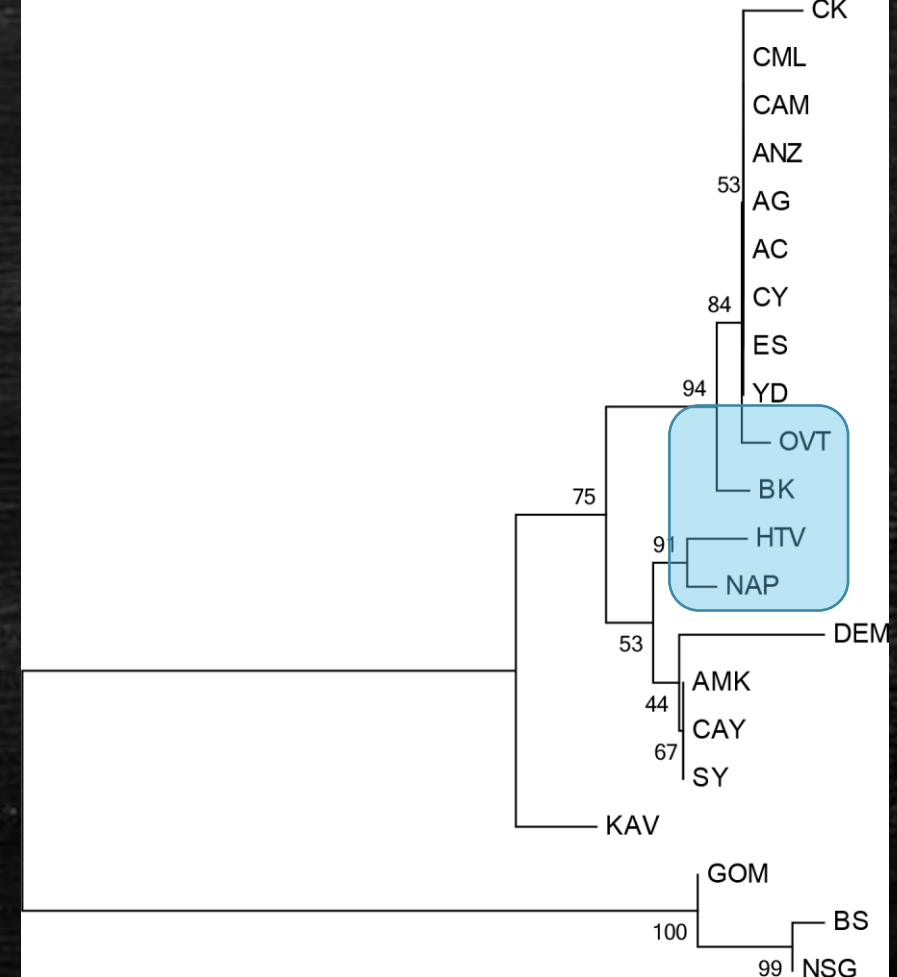
Kesişim Noktası: Birleşme Kuramı

Tür Ağaçları

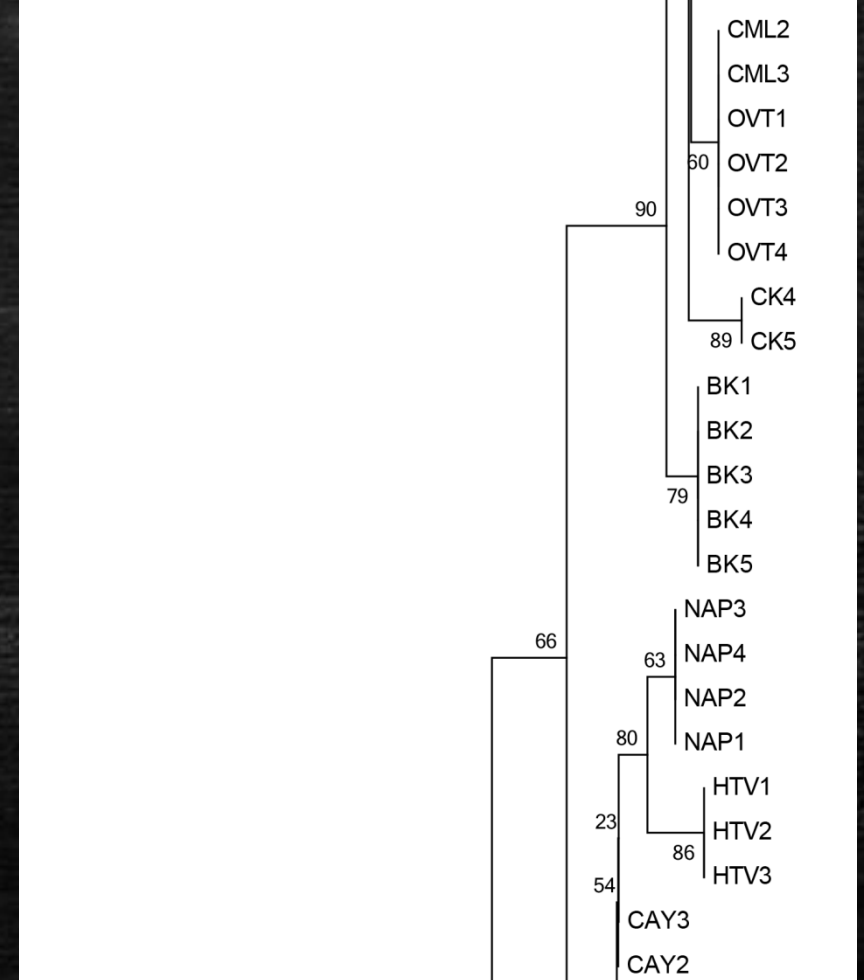
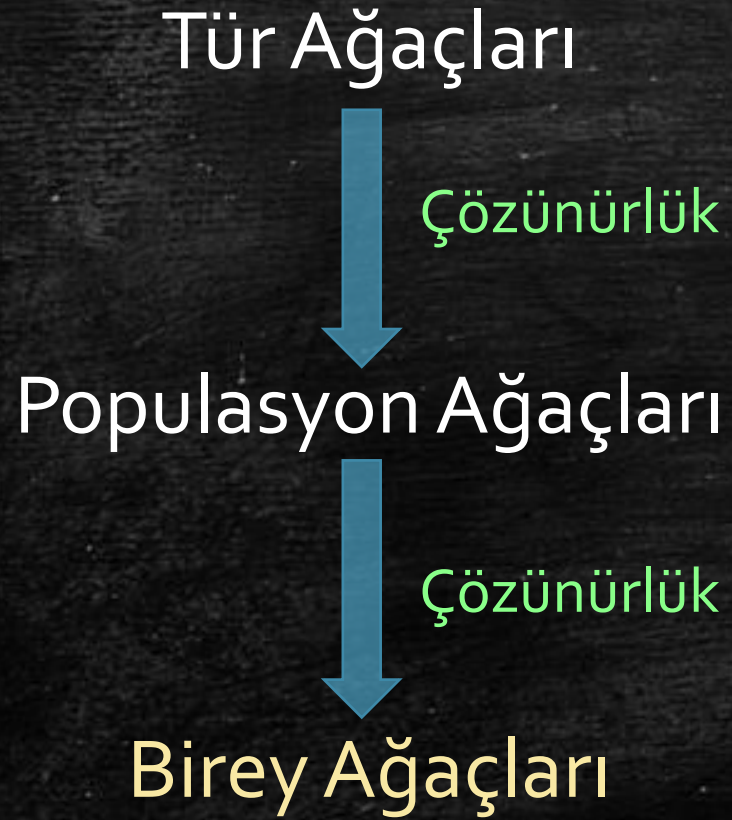


Çözünürlük

Populasyon Ağaçları



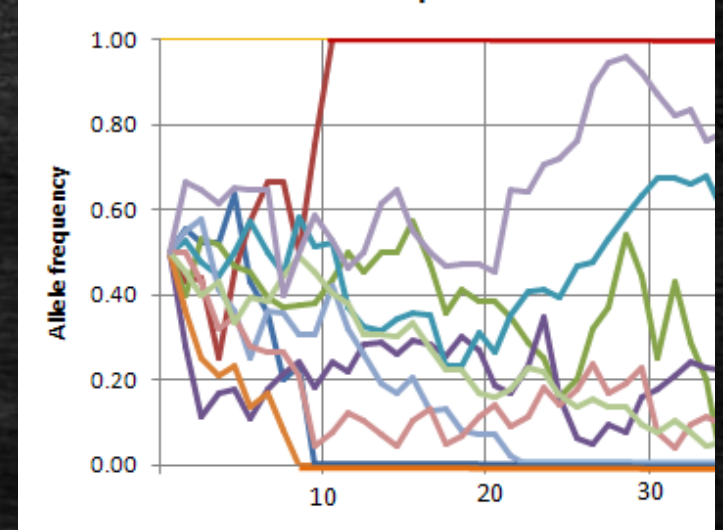
Kesişim Noktası: Birleşme Kuramı



Modellemenmesi Gereken İki Süreç

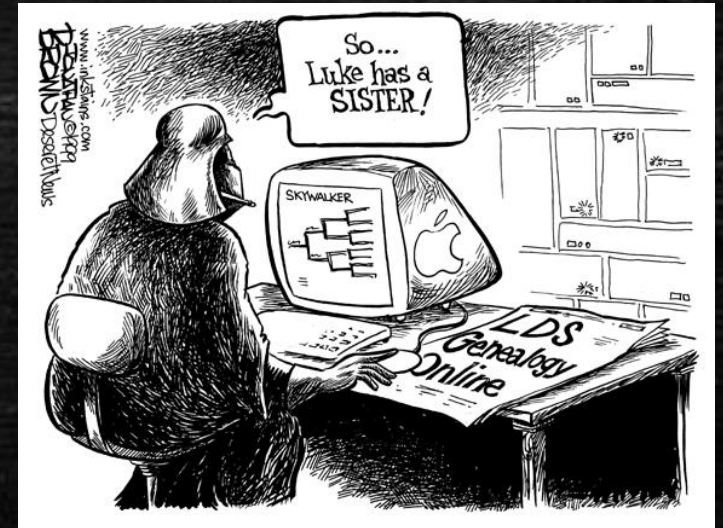
Mutasyon:

Bireyler arasında allelik varyasyonun ortaya çıkışı ve zaman içerisinde nasıl değiştiği.



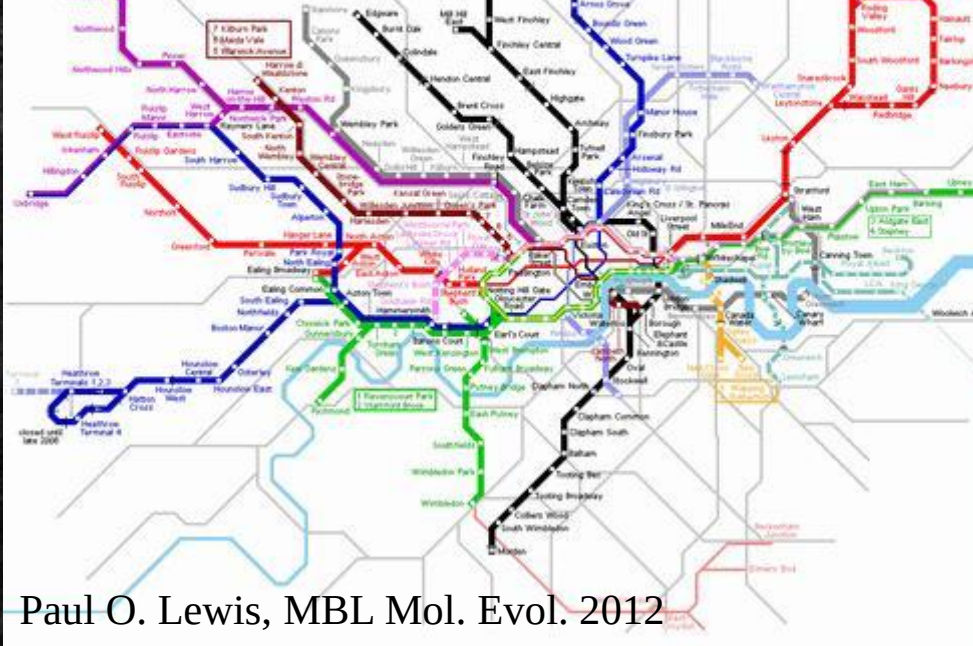
Soy Ağacı:

Bireylerin bir biriyle olan akrabalık ilişkileri, kim kimden geliyor.



Etkin Modelleme ve Basitleştirme

Londra metrosunun oldukça
basitleştirilmiş haritası



Londra metrosunun oldukça
gerçekçi haritası



Varmak istediğiniz yere sizi ulaştıracak güzergahı
bulmanız için hangisi daha *pratik*?



Wright – Fisher Populasyon Modeli

- Bütün bireyler bir kuşak yaşar ve bir sonraki kuşakta yerlerini yavrularına bırakırlar.
- Populasyon büyüklüğü (N) sabittir ve zaman içinde değişmez.
- Üreme rastgeledir ve bütün bireylerin uyumluluğu aynıdır.
- Populasyon büyüklüğü sabit ve üreme rastgele olduğu için bazı bireyler bir sonraki kuşağa yavru bırakmayabilirler.
- Bütün varyasyon nötraldir, genetik soyların kaybı rastgeledir yani genetik sürüklenmeye bağlıdır.
- Varyasyon nötral olduğunda soy oluşturma ile ilgili süreç mutasyon sürecinden bağımsızdır.



Wright – Fisher Populasyon Modeli

Üreme başarısındaki varyasyon allel frekanslarında rastgele salınımlara neden olur.

Allel frekansların denge durumundaki dağılımı aşağıdaki formül ile verilir.

$$p \rightarrow q = \mu \quad q \rightarrow p = \nu$$

$$\Pr(p) = c p^{(4N_e\nu-1)} q^{(4N_e\mu-1)}$$

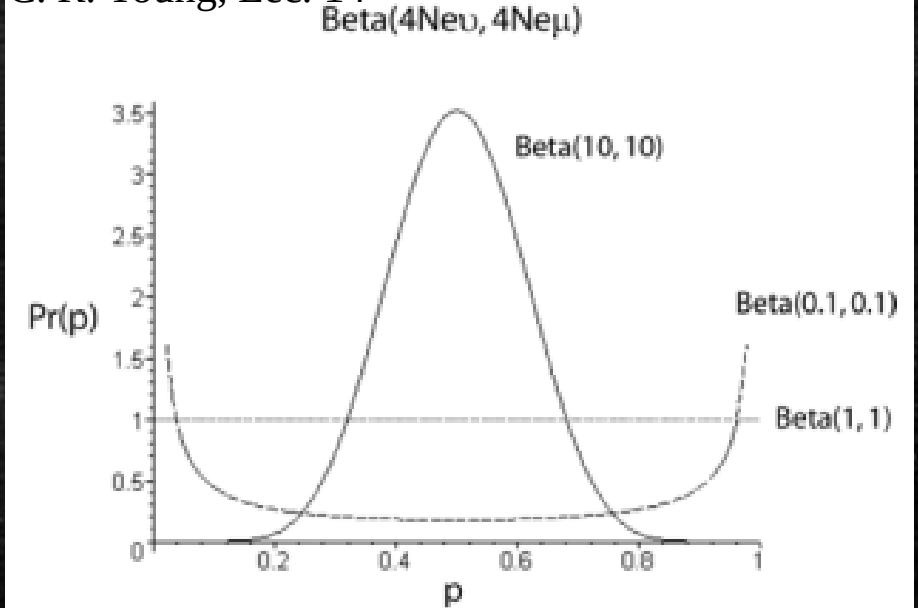
$$E(p) = \frac{4N_e\nu}{4N_e\nu + 4N_e\mu} = \frac{\nu}{\nu + \mu}$$

Beklenen değer (ortalama)
populasyon büyüklüğünden
Bağımsızdır.

$$\Theta = 4N_e\mu$$

(Populasyon büyüklüğüne
oranlı mutasyon parametresi)

C. R. Young, Lec. 14

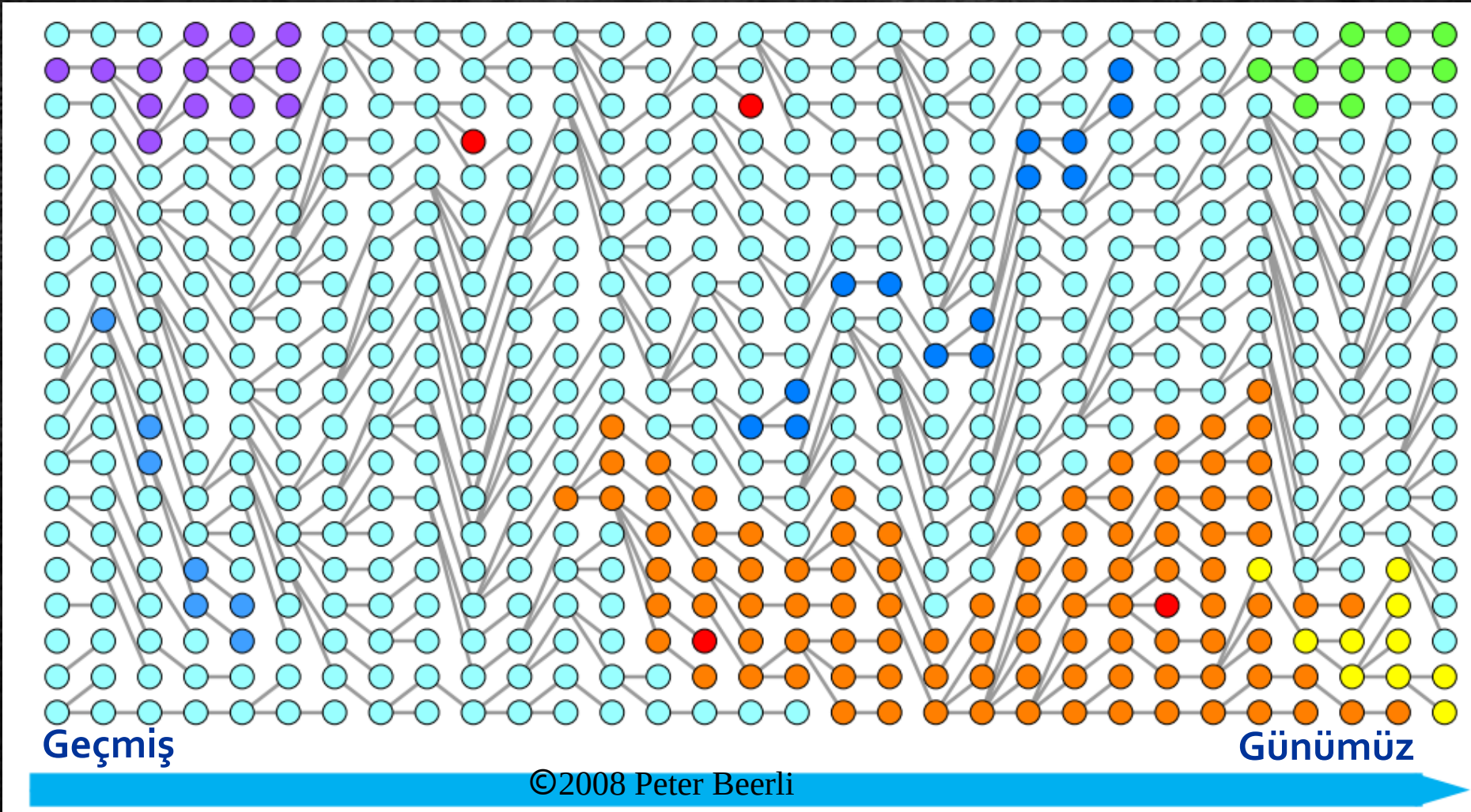


<http://bio.classes.ucsc.edu/bio107>

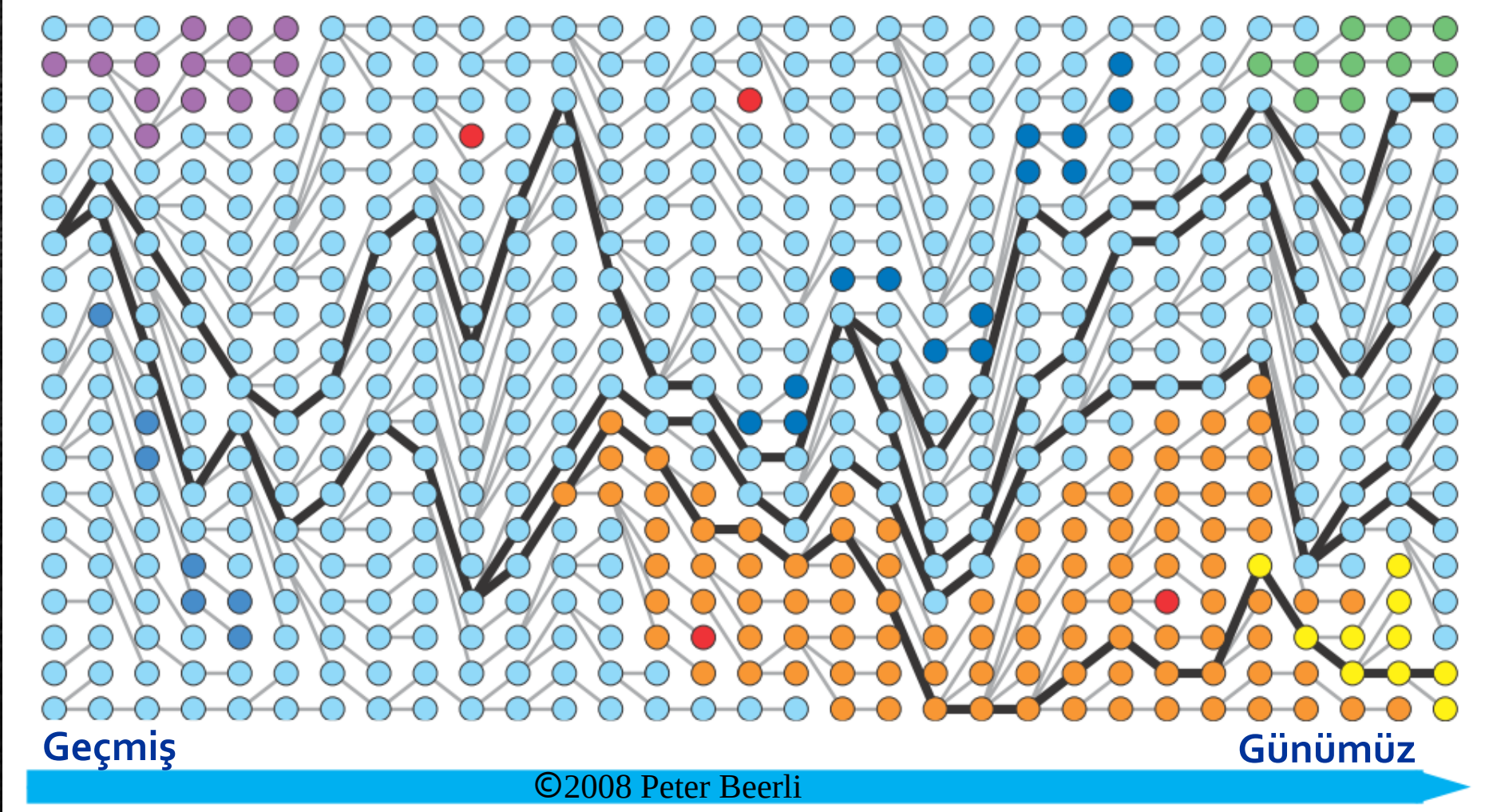
$\Theta \ll 1$ Sürüklenme baskındır.

$\Theta \gg 1$ Mutasyon sabitlenmeyi önler.

Populasyon Modeli = İleri yönlü



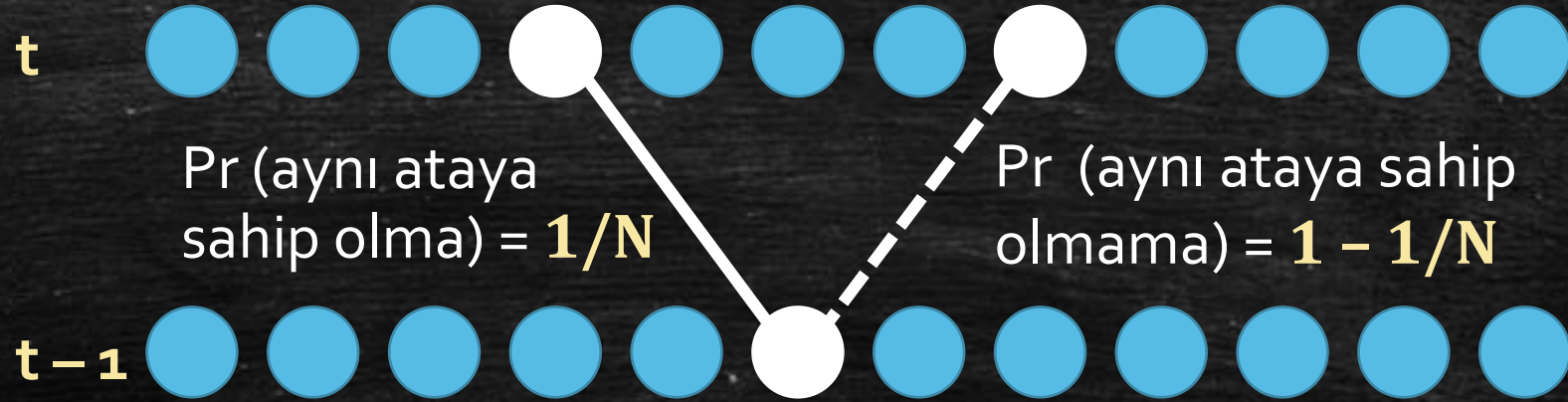
Populasyon Modeli = Geri yönlü



Geri yönlü modelleme ve Birleşme Kuramının Kökenleri



Gustave Malecot rastgele seçilmiş iki bireyin bir kuşak öncesinde ($t - 1$) ortak ataya sahip olma olasılıklarını hesapladı:



İki bireyin ortak
ataya varma zamanı

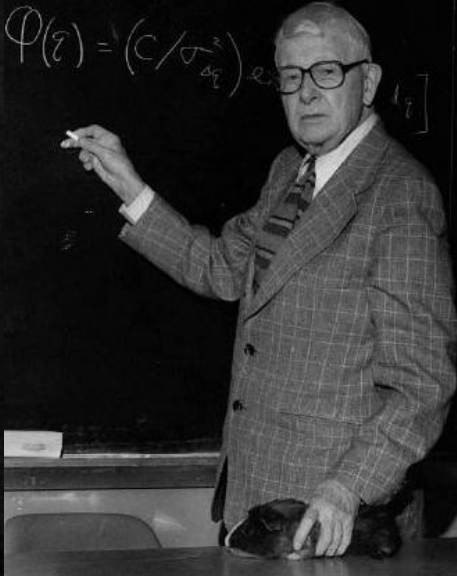
$$\Pr(\tau | N) = \left(1 - \frac{1}{N}\right)^\tau \left(\frac{1}{N}\right)$$

Geometrik Dağılım

Ortalama
bekleme süresi

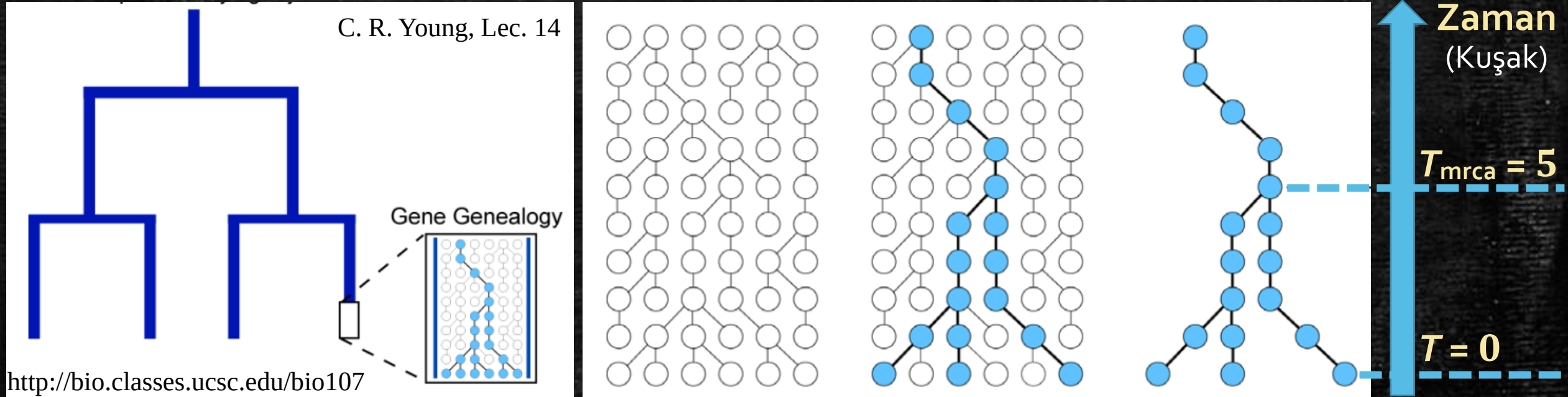
$$E(\tau) = N$$

$$E(T) = 1/p$$



Geriyönlü modelleme, Gen ağaçları ve Birleşme Kuramı

Birleşme Kuramı gen ağaçlarına, **gözlemlediğimiz** veriye dayanır.
(**örneklem üzerinden işler**)

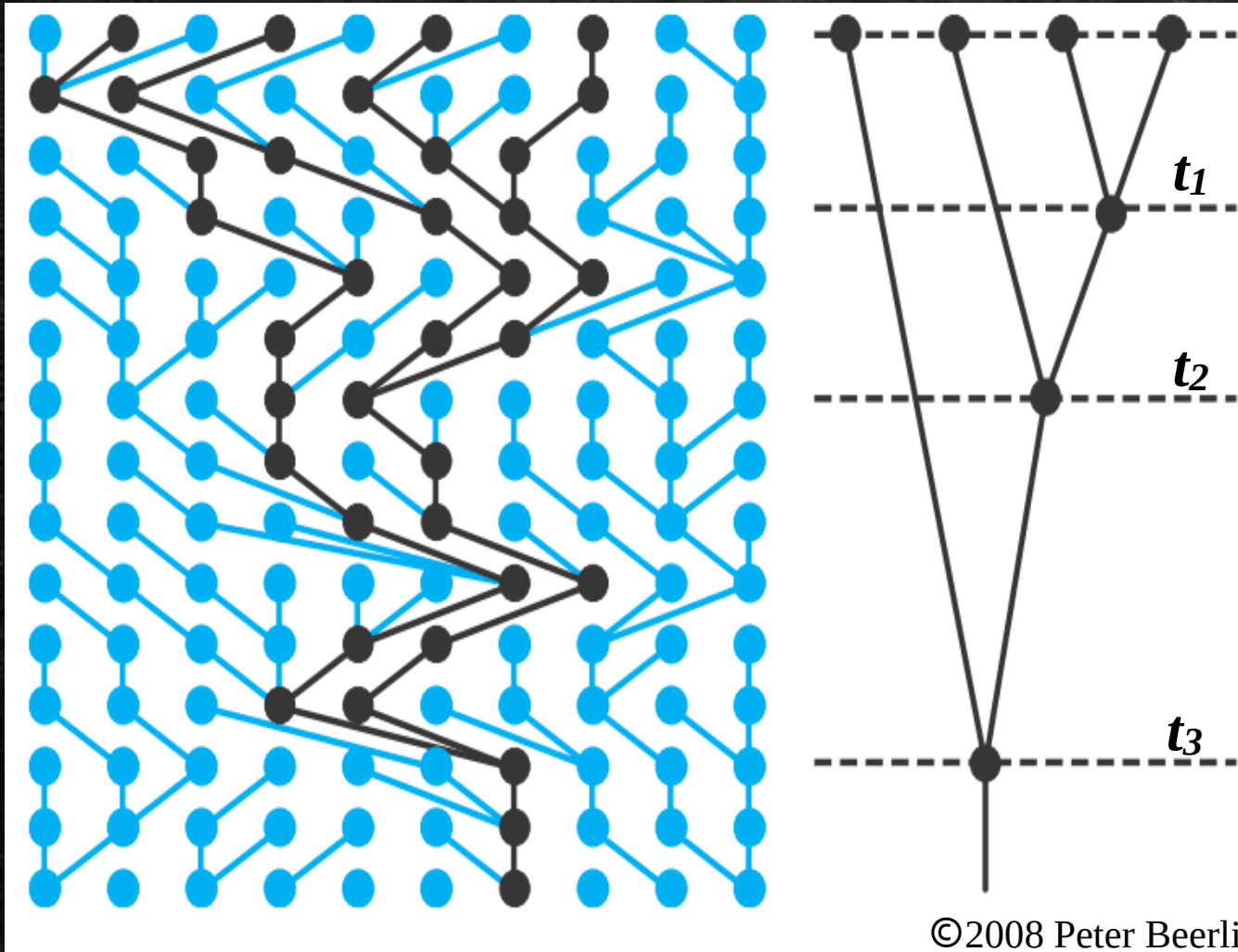


Gen ağaçları iki tip bilgi taşır

1. Ortak ataya varma zamanlar $T(i)$
2. Topoloji (kim kiminle ilişkili)

Ortak Ataya Varma Zamanları

Standart Birleşme Kuramı = Kingman n - birleşme



©2008 Peter Beerli

$$\Pr(2) = \left(1 - \frac{1}{N}\right)$$

$$\Pr(3) = \left(1 - \frac{1}{N}\right) \left(1 - \frac{2}{N}\right)$$

$$\Pr(4) = \left(1 - \frac{1}{N}\right) \left(1 - \frac{2}{N}\right) \left(1 - \frac{3}{N}\right)$$

n bireyin n sayıda
ataya sahip olma olasılığı:

$$\Pr(n) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{N}\right) \approx \frac{n(n-1)}{2N}$$

Ortak Ataya Varma Zamanları

Standart Birleşme Kuramı = Kingman n - birleşme

t inci kuşakta bir birleşme olayının görülme olasılığı:

$$C(t | n) = \Pr(n)' (1 - \Pr(n))$$

$C(t | n)$ Binomiyal Örneklemeyi andırıyor:

$$p^x (1 - p)^y = p' (1 - p)^1$$

Gözlemlediğimiz özel bir tip Binomial Dağılımdır

- $(1 - p)$ nin görülme olasılığı çok düşüktür.
- $(1 - p)$ nin görülmesi için gerekli deneme sayısı $t+1$ çok yüksektir.
- Poisson süreci, **bekleme süresi üssel dağılım** ile modellenir.

$$\Pr(t) = \lambda e^{-\lambda t}$$

Ortak Ataya Varma Zamanları

Standart Birleşme Kuramı = Kingman n - birleşme

$$\Pr(t) = \lambda e^{-\lambda t}$$

Zamanı N kuşak birimleri üzerinden ölçersek, birleşme olayların zamanları ile ilgili genel modellere ulaşırız:

Birleşme olaylarının hızı:

$$\lambda = \frac{n(n-1)}{2N}$$

t anında birleşme görülme olasılığı:

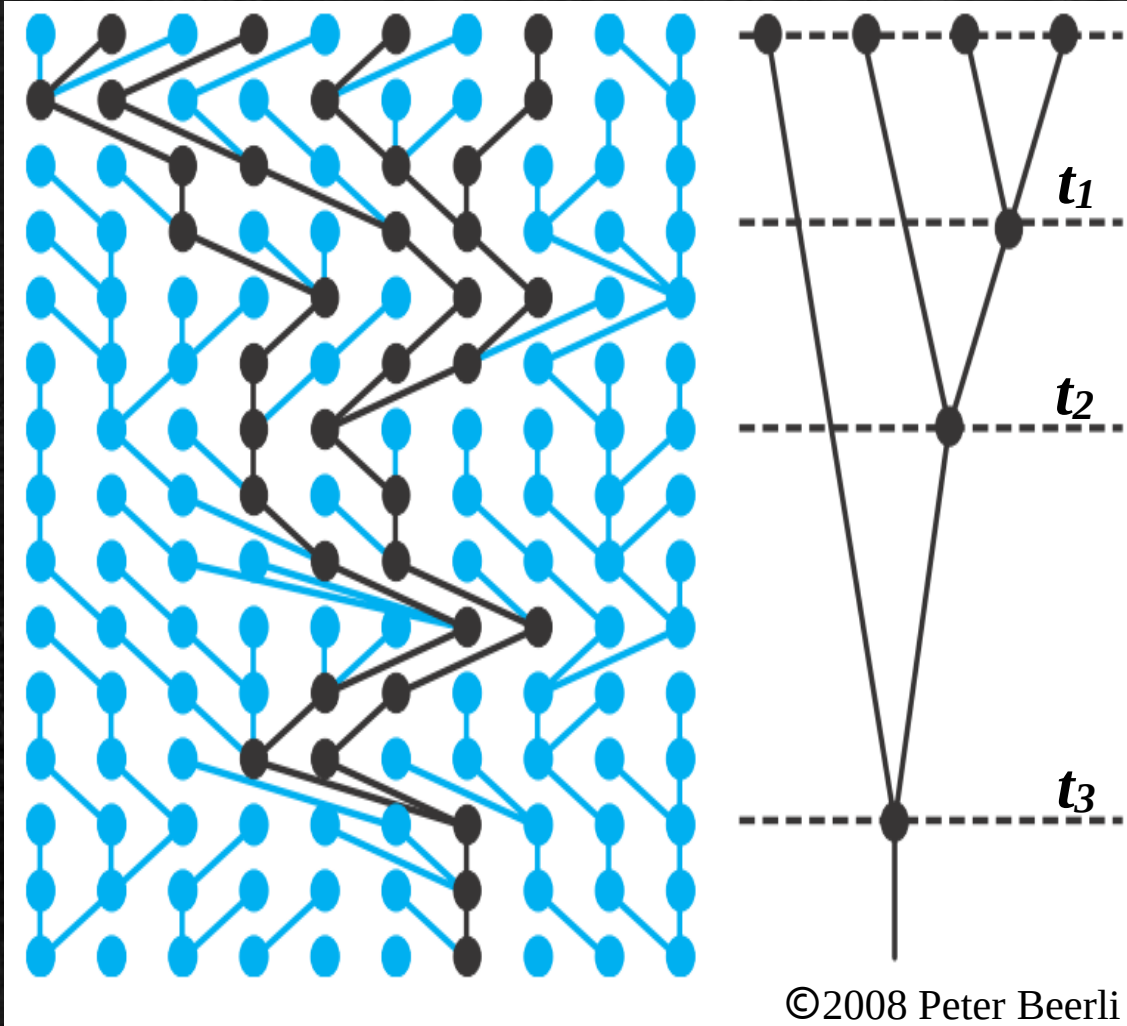
$$C(t | n) = \frac{n(n-1)}{2N} e^{-\frac{n(n-1)}{2N} t}$$

i sayıda soy için birleşme için geçen ortalama zaman:

$$E[T(i)] = \frac{2N}{i(i-1)}$$

Ortak Ataya Varma Zamanları

Standart Birleşme Kuramı = Kingman n - birleşme



Artık ağacımız ile ilgili her birleşme olayını ve ağacın toplam boyunu (ortak ataya varma zamanını, T_{mrca}) hesaplayabiliriz:

$$P(t_j|N) = \lambda e^{-\lambda t}$$

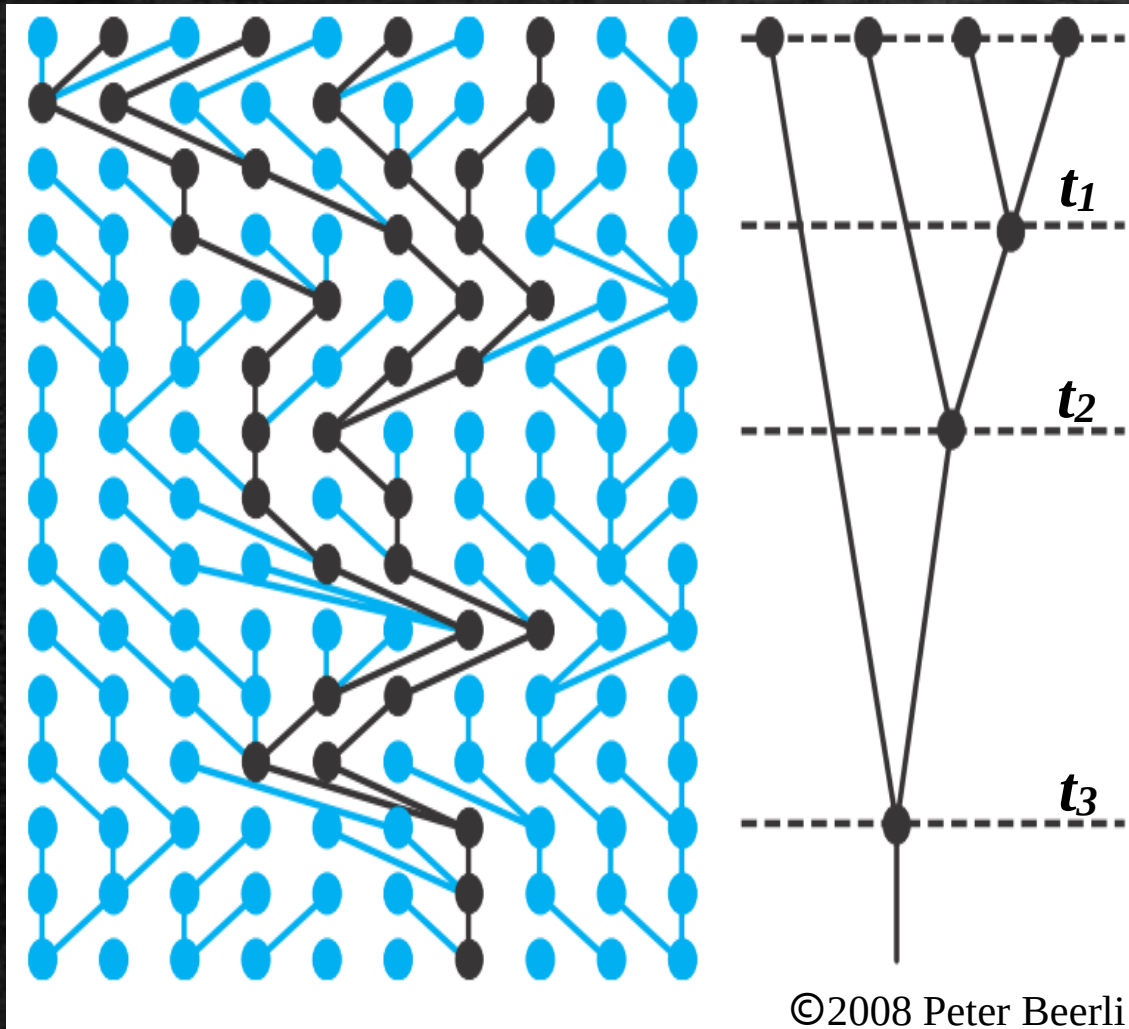
t_i zamanında birleşme olayının olasılığını gösterir.

Fakat hangi iki bireyin birleştiğini bilmiyoruz dolayısıyla formüle rastgele iki birey seçmenin olasılığını ekleriz:

$$P(t_j|N, i_1, i_2) = P(t_j|N) \frac{1}{\binom{n}{2}}$$

Ortak Ataya Varma Zamanları

Standart Birleşme Kuramı = Kingman n - birleşme



$$P(t_j | N, i_1, i_2) = P(t_j | N) \frac{1}{\binom{n}{2}}$$



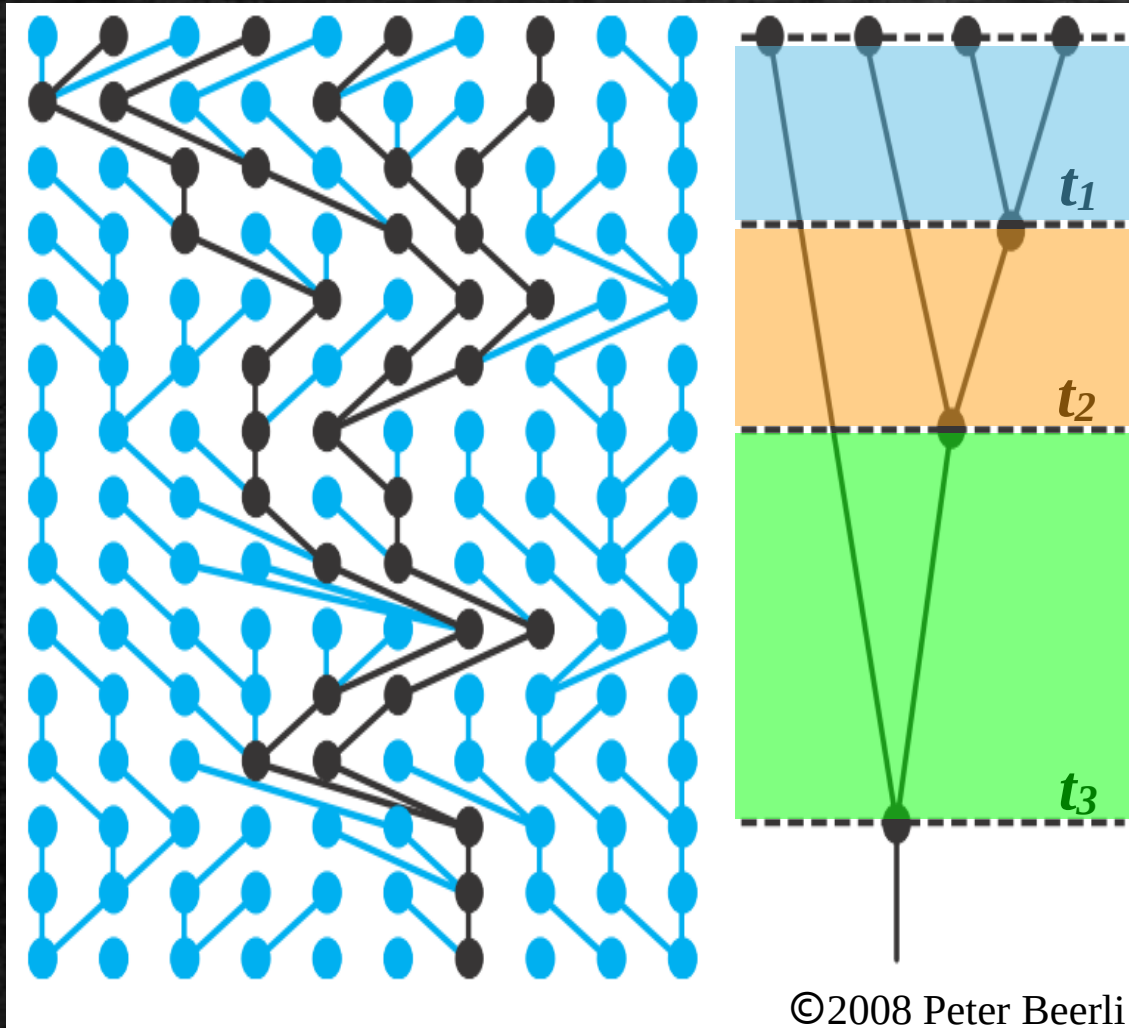
$$P(t_j | N, i_1, i_2) = \left[\frac{n(n-1)}{2N} e^{-t_j \frac{n(n-1)}{2N}} \right] \frac{2}{n(n-1)}$$



$$P(t_j | N, i_1, i_2) = e^{-t_j \frac{n(n-1)}{2N}} \frac{2}{2N}$$

Ortak Ataya Varma Zamanları

Standart Birleşme Kuramı = Kingman n - birleşme



Artık bütün ağacımız olasılığını (**Geneology, G**) ve ağacın toplam boyunu (ortak ataya varma zamanını, T_{mrca}) hesaplayabiliriz. Her bir birleşmenin diğerlerinden bağımsız olduğunu varsayarsak:

$$P(G | N) = P(t_1 | N, i_1, i_2)$$

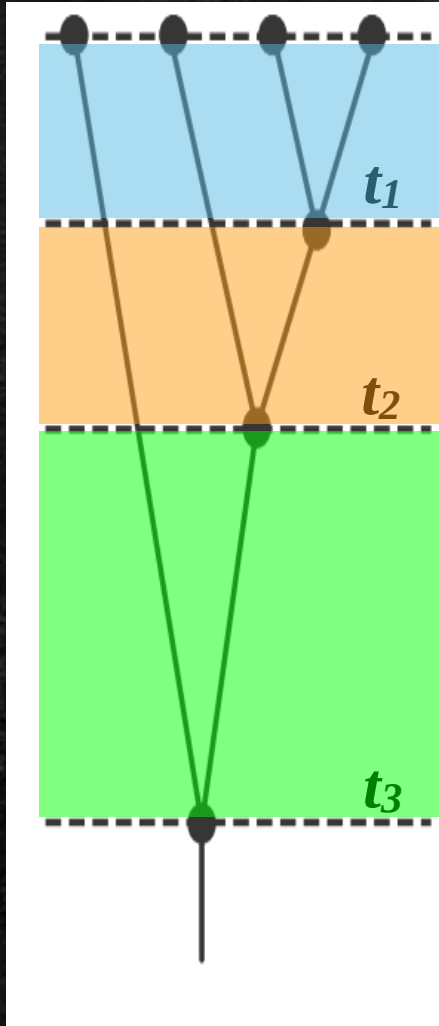
$$\times P(t_2 | N, i_{1,2}, i_3)$$

$$\times P(t_3 | N, i_{1,2,3}, i_4)$$

$$P(G|N) = \prod_{j=0}^T e^{-t_j \frac{n_j(n_j-1)}{2N}} \frac{2}{2N}$$

Ortak Ataya Varma Zamanları

Standart Birleşme Kuramı = Kingman n - birleşme



$$P(G|N) = \prod_{j=0}^T e^{-t_j \frac{n_j(n_j-1)}{2N}} \frac{2}{2N}$$

Bu olasılığın beklenen değeri (ortalaması) her bir birleşme aralığının beklenen değerlerinin toplamıdır.

Her aralığın beklenen değeri:

$$E(t) = \frac{2N}{n(n-1)}$$

Buradan en yakın ortak ataya olan ortalama bekleme süresi (T_{mrca}) hesaplanabilir:

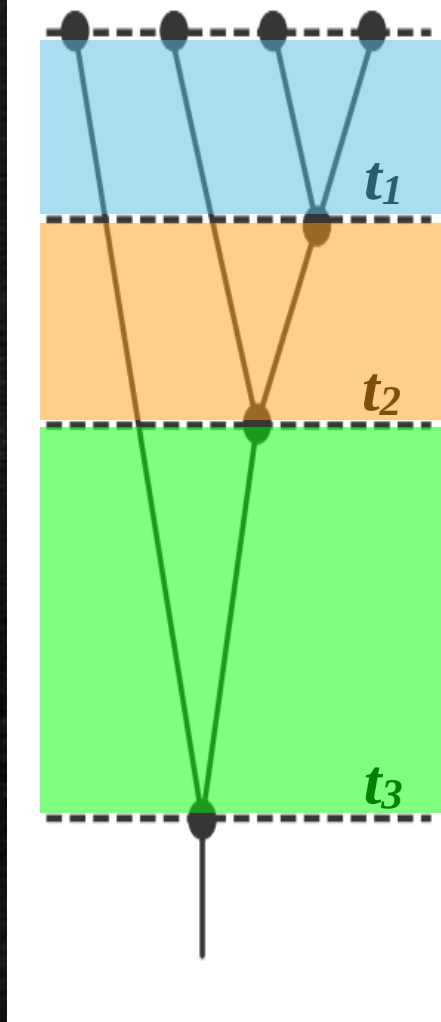
$$E(T_{mrca}) = \sum_{j=0}^j \frac{2N}{n_j(n_j-1)}$$

$$\lim_{n \rightarrow \infty} E(T_{mrca}) = N + \frac{N}{3} + \frac{N}{6} + \frac{N}{10} + \dots = 2N$$

$$\lim_{n \rightarrow \infty} \sigma(T_{mrca}) = 2N$$

Ortak Ataya Varma Zamanları

Standart Birleşme Kuramı = Kingman n - birleşme



$$\lim_{n \rightarrow \infty} \sigma(T_{mrca}) = 2N$$

ortak ataya varma
zamanı doğrudan
populasyon
büyükliğüne bağlıdır!!

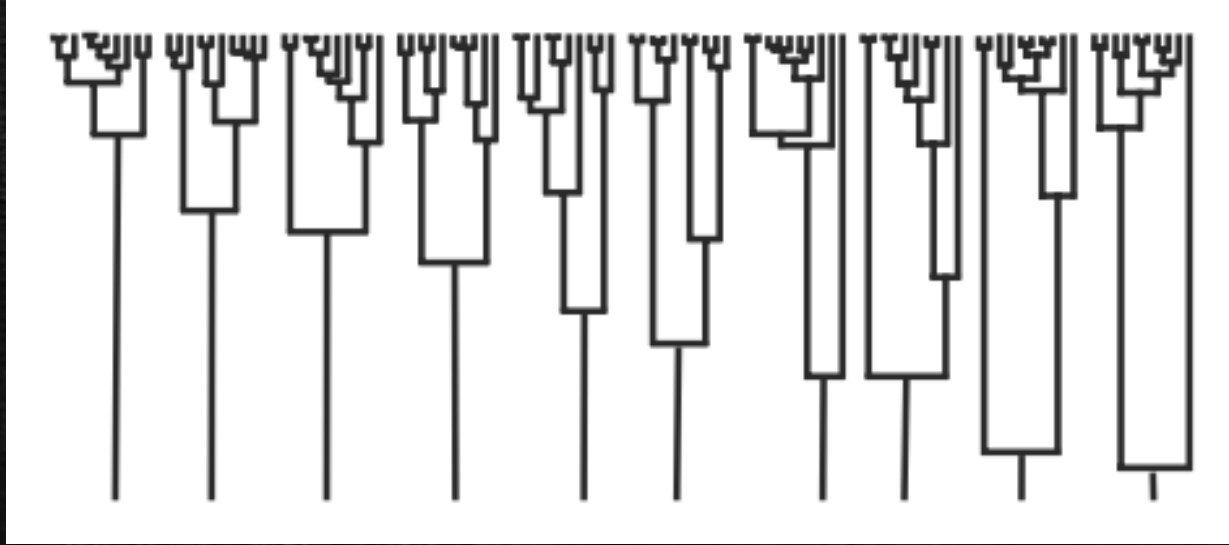
i sayıda bireyin ortalama birleşme zamanları $E[T(i)/N]$

i	$N = 100$	$N = 200$	$N = 1000$
5	10	20	100
4	17	33	167
3	33	67	333
2	100	200	1000

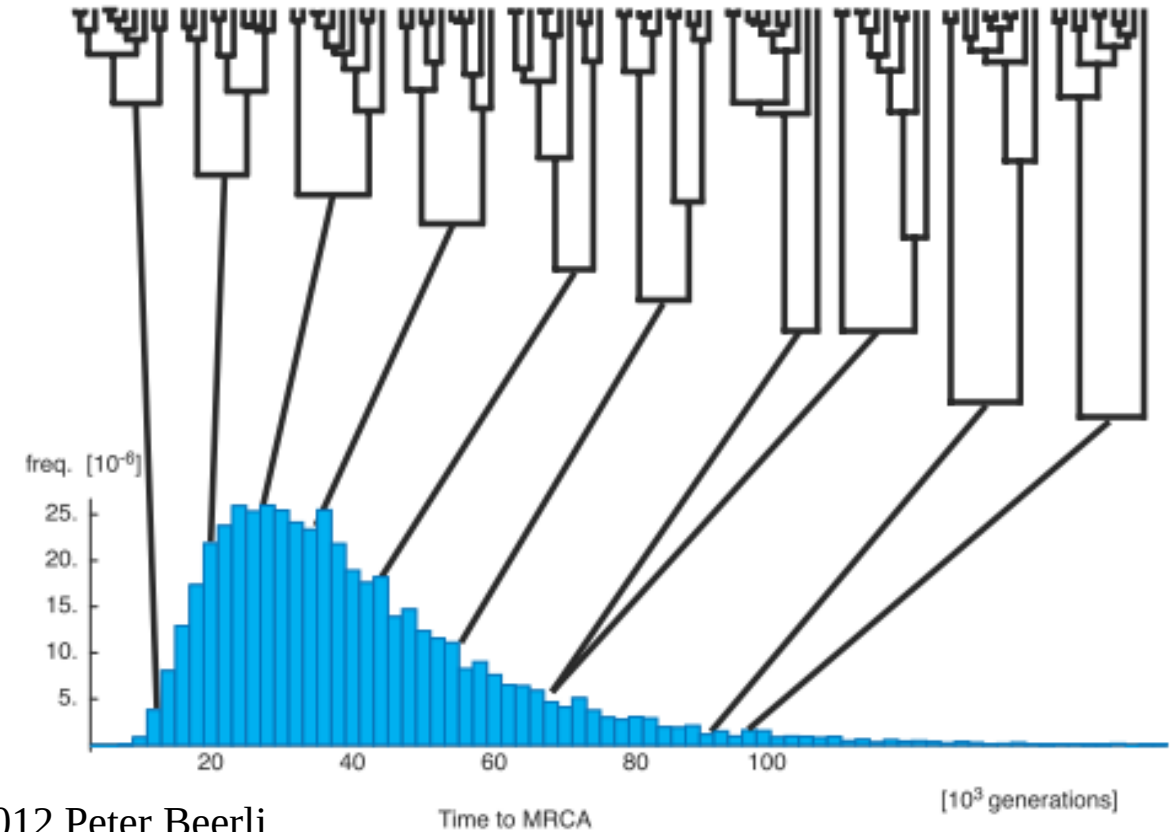
Çizelgeye bakarak 3 temel bilgiye ulaşabiliriz:

1. Birleşme zamanları (kuşal olarak ölçülür) küçük N için daha kısadır.
2. Çok sayıda birey olduğunda ($n \gg 2$) birleşme az sayıda bireyin olduğu popülasyonlara oranla çok hızlı gerçekleşir.
3. Populasyon büyüklüğü gen ağacının toplam yüksekliğini etkiler, birleşme olaylarının görece zamanları populasyon büyüklüğünden bağımsızdır.

Birleşme Sürecinde Stokastisite



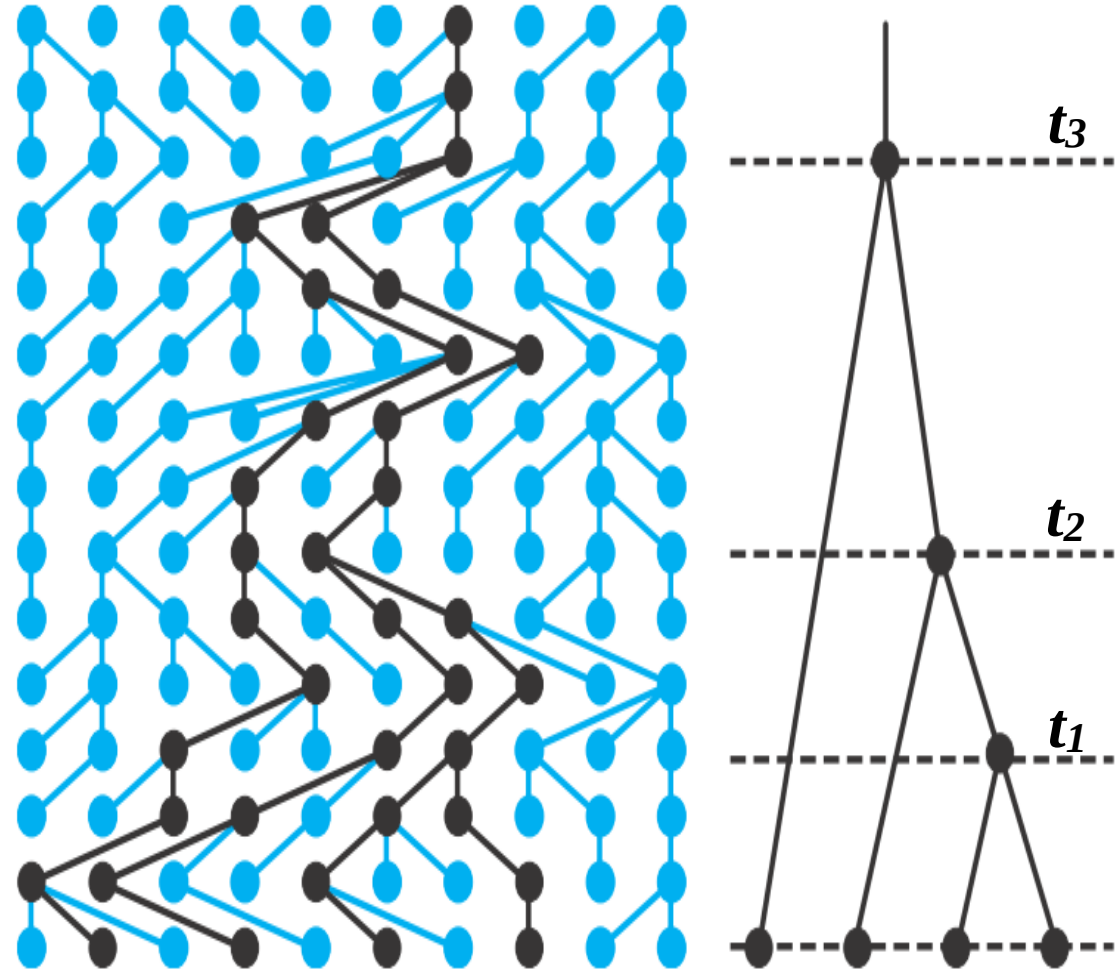
Bütün Gen ağaçları $N_e = 10,000$ ile simüle edilmiştir



©2012 Peter Beerli

Birleşme Kuramı: Topoloji

©2008 Peter Beerli

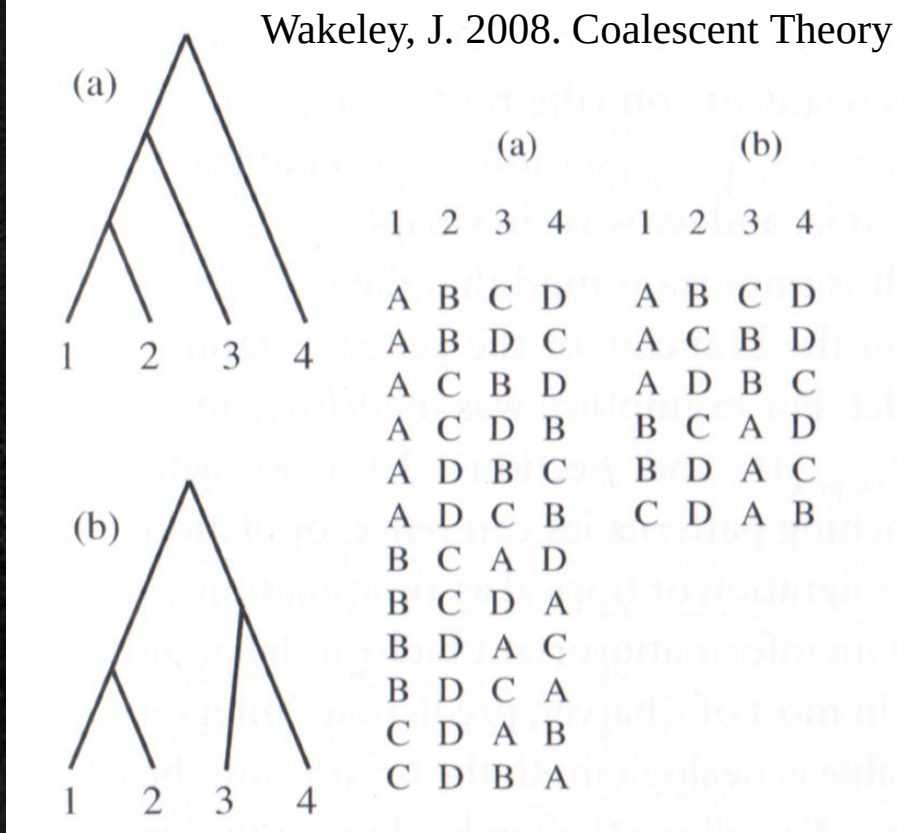


Birleşme kuramının geriye yönelik sürecini düşündüğümüzde gen ağaçlarını rastgele birleşme gösteren ikili ağaçlar olarak düşünebiliriz.

Ağaç yapısının krulmasında rastgelelik gerçek hayatta kimin kimle akraba olduğunu bilmememizden ileri gelir.

Birleşme Kuramı: Topoloji

$n = 4$ 'lük bir örneklem için olası bütün gen ağaçları



n büyüklüğünde bir örneklem için olası bütün ağaç yapıları her adımdaki birleşme olasılıkları kullanılarak hesaplanabilir.

$$\prod_{i=2}^n \binom{i}{2} = \frac{n! (n-1)!}{2^{n-1}}$$

Örneklem büyüklüğü arttıkça taranması gereken ağaç uzamı çok yüksek sayılara ulaşır.

n	Olası AGaç sayısı (Kölendirilmiş)
2	1
4	18
8	1,587,600
10	2,571,912,000
100	1.37×10^{284}
1000	3.02×10^{4831}

Birleşme Kuramı: Mutasyon

- Elimizdeki veri, soylar değil genetik varyantlardır!!!
- Genetik varyantlar arasında akrabalık ilişkileri kurmak için mutasyonları kullanırız.
- Peki mutasyon süreci (modelleri) birleşme kuramına nasıl eklenir?

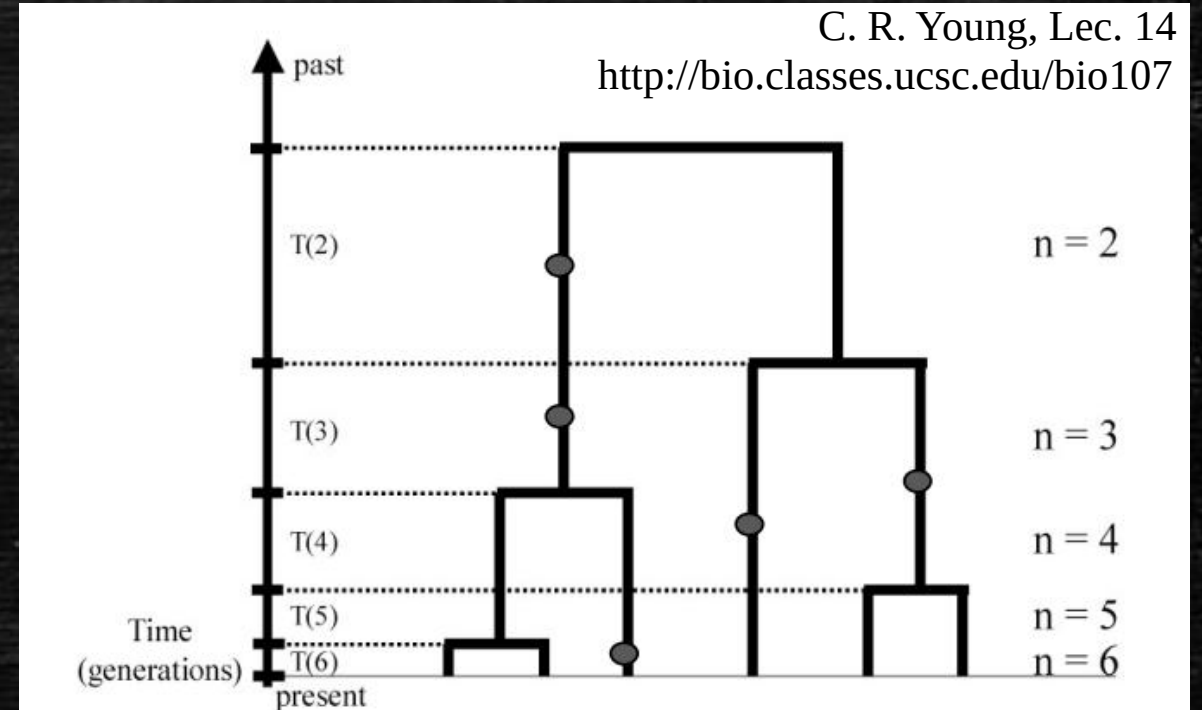
Awillemsei	ATGTTAAAGTATTTATTTATAATTATTTTTATGATTCCTCCTTGTTTTTTTGTT---AATTGTTGGTGGTTAGTT
CitalicusND4	ATGTTAAAGTATTTATTTATAACTATTTTTTTGATCCCTCCTTGTTTATTGAAA---AATTGTTGATGGTTGGTC
ChrtchinND4	ATGTTAAGTTATTTGTTTATGATTAGTTTTTTAATCCCTCCTTGTTTTTTATAT---AATTGTTGATGATTAAAT
EchrfusND4	ATGTTAAGGTATTTGATAATAATTGTTTTCTTGACCCCTCCTTGTTTTTTATAT---GGTTGTTGATGGTTGATT
LMIGRATND4	ATGCTAAAGTACTTATTTATAATTATTTTTTTGATCCCTATTTGTTTATTAGGT---AATTGTTGATGATTAGTT
OchnensND4	ATGTTAAGTTATTTATTTATGATTGTTTTTTTGATTCCTCCTTGTTTATTAAAT---AATTCCTTGATGATTGGTT
PALBONMND4	ATGTTAAAGTATTTATTTATGATTATTTTTTTGATCCCTCCTTGTTTACTAAAA---AATTGTTGGTGGTTGGTT
SGREGND4	ATGTTAAAGTATTTATTTATAACTATTTTTTTGATCCCTCCTTGTTTATTAAAT---AATTGTTGATGGTTGGTT
AsinensisND4	ATGTTAAGGTATTTATTTGTTCTTCTTTTTATGATCCCTCCTTCTTTAGTTGAT---TGTTTATGGTGAATTTTT
MxznGnsND4	ATGCTAAAGTACTTATTTATAATTATTTTTTTGATCCCTCTCTCTTCCCTTTGA---TGTTTTTGATGAATTTGTT
MMANNIND4	ATGTTAAGGATTATTTTAATAATTAGTATTACGATCCCATTAATGTTTGGTGAGA---ATAGGGTGATGGTTGGTT
TEMMAND4	ATGTTAAATTTATTTTATTAATAATGTTTATGATCCCTCCTTCTTTTTTAAATGGAAATTTTTGATGGGTGGTT

Birleşme Kuramı: Mutasyon

- Mutasyonları nötral kabul ettiğimiz için mutasyon süreci ile birleşme sürecini bağımsız Poisson süreçleri olarak kabul edebiliriz.
- Mutasyonlar dal uzunluklarına orantılı bir şekilde, dallar üzerine rastgele sepiştirilirler.

Toplam ağaç uzunluğu T , her bir birleşme aralığının uzunluğunun $T(i)$ o uzunluğa sahip soy sayısı, i , ile çarpımlarının toplamı olarak ifade edilebilir.

$$T = \sum_{i=2}^n iT(i)$$



Birleşme Kuramı: Theta (θ)'nın tahminlenmesi

Kuşak başına mutasyon oranını sonsuz-bölge modeli atında μ olarak kabul edersek ve $\theta = 2N\mu$ olarak tanımlarsak.

Beklenen polimorfik bölge sayısı

$$E[S] = N\mu T$$

=

$$E[S] = \frac{\theta}{2} \sum_{i=2}^n iT(i)$$

$$E[S] = \frac{\theta}{2} \sum_{i=2}^n i \frac{2}{i(i-1)}$$

$$E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

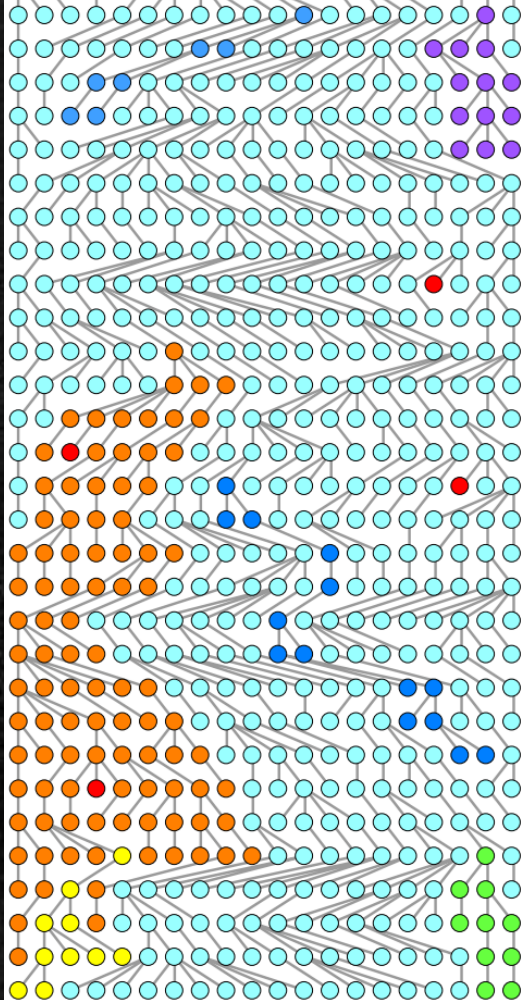
=

$$\hat{\theta}_s = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

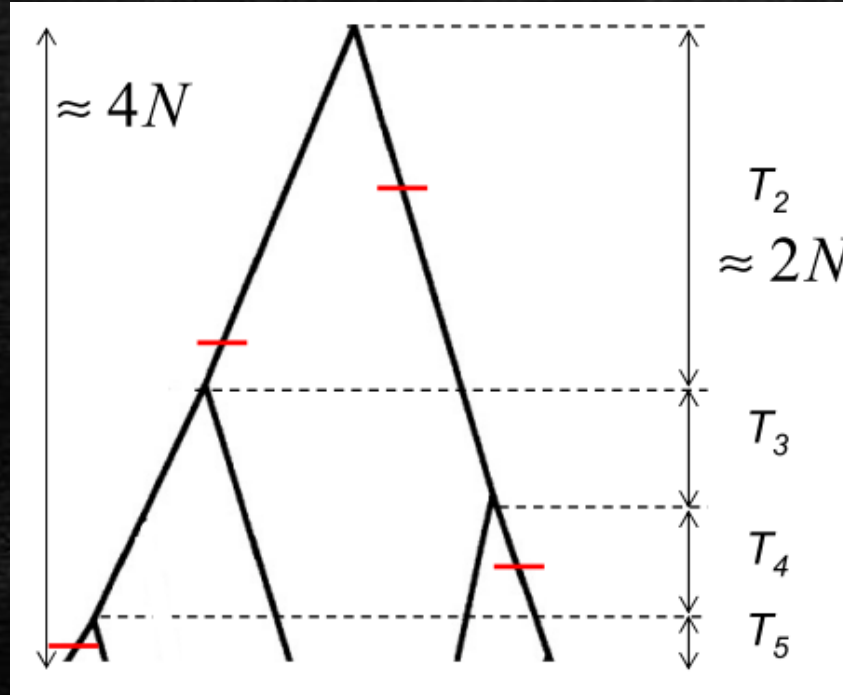
Watterson's
theta

Standard Birleşme Kuramı

©2008 Peter Beerli



Awillemsei	ATGTTAAAGTATTTATTTATAA
CitalicusND4	ATGTTAAAGTATTTATTTATAA
ChrtchinND4	ATGTTAAGTTATTTGTTTATGA
EchrfusND4	ATGTTAAGGTATTTGATAATAA
LMIGRATND4	ATGCTAAAGTACTTATTTATAA
OchnensND4	ATGTTAAGTTATTTATTTATGA



$$E[T(i)] = \frac{2N}{i(i-1)}$$

$$E(T_{mrca}) = \sum_{j=0}^j \frac{2N}{n_j (n_j - 1)}$$

$$E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

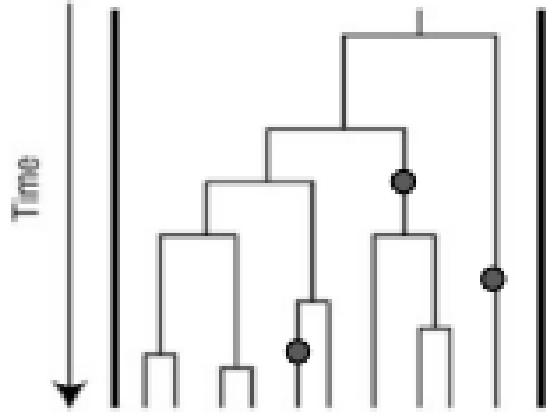
Ağaç oluşturma
Parametre tahminleme
Simülasyon

Birleşme Kuramının Uzantıları

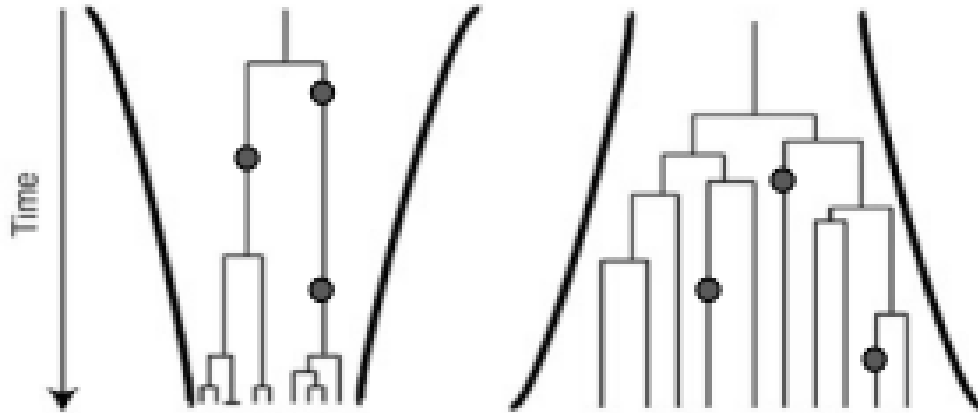
- Standard birleşme kuramı bir çok basitleştirmenin kabul edildiği idealize Wrigth-Fisher populasyon modeli üzerine kurgulanmıştır.
- Son yıllarda yapılan bir çok çalışma WF modelinde kabul edilen varsayımların gevşetilebilmesine neden olmuştur.
- **Gevşetilen varsayımlar arasında**
 - 1.Populasyon büyüklüğünde salınımlar.
 - 2.Göç/izolasyon modelleri (yapılı birleşme).
 - 3.Rekombinasyon (rekombinasyon yolları).
 - 4.Seçilim (atasal seçilim yolları).
 - 5.Metapopulasyonlar (yok olma/tekrar işgal etme).

Birleşme Kuramının Uzantıları: Üssel Büyüme

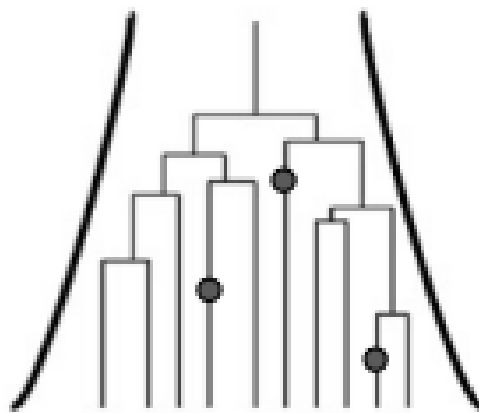
C. R. Young, Lec. 14 <http://bio.classes.ucsc.edu/bio107>



(a) Constant



(b) Declining



(c) Expanding

Populasyonlar nadiren zaman için sabittirler.

Populasyon büyümesi üssel olarak modellenenebilir:

$$\frac{dN}{dt} = rN \quad N_t = N_0 e^{-rt}$$

Populasyon büyüklüğündeki değişimler gen ağaçlarının **şeklinin** değişmesine neden olur.

Büyüyen Populasyonlar

- Uzun terminal dallara.
- Artık sayıda tekli mutasyon.

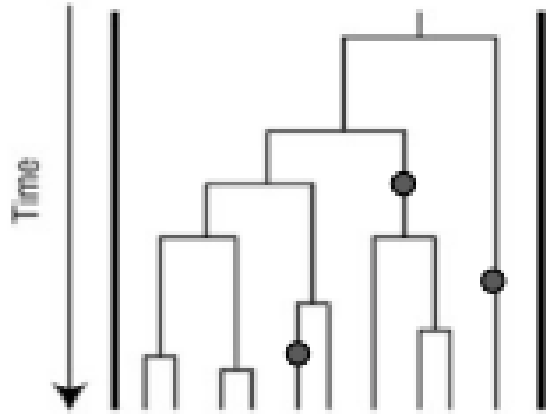
$$D = \frac{\hat{\theta}_{\pi} - \hat{\theta}_S}{\sqrt{\text{Var}(\hat{\theta}_{\pi} - \hat{\theta}_S)}}$$

Küçülen Populasyonlar:

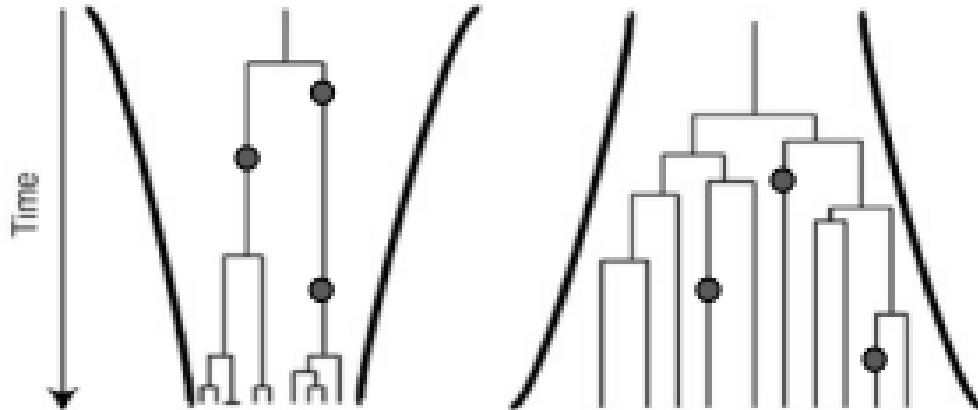
- Uzun iç dallar.
- Tekli mutasyon sayısında azalma.

Birleşme Kuramının Uzantıları: Üssel Büyüme

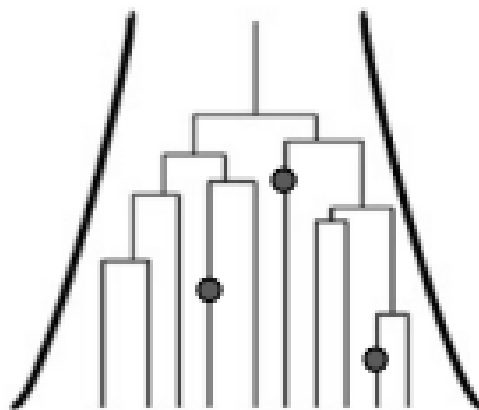
C. R. Young, Lec. 14 <http://bio.classes.ucsc.edu/bio107>



(a) Constant



(b) Declining



(c) Expanding

Ağaçlarındaki bu sinyalleri kullanarak üssel büyüme hızını (g), populasyon büyüklüğü (θ) ile beraber hesaplayabiliriz.

Sabit Populasyon

=

$$P(G|\theta) = \prod_j e^{-t_j \frac{n(n-1)}{\theta}} \frac{2}{\theta}$$

$$N_t = N_0 e^{-rt}$$

=

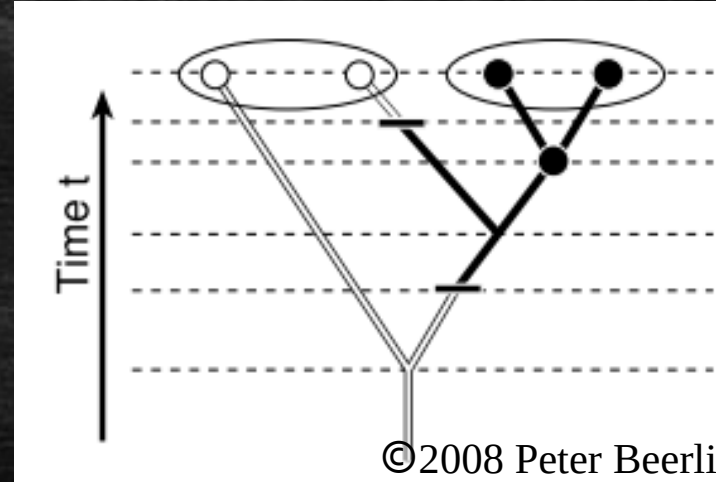
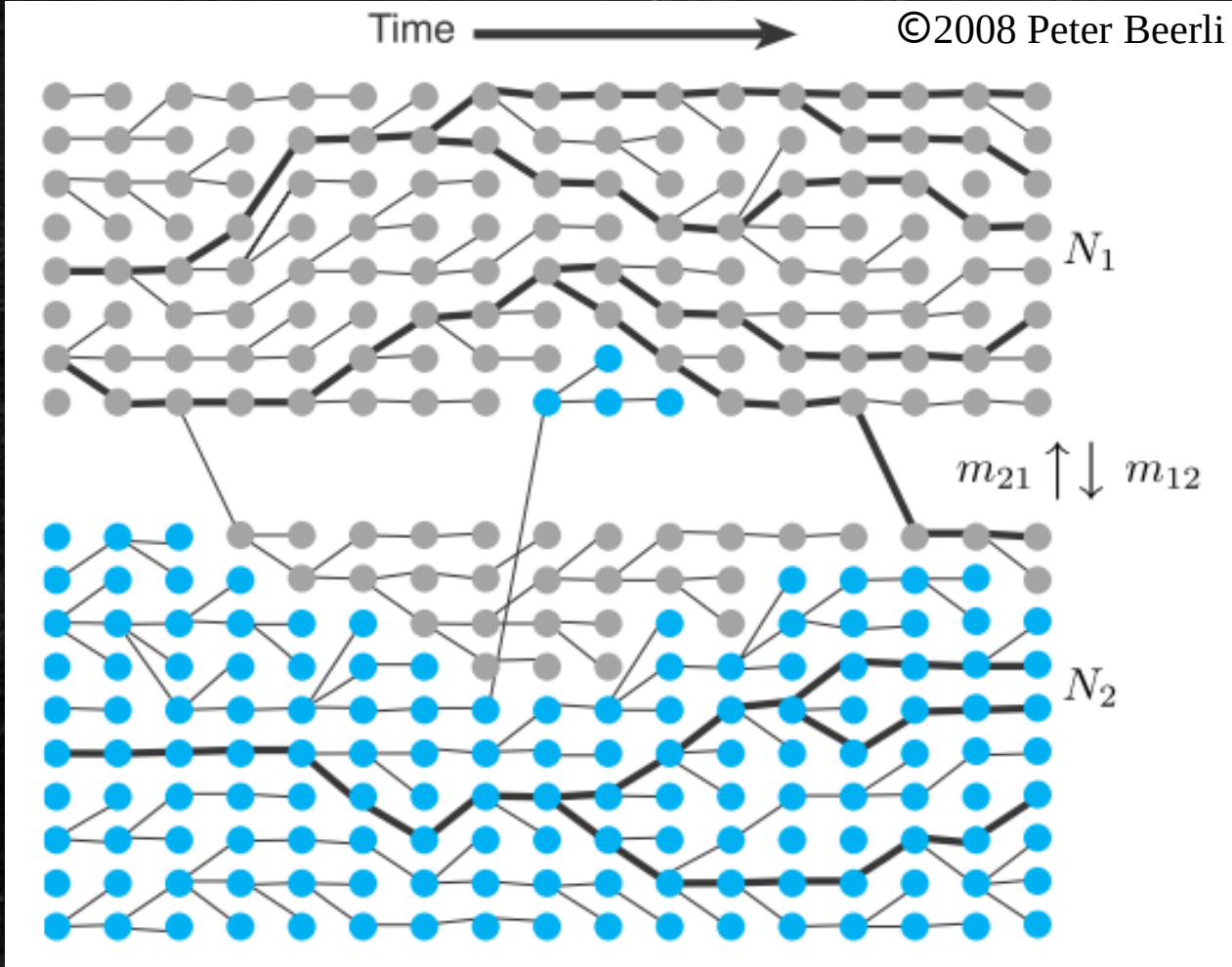
$$\theta_t = \theta_0 e^{-\left(\frac{r}{\mu}\right)t}$$

Değişen
Populasyon

=

$$P(G|\theta_0, g) = \prod_j e^{-(t_j - t_{j-1}) \frac{n(n-1)}{\theta_0 e^{-gt}}} \frac{2}{\theta_0 e^{-gt}}$$

Birleşme Kuramının Uzantıları: Göç



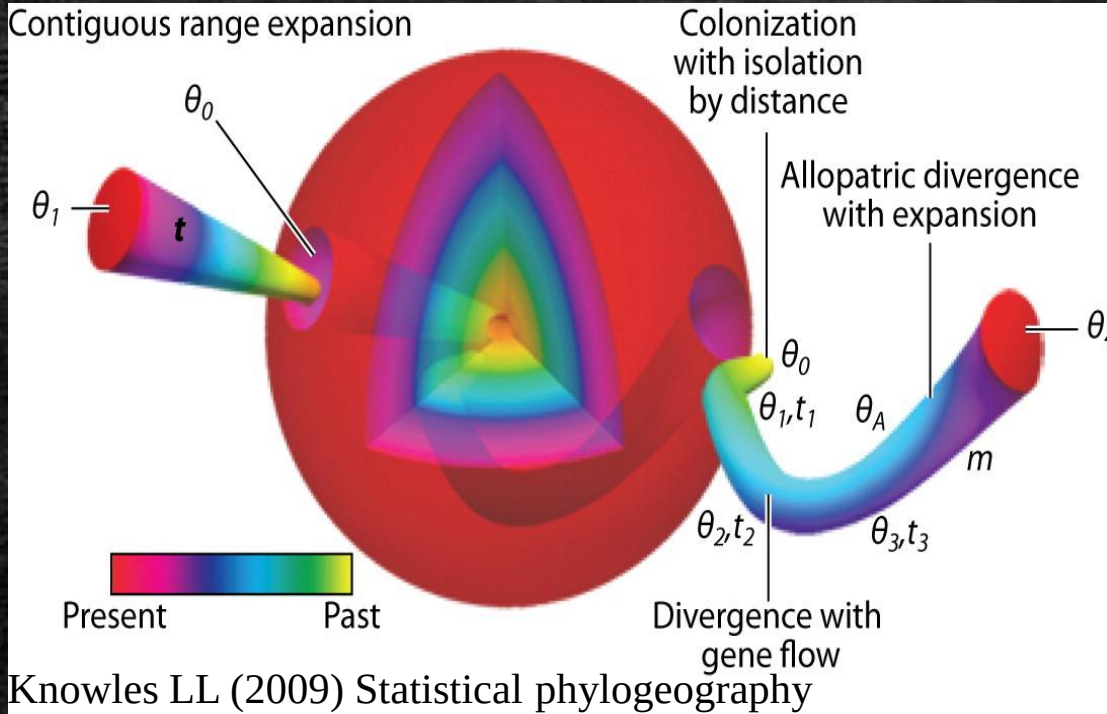
Nokta
olasılık

Bekleme zamanı

$$P(G|N, m) = \prod_{t=1}^T \exp \left(-u_t \left[\sum_i^d \left(\frac{k_{ti}[k_{ti} - 1]}{4N_i} \right) + k_{ti} \sum_{j \neq i}^d m_{ji} \right] \right) \delta_t$$

$$\delta_t = \begin{cases} \frac{1}{2N_i} & \text{\textit{i}. populasyonda birleşme olayı} \\ m_{ij} & \text{\textit{j} den \textit{i}'ye göçü tanımlar} \end{cases}$$

Birleşme Kuramı altında çözümleme



- Demografik süreçlerin genetik varyasyon üzerindeki etkisini modelleyebilmek için son derece kuvvetli bir yöntem.
- Temel amaç ağaç çizmek değil, görülen ağaç ile ilişkili **parametreleri tahminlemektir**.
- Genetik örüntüleri tanımlamaktan çok bu örüntülere neden olan süreçler ortaya koyar.

Çözümlemek istediğimiz olgu gözlemlediğimiz verinin (genetik varyantların) hangi tarihsel, demografik ve evrimsel süreçler sonucunda ortaya çıktığıdır.

Birleşme Kuramı altında çözümleme

Awillemsei	ATGTTAAAGTATTTATTTATAATTATTTTATG
Citalicus	ATGTTAAAGTATTTATTTATAACTATTTTTTTG
Chrtchin	ATGTTAAGTTATTTGTTTATGATTAGTTTTTTA
Echrfus	ATGTTAAGGTATTTGATAATAATTGTTTTCTTG
LMIGRAT	ATGCTAAAGTACTTATTTATAATTATTTTTTTG
Ochnens	ATGTTAAGTTATTTATTTATGATTGTTTTTTT
PALBONM	ATGTTAAAGTATTTATTTATGATTATTTTTTTG
SGREG	ATGTTAAAGTATTTATTTATAACTATTTTTTTG
Asinensis	ATGTTAAGGTATTTATTTGTTCTTCTTTTTATG
MxznGns	ATGCTAAAGTACTTATTTATAATTATTTTTTTG
MMANNI	ATGTTAAGGATTATTTTAAATATTAGTATTTACG
TEMMA	ATGTTAAAATTTATTTTATTATTAATGTTTATG

Akrabalık ilişkileri (gen ağacı).

Dal uzunlukları.

Birleşme zamanları (Farklılaşma zamanları).

Ağacın toplam boyu (Tmrca).

Büyüme/Küçülme var mı? (g)

Populasyonlar arası göç var mı? (m)

Eldeki veri altında modele ilişkin parametrelerin olasılıklarını tahmin etmek istiyoruz

$$P(\text{model} \mid X)$$

Birleşme Kuramı ve Olabilirlik (Likelihood)

$P(\text{model} | X)$ Nasıl hesaplanır?

Birleşme: Populasyon genetik süreçlerini tanımlar (m, g, Tdiv vs.).

Mutasyon modeli: Genetik varyantların zaman içinde değişimini tanımlar.

$P(\text{model} | X)$ 'in *Ardıl Dağılımını* Bayes kuralını kullanarak hesaplarız.

$$p(\theta | D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

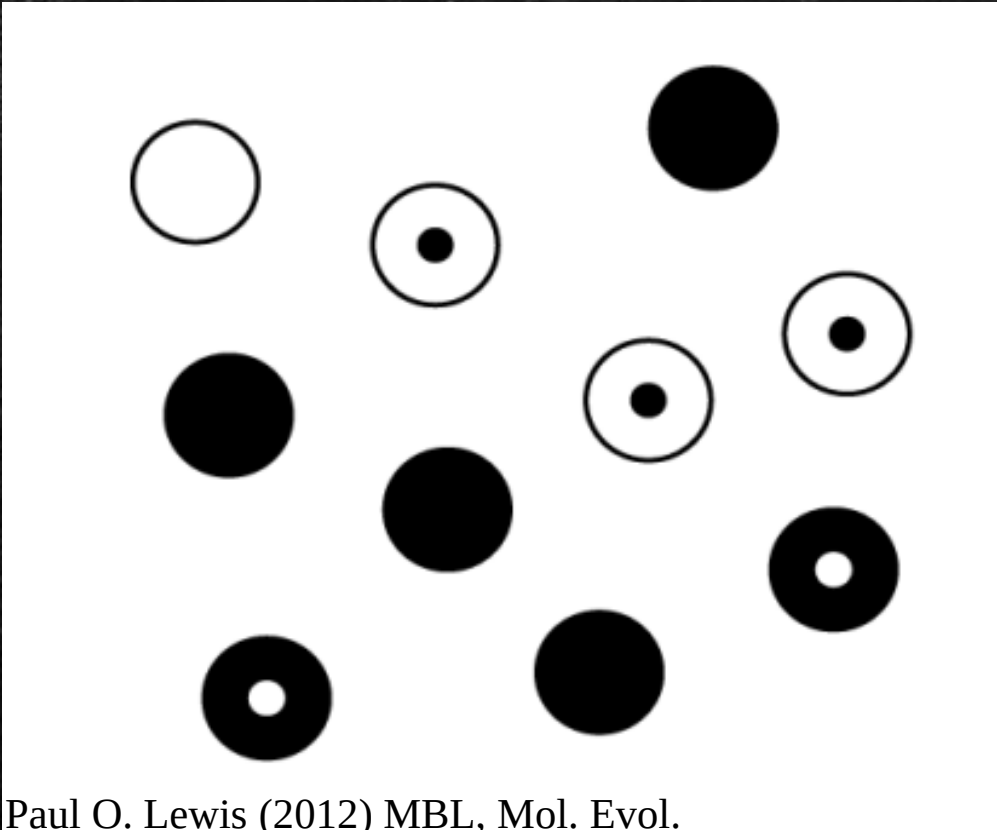
$$p(D | \theta) = \begin{cases} \text{parametrelerin olabilirliği (likelihood).} \\ \text{verinin parametreler altında olasılığı.} \end{cases}$$



Veri altında parametrelerin olasılığı.

Birleşme Kuramı ve Bayes Kuralı

Soru: Noktalı olma koşu ile siyah çekme olasılığım nedir?



Paul O. Lewis (2012) MBL, Mol. Evol.

$Pr(S, N)$

$$Pr(S) Pr(S | N) = Pr(N) Pr(N | S)$$

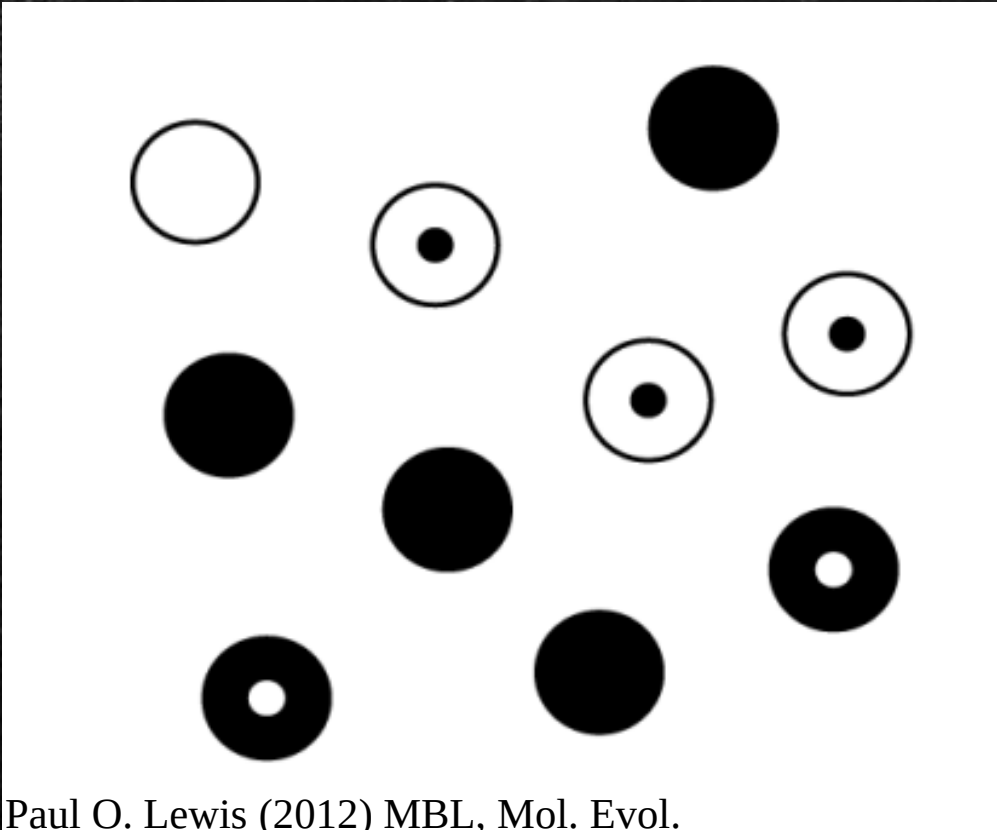
$$1/2 \times 2/5 = 3/5 \times 1/3$$

$$P(S|N) = \frac{p(S)p(N|S)}{p(S)}$$

$$= \frac{\frac{3}{5} \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{5}$$

Birleşme Kuramı ve Bayes Kuralı

Soru: Noktalı olma koşu ile siyah çekme olasılığım nedir?



Paul O. Lewis (2012) MBL, Mol. Evol.

$$Pr(S|N) = \frac{Pr(S)Pr(N|S)}{Pr(N)}$$

$$Pr(S|N) = \frac{Pr(N, S)}{Pr(N, S) + Pr(N, B)}$$

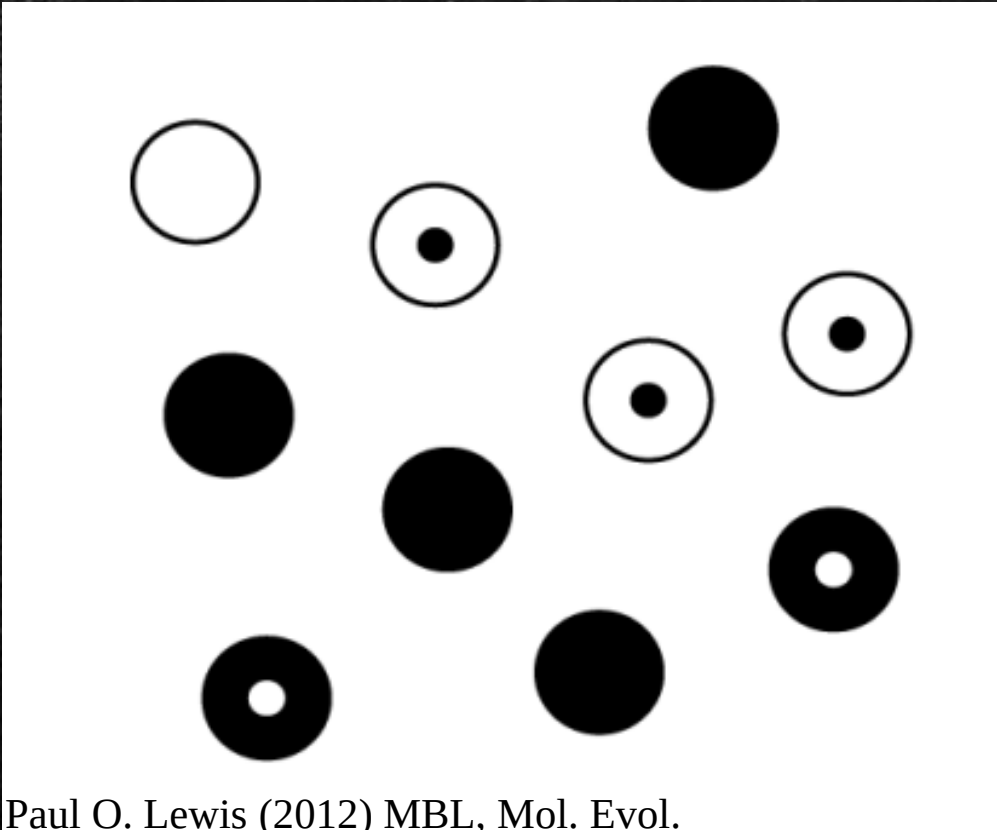
Pr(N) noktalı olmanın bileşen olasılığıdır.

Hesaplayabilmek için bütün renklerin bileşimini alırız.

Matematiksel olarak normalizasyon katsayısıdır ve **Pr(S | N) + Pr(B | N) = 1.0** olmasını garantiler.

Birleşme Kuramı ve Bayes Kuralı

Soru: Noktalı olma koşu ile siyah çekme olasılığım nedir?



Paul O. Lewis (2012) MBL, Mol. Evol.

$Pr(S, N)$

$$= \frac{Pr(S)Pr(N|S)}{Pr(N, S) + Pr(N, B)}$$

$$= \frac{Pr(S)Pr(N|S)}{Pr(S)Pr(N|S) + Pr(B)Pr(N|B)}$$

$$= \frac{Pr(S)Pr(N|S)}{\sum_{\theta \in \{S, B\}} Pr(\theta)Pr(N|\theta)}$$

Birleşme Kuramı ve Bayes Kuralı

The diagram shows the Bayes' Theorem equation with color-coded components and arrows pointing to their meanings:

$$\Pr(\theta|D) = \frac{\Pr(D|\theta) \Pr(\theta)}{\sum_{\theta} \Pr(D|\theta) \Pr(\theta)}$$

- $\Pr(\theta|D)$ is labeled **θ hipotezinin ardıl olasılığı** (Posterior probability of θ hypothesis).
- $\Pr(D|\theta)$ is labeled **θ hipotezinin olabilirliği** (Likelihood of θ hypothesis).
- $\Pr(\theta)$ is labeled **θ Hipotezinin ön olasılığı** (Prior probability of θ hypothesis).
- The denominator $\sum_{\theta} \Pr(D|\theta) \Pr(\theta)$ is labeled **Verinin bileşen olasılığı (hipotezlerin bileşkesi)** (Component probability of the data (mixture of hypotheses)).

D gözlemlerimiz (yani verilerimiz).

θ bir veya birden çok gözlemleyemediklerimiz (model veya model parametreleri):

- ağaç modelleri (i.e. Ağaç topolojisi)
- mutasyon modelleri (e.g. JC, F84, GTR, vs.)
- göç matrisleri (m_{ij})
- üssel büyüme (g)
- Birleşme zamanları (t_i veya T_{mrca})

Birleşme Kuramı ve Tam Olabilirlik (Full Likelihood)

$$p(\mathbf{D} \mid \theta) = \left[\begin{array}{l} \text{Elimdeki veriyi gözlemleme olasılığını en} \\ \text{yüksek kılan parametre değerleri nedir?} \end{array} \right.$$

$$p(\mathbf{D} \mid \theta) = \int_{\mathbf{G}} p(\mathbf{G} \mid \theta) p(\mathbf{D} \mid \mathbf{G}) d\mathbf{G}$$

$$p(\mathbf{G} \mid \theta)$$



Parametre değerleri altında gen ağacının olasılığı.

$$p(\mathbf{D} \mid \mathbf{G})$$



Eldeki gen ağacı altında verinin olasılığı.
Filogenetikte ağaç-olabilirliği (tree-likelihood).

Birleşme Kuramı ve Tam Olabilirlik (Full Likelihood)

Terminal Takson	Tarihçe Sayısı
3	3
4	18
5	180
7	56700
8	1587600
10	2571912000
15	69.5806×10^{18}
20	56.4481×10^{28}
30	43.6846×10^{52}
50	32.8632×10^{111}
100	13.7416×10^{283}

$$p(\mathbf{D} | \theta) = \int_{\mathbf{G}} p(\mathbf{G} | \theta) p(\mathbf{D} | \mathbf{G}) d\mathbf{G}$$

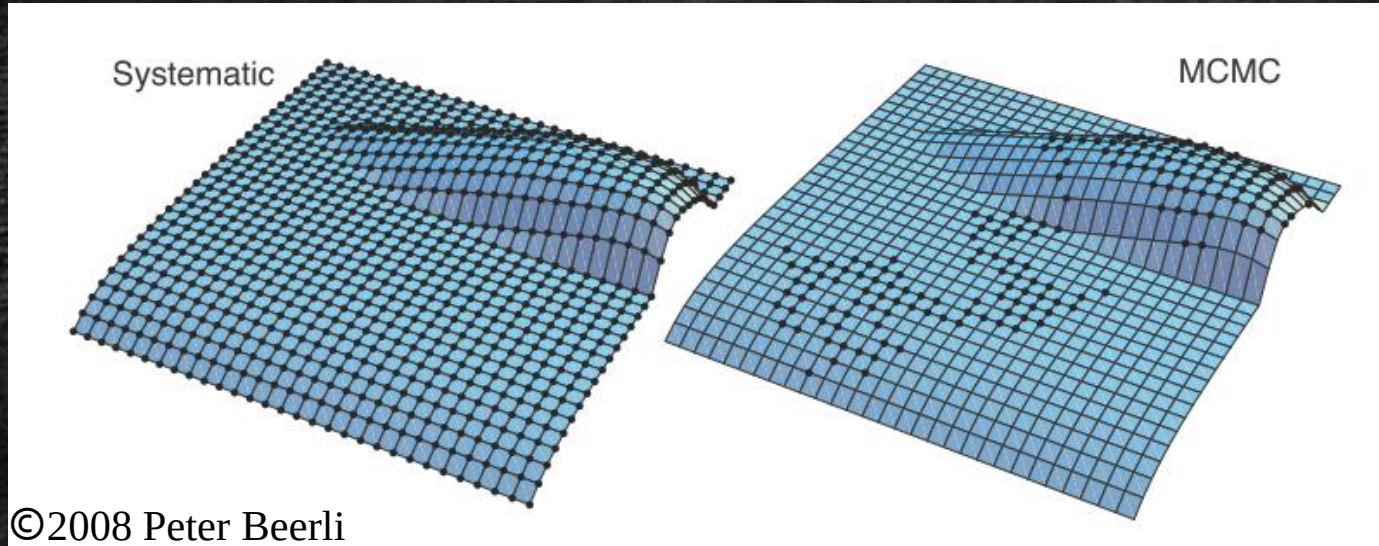
Olası gen ağaç sayısı
olağanüstü yüksek

Markov Chain Monte Carlo

Milli piyangoyu kazanma olasılığı = $1/10,000,000$

Evrende bilinen yıldız sayısı = $3 \times 10^{22} - 10^{24}$

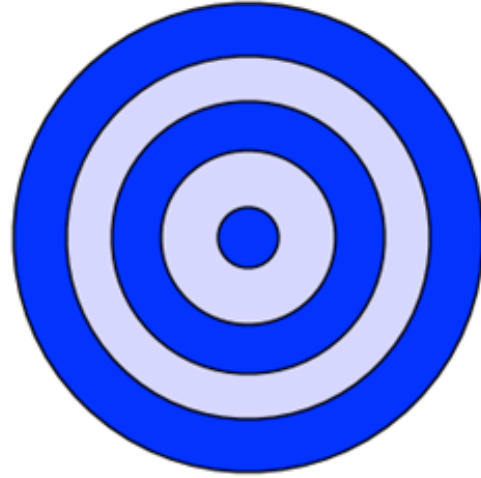
Metropolis – Hastings Algoritması



- Markov zinciri parametre uzamın ilginç noktalarının hepsine uğramalı.
- İlginç bölgelere sürekli uğranmalı, hatta zincir sonuza kadar götürülürse sonsuz kere uğramalı.
- Örneklemin ortalaması beklenen değere birleşmeli.

Parametre Uzamını Taramak = Öncül Dağılımlar

Paul O. Lewis (2012) MBL, Mol. Evol.



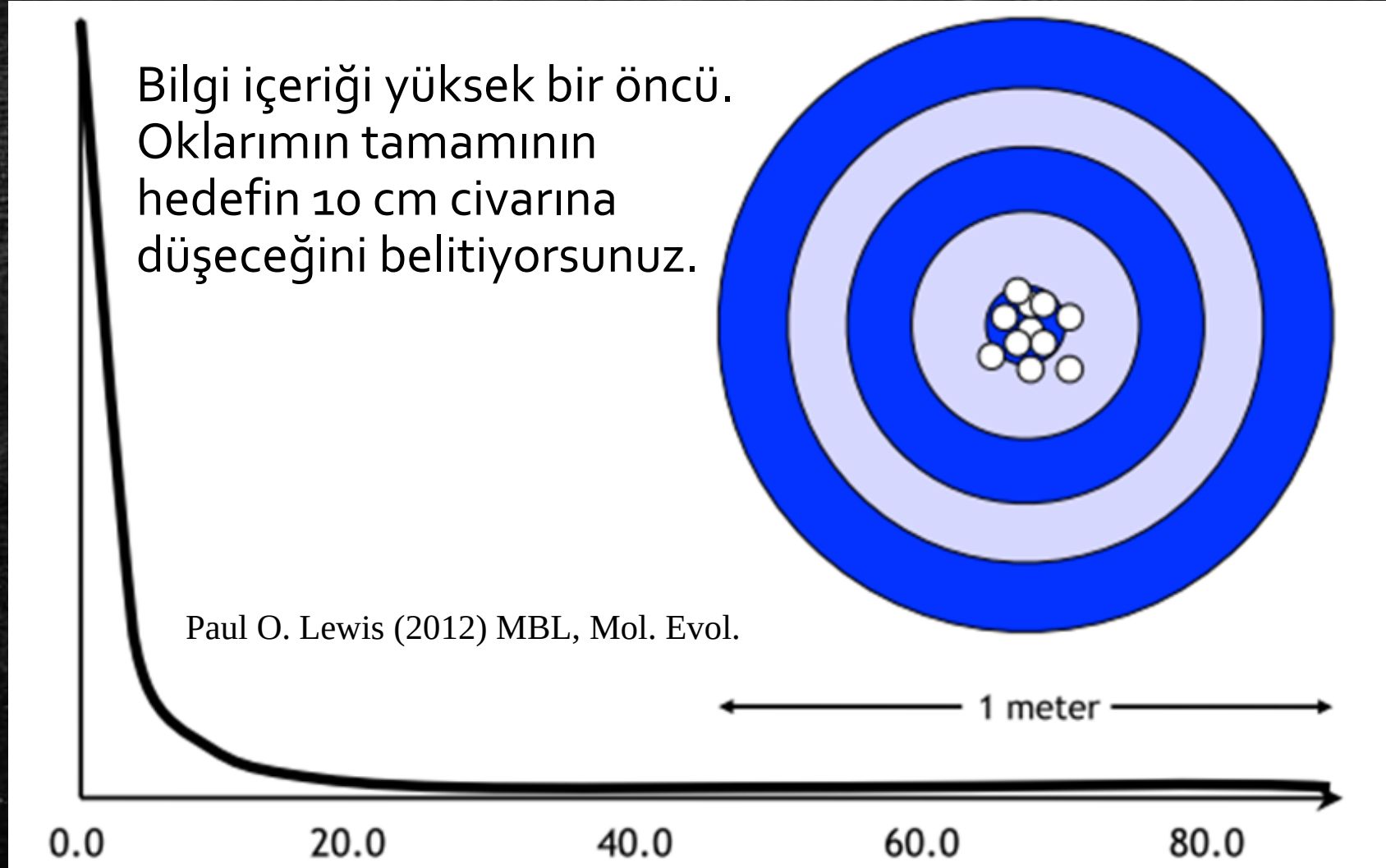
Benim okçuluk yeteneğim hakkında hiç bir şey bilmeden 50 metreden attığım okların hedefden ne kadar mesafe uzağa (d) düşeceği üzerine tahminlerinizi gösteren bir eğri çizmenizi istiyorum.

0.0

d

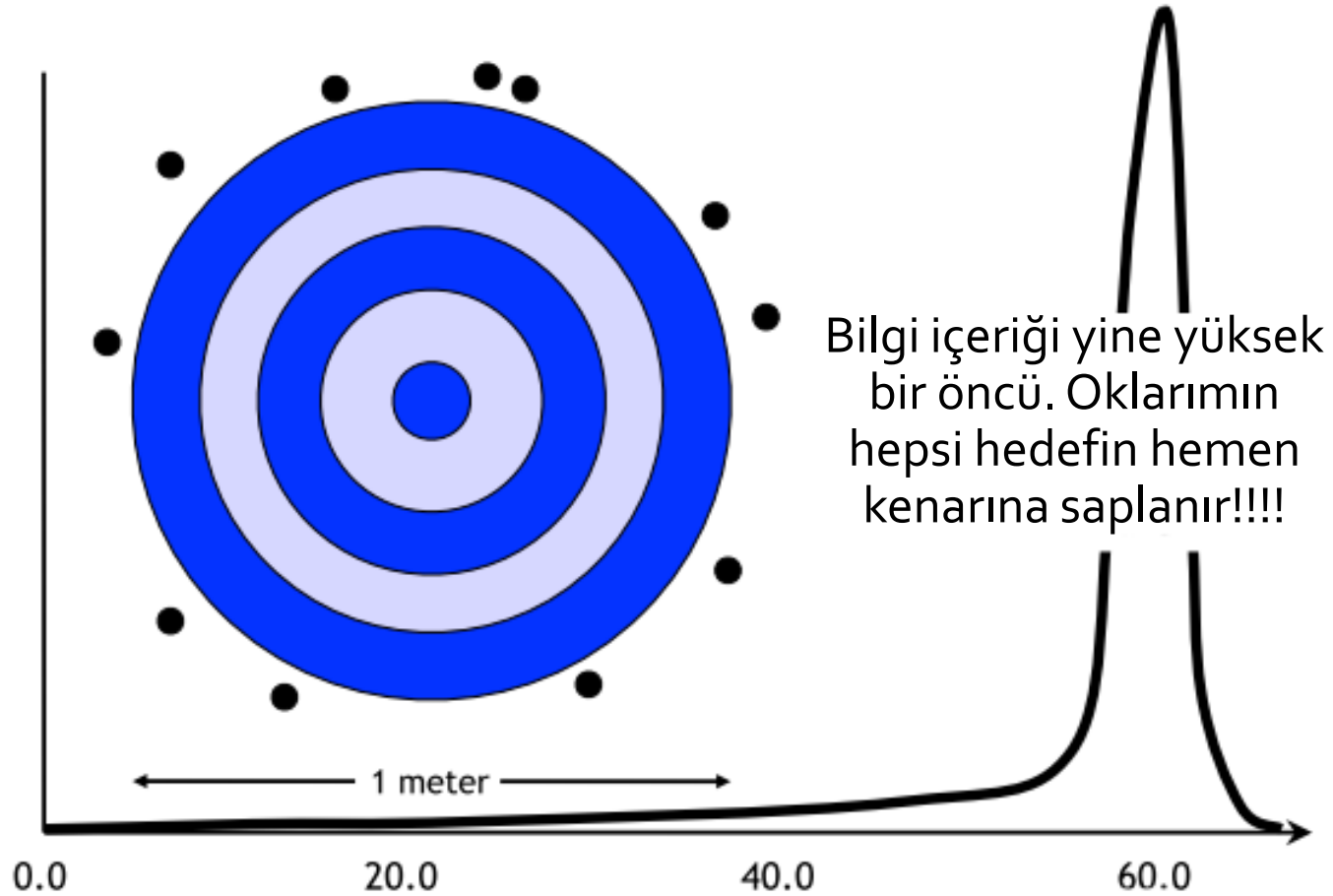
∞

Parametre Uzamını Taramak = Öncül Dağılımlar

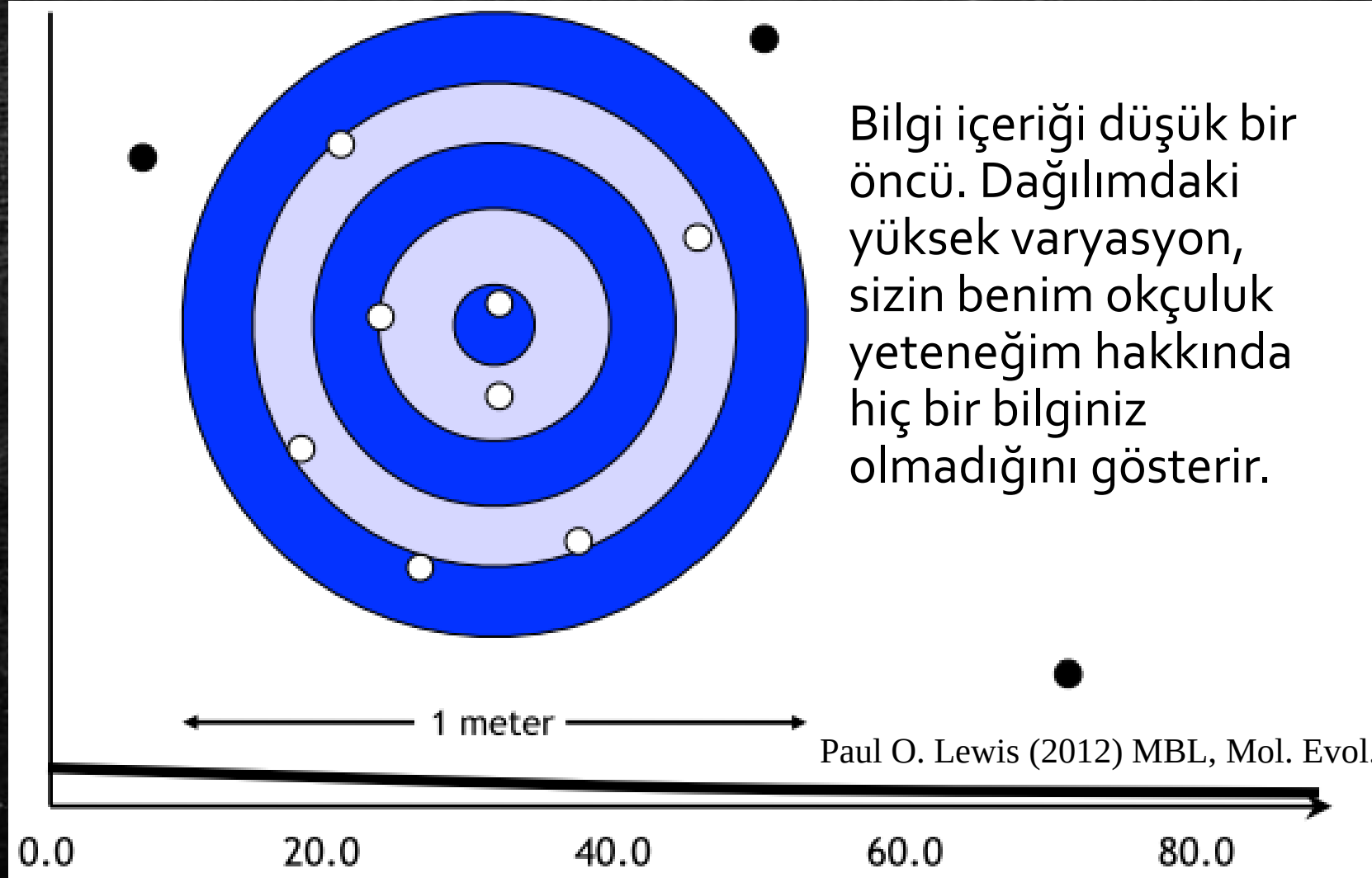


Parametre Uzamını Taramak = Öncül Dağılımlar

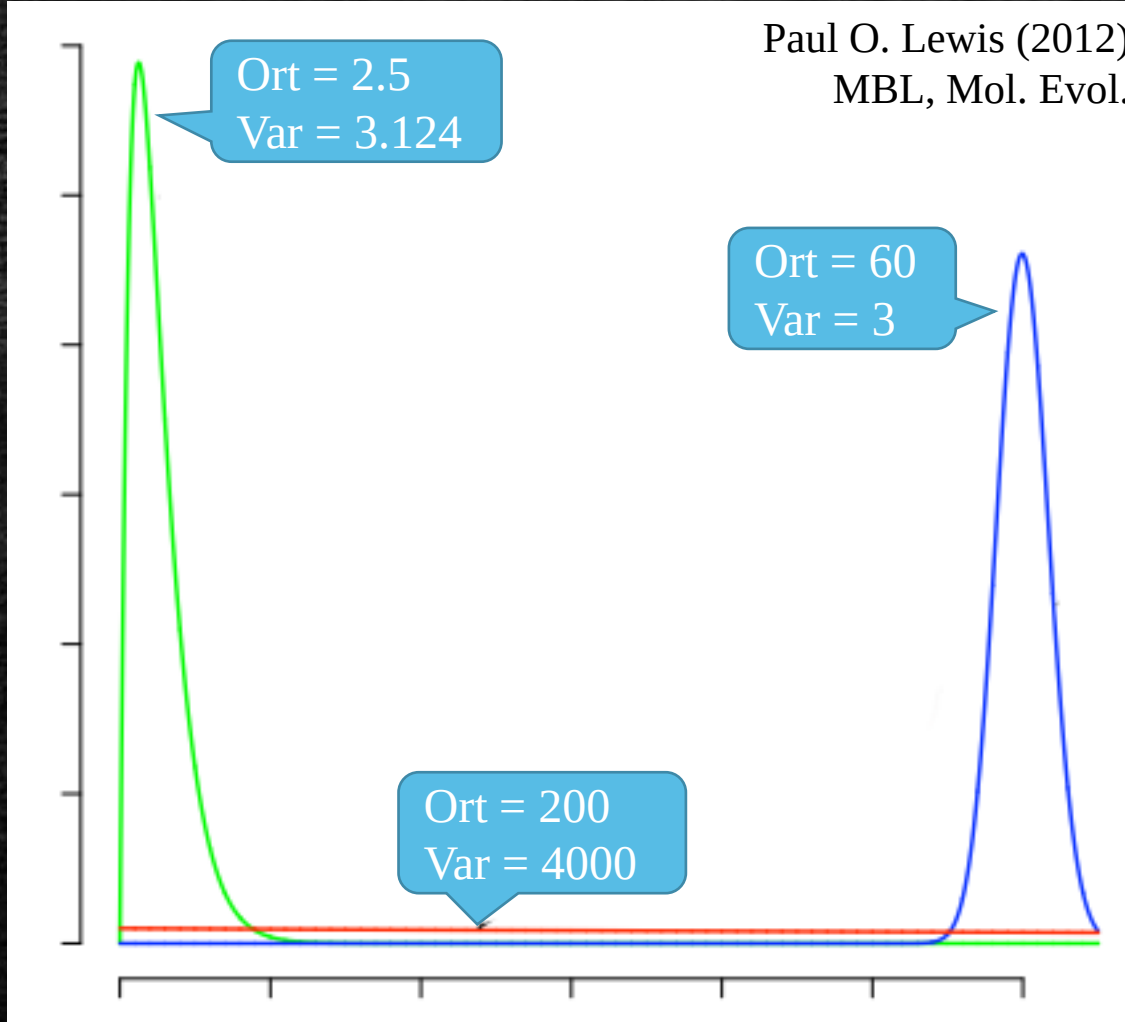
Paul O. Lewis (2012) MBL, Mol. Evol.



Parametre Uzamını Taramak = Öncül Dağılımlar

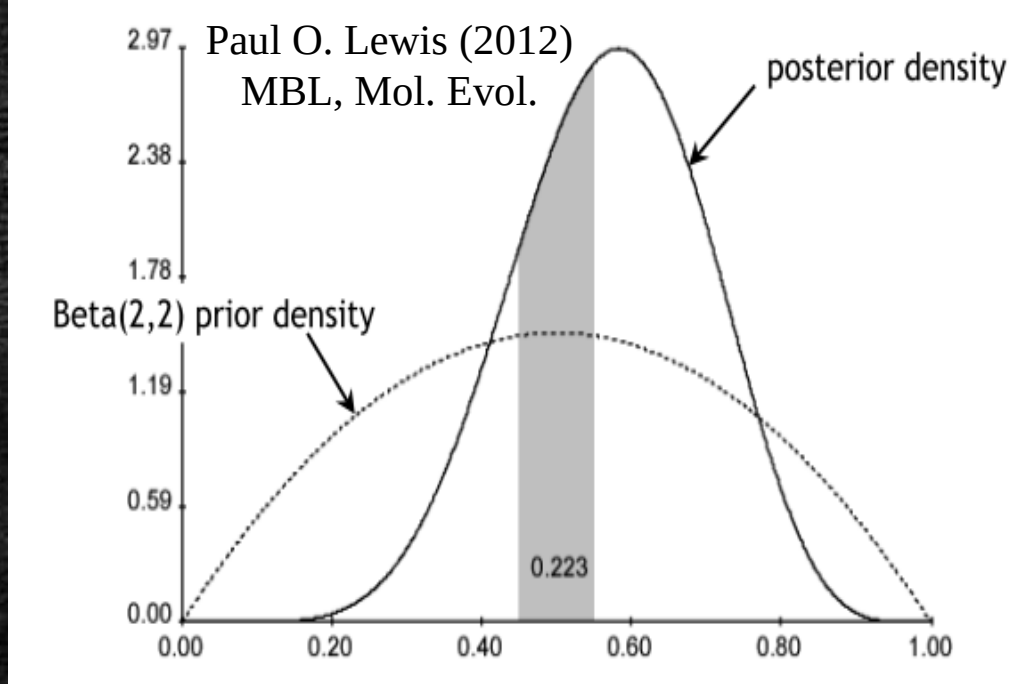


Parametre Uzamını Taramak = Öncül Dağılımlar



Bu eğrilerin hepsi gamma olasılık dağılımının türevleridir. Okçuluk örneğinde başarı değerlerini bulamak için gamma dağılımını ön olasılık aralığını (dağılımını) belirlemek için kullanabilirdik. Fakat bütün dağılımlar aynı derecede bilgi içermiyor. Yüksek varyans düşük bilgi demek!!!!

Parametre Uzamını Taramak = Öncül Dağılımlar



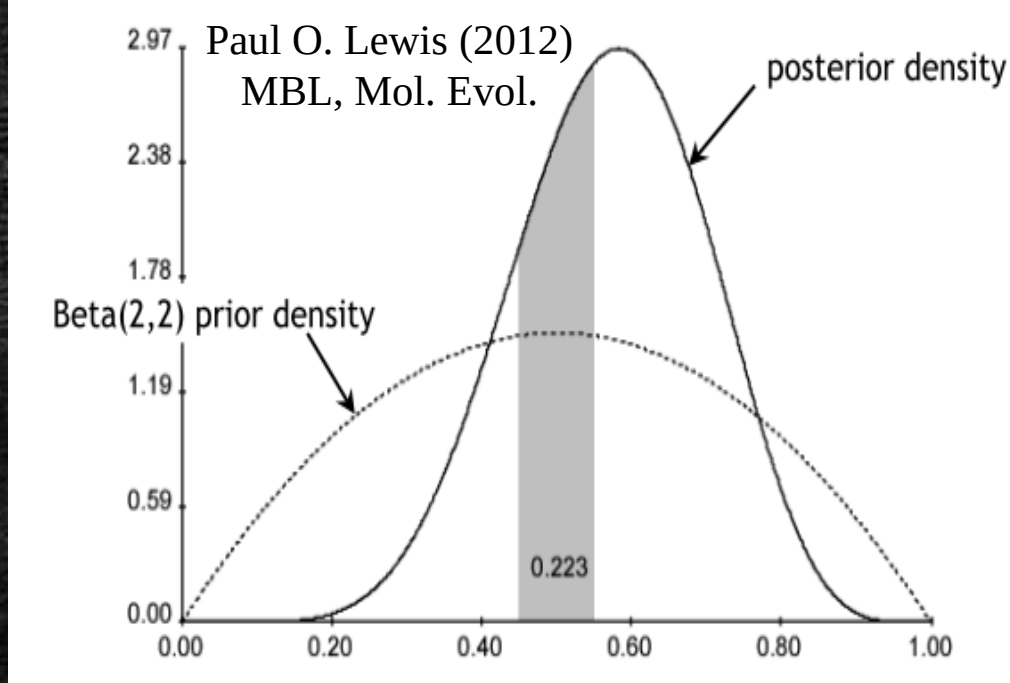
Amacımız ön dağılım olasılık eğrilerini kullanarak tahmin etmek istediğimiz parametrenin ortalama değeri hakkında bize bilgi verebilecek ardıl olasılık eğrilerini bulmak.

Parametre Uzamını Taramak = Öncül Dağılımlar

Sıklıkla Kullanılan Öncüller

- **Tekdüze (Uniform)**, Ağaç topolojileri için sıklıkla kullanılır.
- **Beta**, yüzde şeklinde dağılan veriler için.
- **Gamma yada Log-normal**, dal uzunlukları veya $[0, \infty)$ aralığında dağılan parametreler için.
- **Dirichlet**, baz frekansları ve GTR matriksindeki görece değişim hızları için.

Parametre Uzamını Taramak = Ardıl Dağılımın Tahmini



Amacımız ön dağılım olasılık eğrilerini kullanarak tahmin etmek istediğimiz parametrenin ortalama değeri hakkında bize bilgi verebilecek ardıl olasılık eğrilerini bulmak.

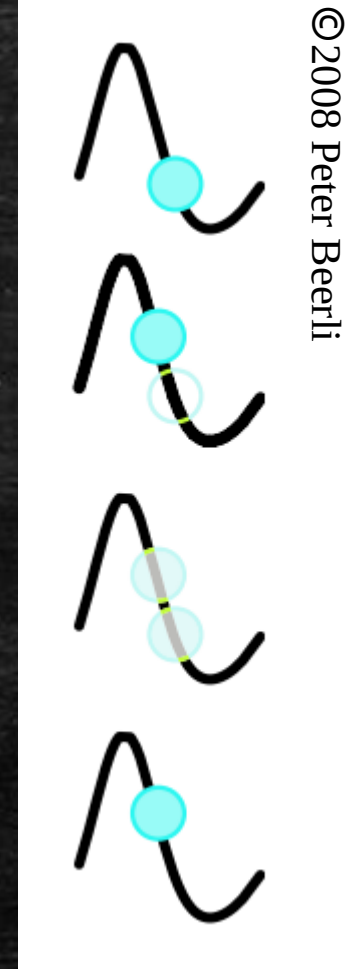
Bilgi içeriği yüksek ardıl olasılığı bulmak için ön olasılık uzamı nasıl taranmalı.

Markov Chain Monte Carlo (MCMC)

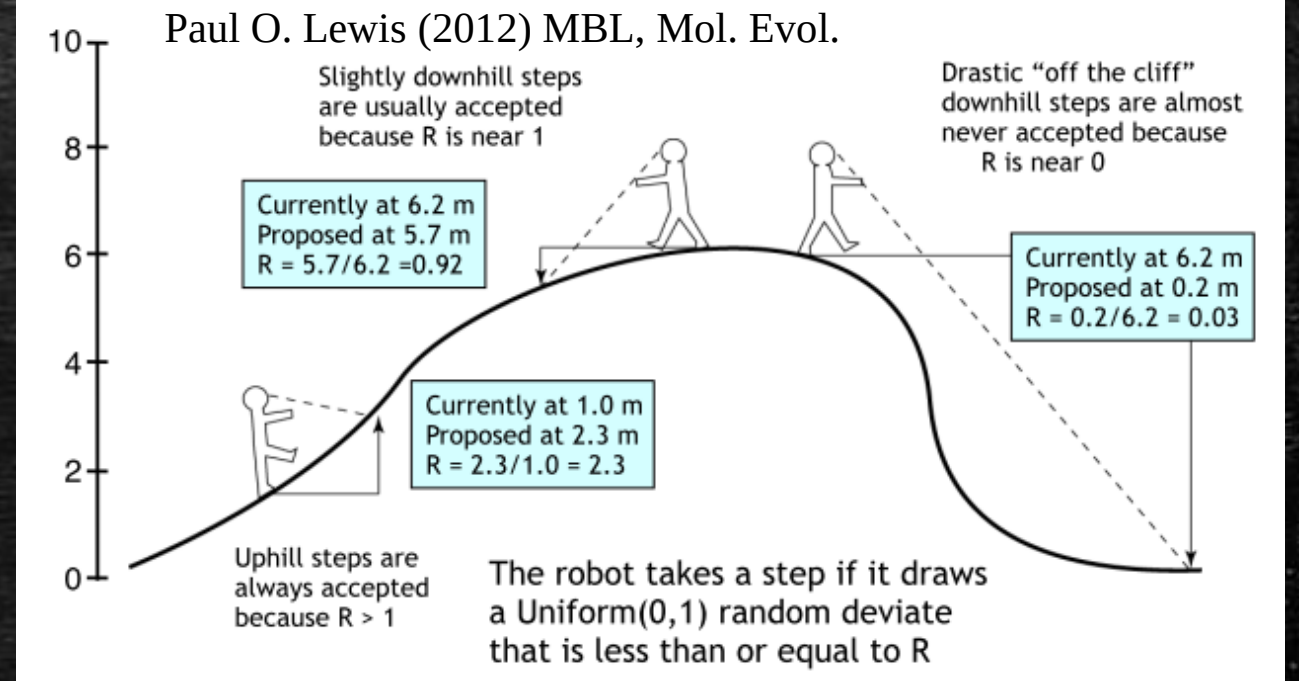
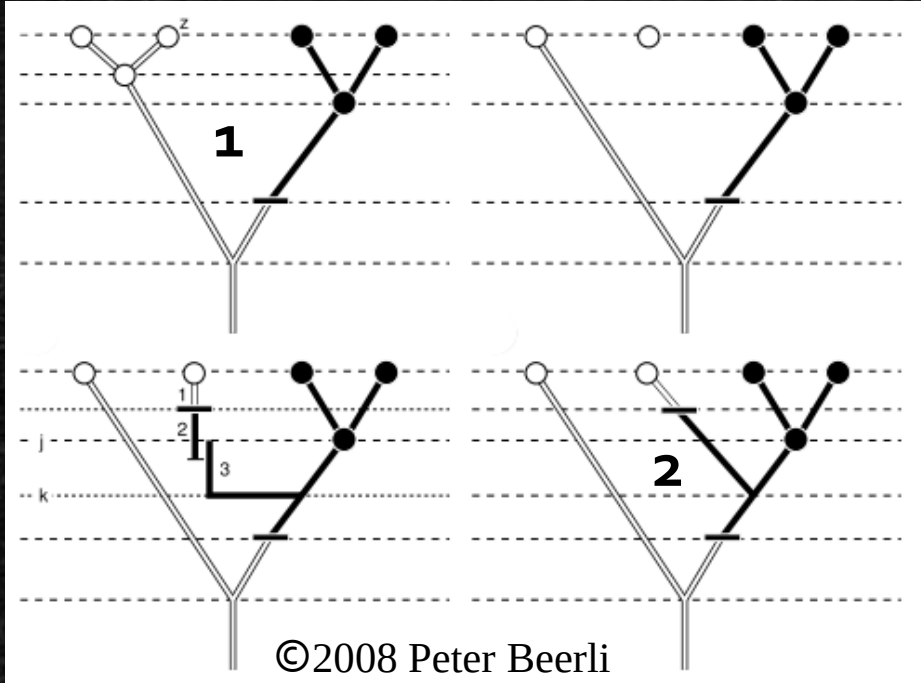
Ardıl Dağılımın Tahmini MCMC Yöntemi

Metropolis tarifi:

1. ilk durum
2. Eski durumu yeni durum ile değiştir ve yeni durumun olasılığını hesapla.
3. Yeni durumun eski durumdan daha iyi olup olmadığını değerlendir : yeni ile eskinin oranını 0 ile 1 arasında seçilen bir sayıya eşit veya yüksek ise kabul et.
4. Kabul ise yeni duruma geç yoksa eski durumda kal.
5. 1. basamağa geri dön.



Ardıl Dağılımın Tahmini MCMC Yöntemi



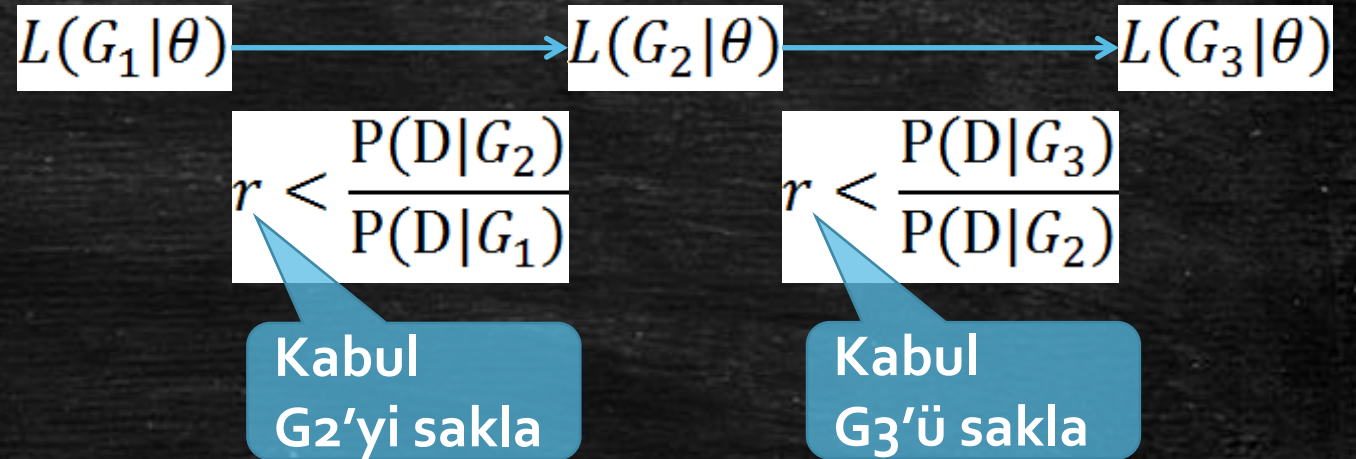
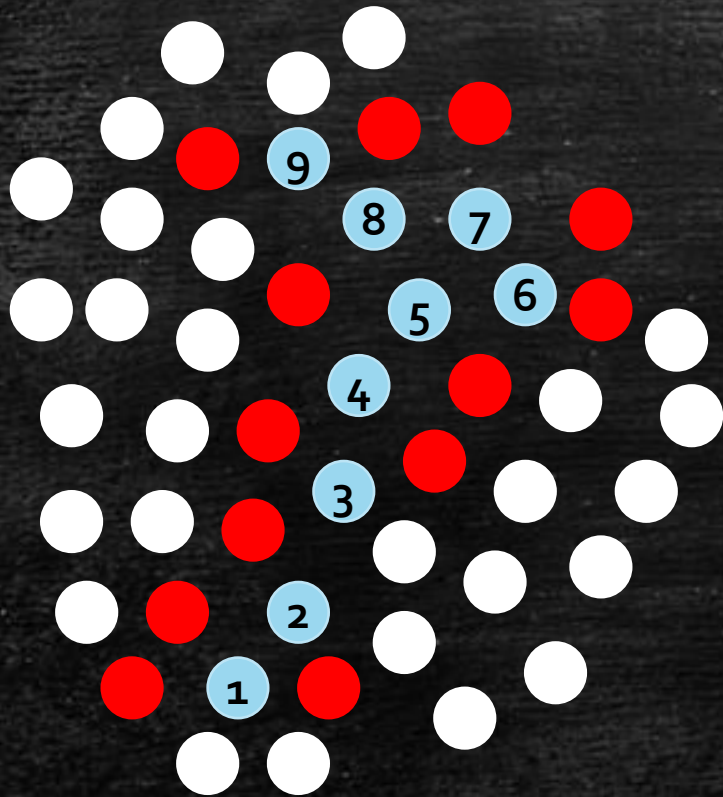
$$\frac{f(\theta^*|D)}{f(\theta|D)} = \frac{\frac{f(D|\theta^*)f(\theta^*)}{\cancel{f(D)}}}{\frac{f(D|\theta)f(\theta)}{\cancel{f(D)}}} = \frac{f(D|\theta^*)f(\theta^*)}{f(D|\theta)f(\theta)}$$

Ön olasılık

Olabilirlik
oranı

Ardıl Oran

Ardıl Dağılımdan Olabilirliğin Hesaplanması



Örneklenen ağaçlar (veya parametreler) en sonunda parametre değerlerinin olasılığını $L(\Theta)$ hesaplamak için kullanılır.

MCMC Zincirinin Uzunluğu ne Olmalı?

Doğru ardıl olasılık dağılımına ulaştığımı nasıl bileceğim?

- MCMC sonsuza kadar işlemesine izin verildiğinde sorunsuz çalışıyor ve kesinlikle doğru ardıl dağılım aralığını bulacaktır !!!!!
- Zincirlerimiz sonludur ve zincirin istenilen dağılıma ulaşp ulaşmadığını ve bütün ilgili bölgeleri dolaşıp dolaşmadığını anlamak oldukça zordur.
- Etkin tarama ve örneklemin oluştuğunu anlayabilmek için çeşitli diagnostikler vardır. Fakat hiç biri mükemmel değildir.

MCMC Zincirinin Uzunluğu ne Olmalı?

Etkin tarama ve örneklem için diagnostikler

Etkin Örneklem Sayısı (EÖS): Eldeki verinin olasılığını en yüksekte tutan bölgeden çekilen örnek sayısı. Genellikle 1000'in üzerinde olması tercih edilmelidir.

Tarama izlerinin arka plan parazitini andırması istenilir

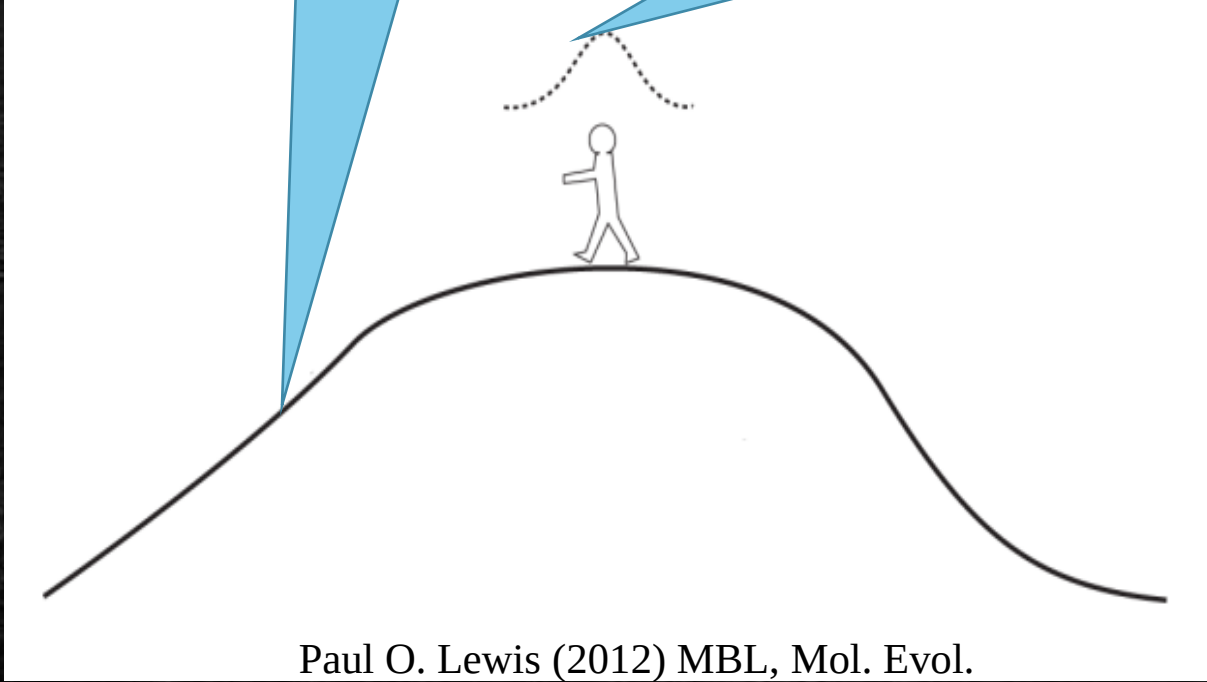
Metropolis-coupled MCMC yöntemi kullanılarak olasılık uzamında farklı tepeler ulaşılması sağlanır.

Farklı **bağımsız zincirler** yürütülür ve bu zincirlerin aynı ardıl olasılık dağılımına ulaşp ulaşmadığına bakılır!!!!!! (Bunu yapmadan yürütülen hiç bir MCMC analizine güvenemezsiniz!!!!!!)

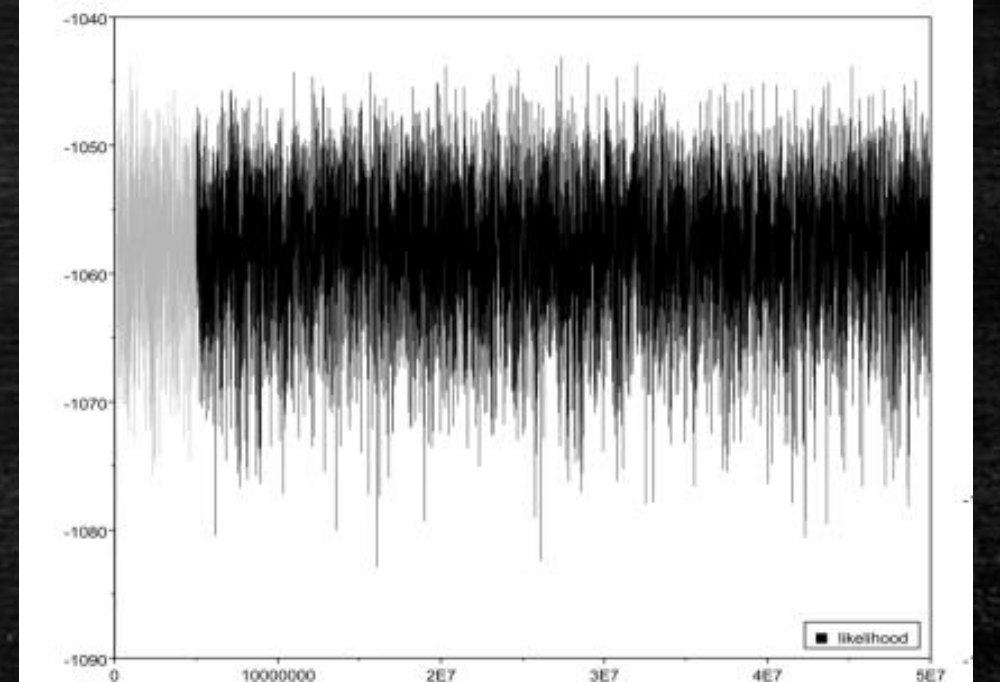
Etkin Tarama ve Örneklem için Diagnostikler

Hedef Dağılım

Önerilen Dağılım



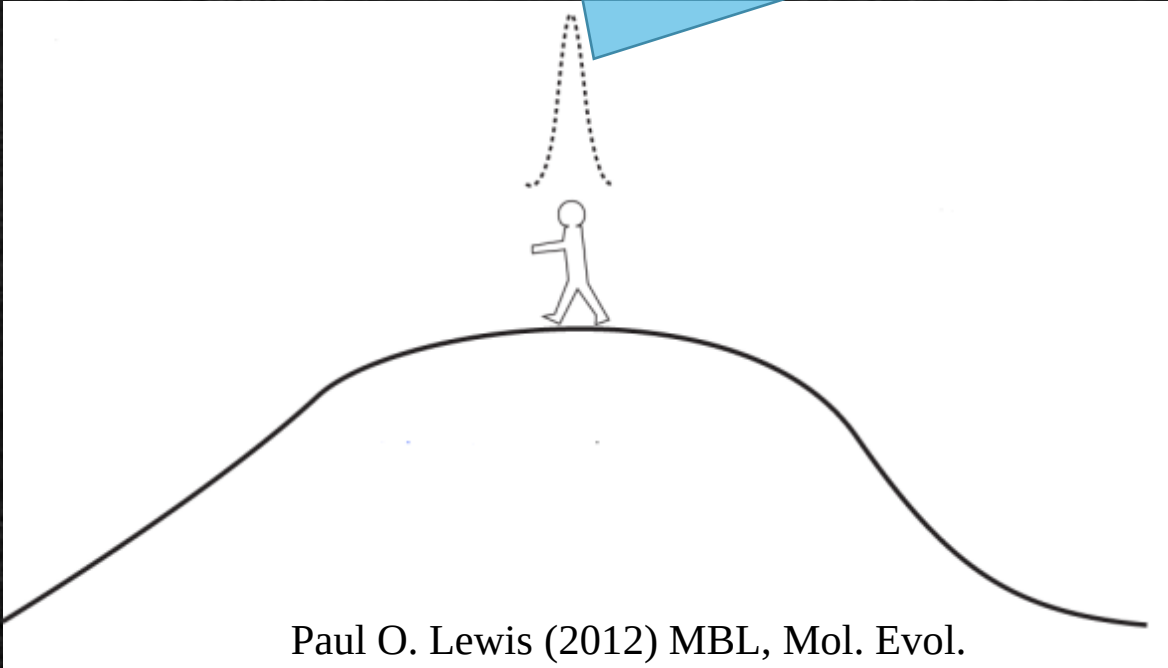
Önerilen bu dağılımın **iyi karışım** neden olduğunu varsayalım. **İyi karışım** ne demektir?



Tarama izlerinin beyaz sesi andırması iyi karışım demektir.

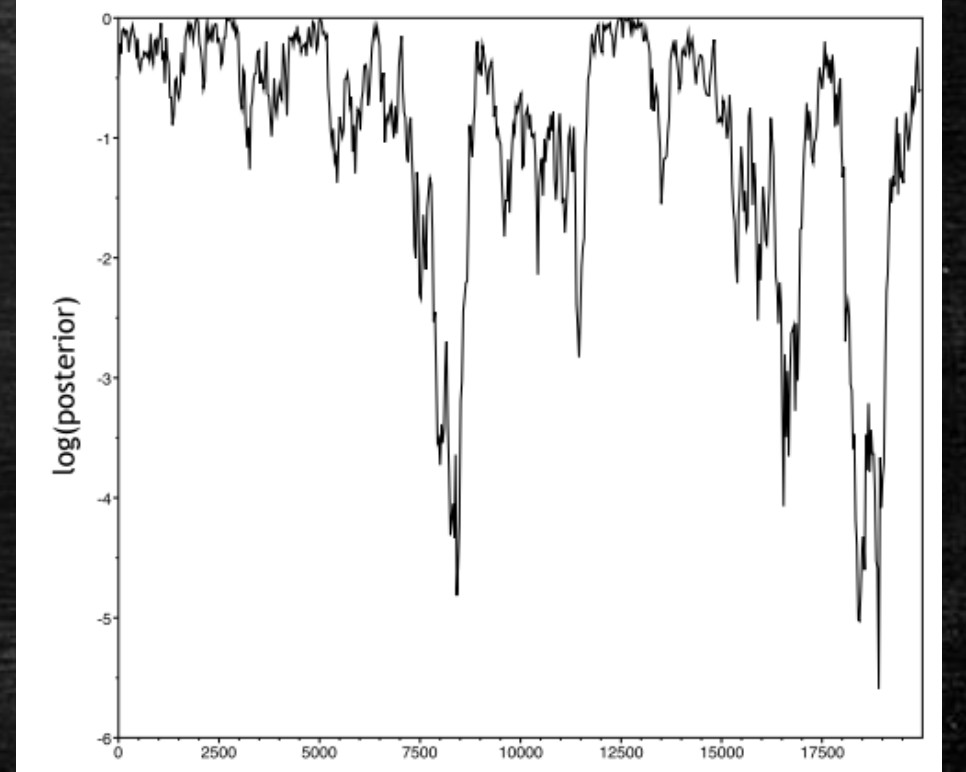
Etkin Tarama ve Örneklem için Diagnostikler

Önerilen Dağılımın varyansı düşük



Dezavantaj: daha küçük adımlar, aynı alanı taramak için daha uzun süre.

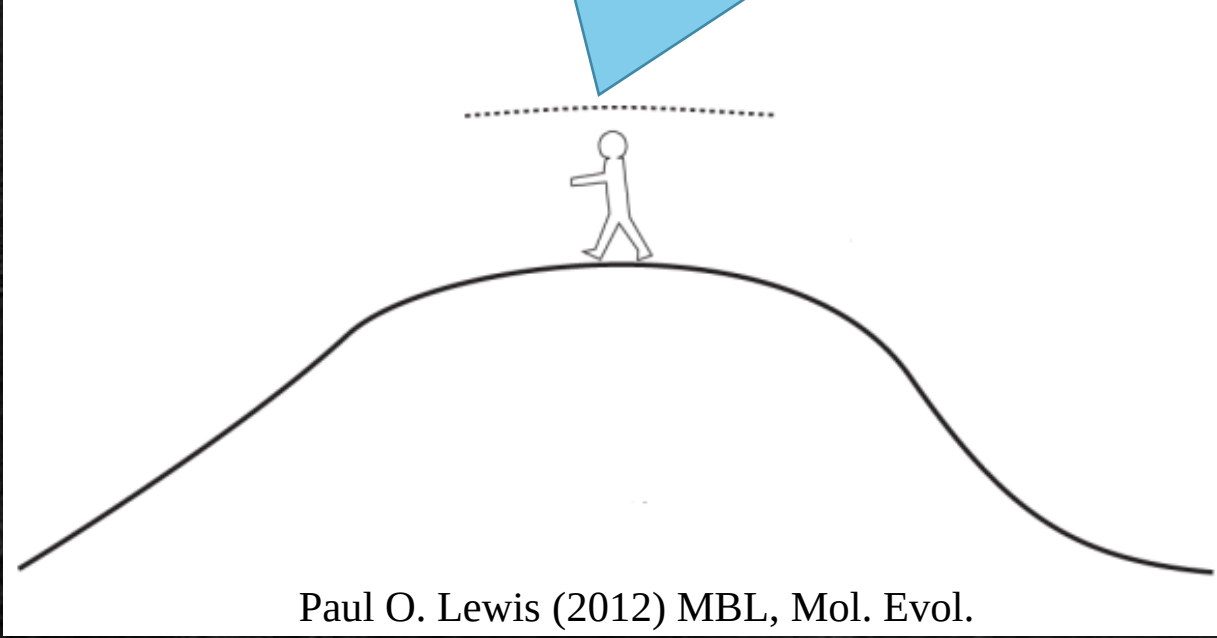
Avantaj: Önerilen yeni adımın red oranı çok düşük.



Eğer adımlar arası mesafeler çok kısa ise, **uzun süreli eğilimler** görülür.

Etkin Tarama ve Örneklem için Diagnostikler

Önerilen Dağılımın varyansı yüksek



Dezavantaj: adımların bazıları vadilere denk gelir bu adımlar red edilir.

Avantaj: Hızla büyük alanlar ziyaret edilebilir.



Zincirin **takılması** demek önerilen adımların bir çoğunun vadilere denk geldiğini gösterir.

MCMC Tarama = Büyük Adım vs Küçük adım

Ödün vermek yada vermemek

- Büyük adımlar atmak ardıl dağılım uzamında bir adadan diğerine zıplamaya yardımcı olur.
- Daha küçük adımlar ile ilerlemek daha iyi karışma ile sonuçlanır.
- Peki ödün vermeden her iki yönteminde iyi taraflarını alabilir miyiz?

EVET: METROPOLIS-COUPLED MCMC

MCMCMC

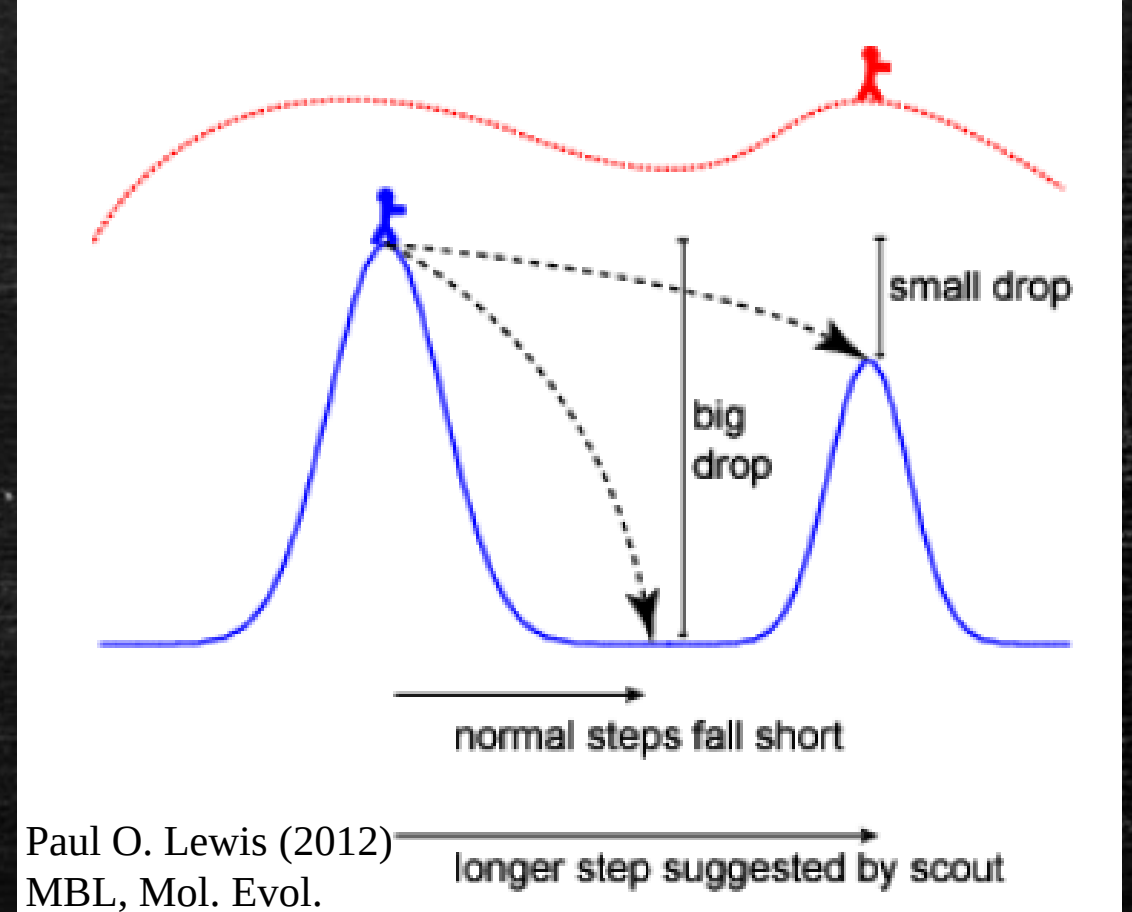
Metropolis-Coupled MCMC = MCMCMC

MCMCMC birkaç zinciri aynı anda yürütme prensibine dayanır.

Soğuk zincir önemli zincirimizdir (yani tahminlerimizi bu zincirden alırız).

Geri kalan zincirler **ısıtılmış zincirlerdir** ve izci görevi görürler.

İzciler ardıl dağılım uzamında farklı tepeler bulunca soğuk zincire haber verir, soğuk zincir bu alana zıplayarak o bölgeye tarar.



Tam Olabilirlik atında Filocoğrafya

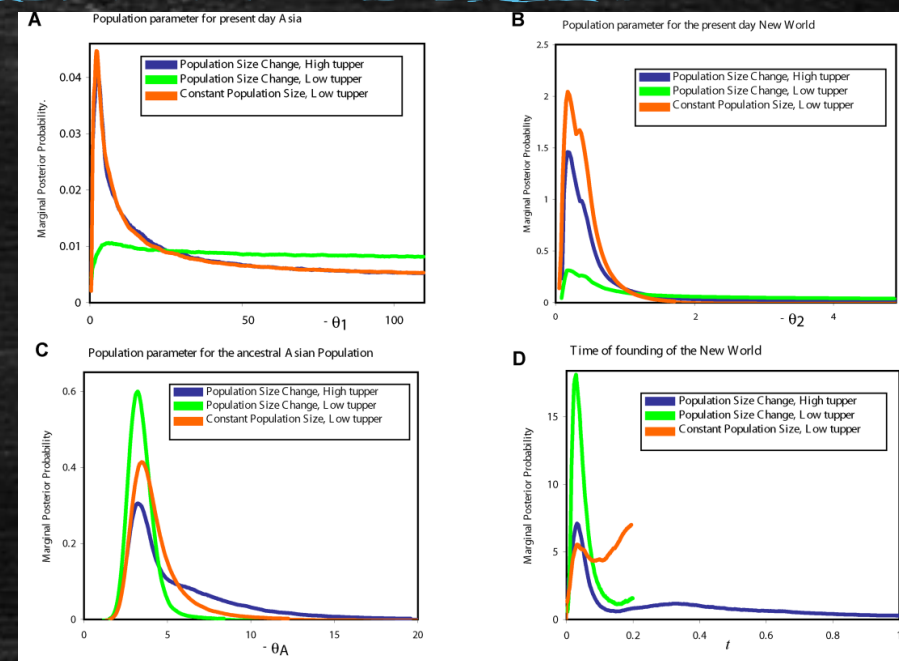
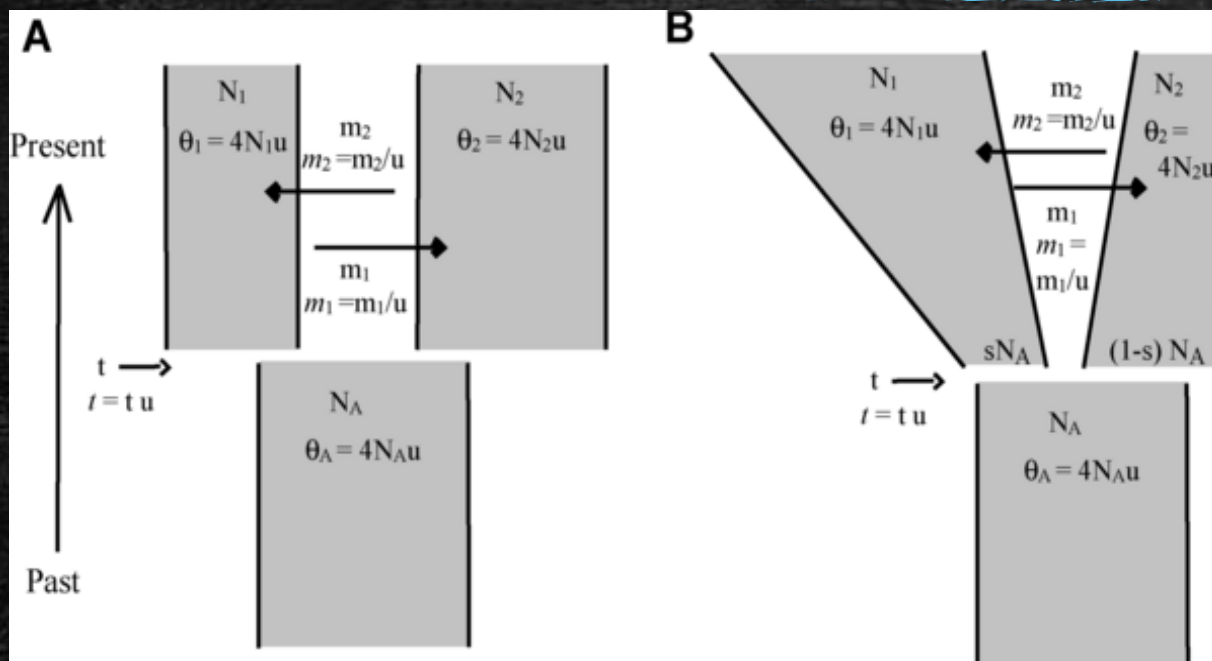
Birleşme kuramı ve var olan filogenetik yöntemler elimizdeki genetik varyasyona neden olan parametreleri ve süreçleri etkin bir şekilde modellememize izin verir.

Bayes, Likelihood ve **MCMC** yöntemleri ise modellere ilişkin parametre uzamını gezmemize ve etkin parametre tahmini yürütmemizi sağlar.

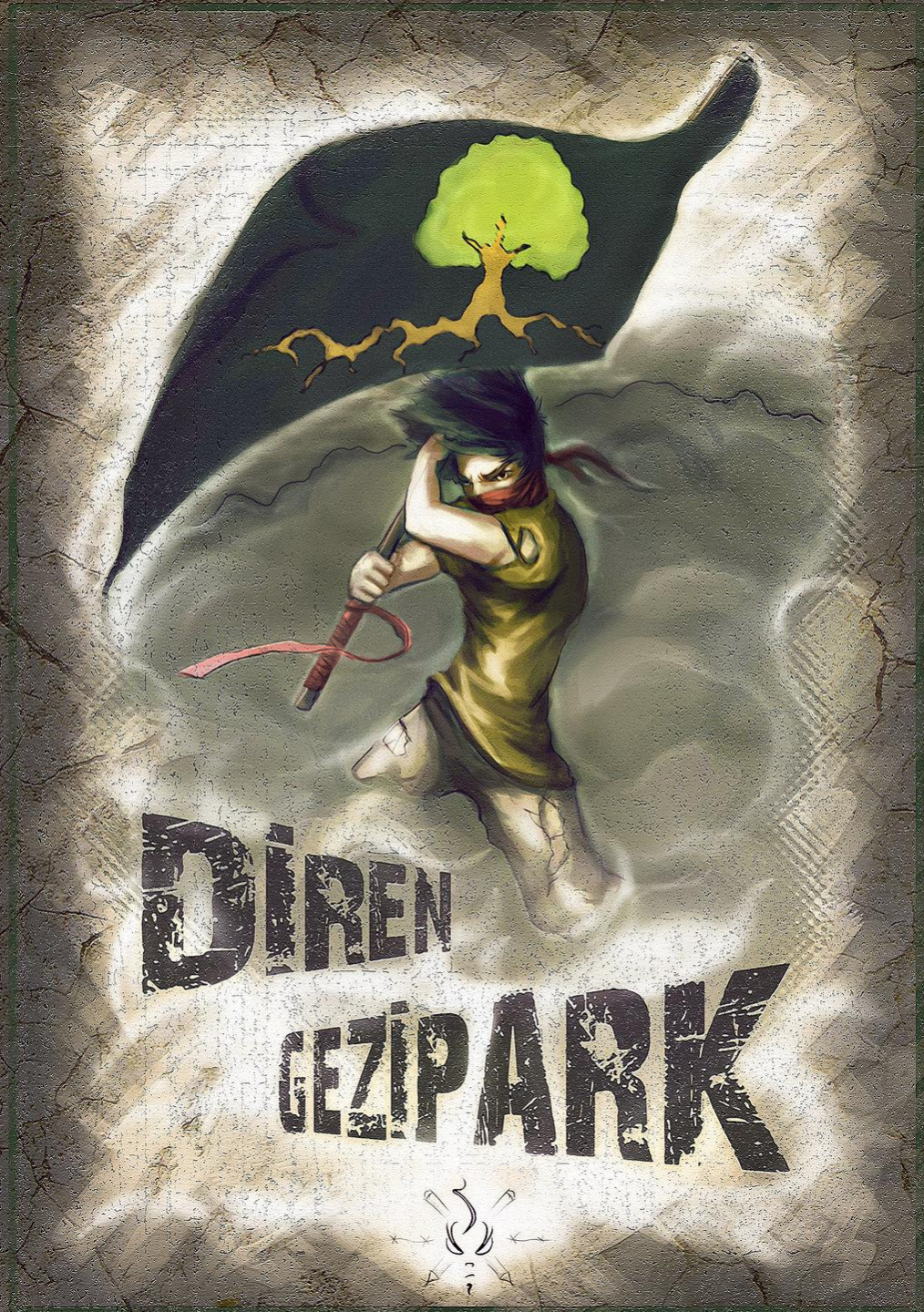
İSTATİSTİKSEL FİLOCOĞRAFYA

Zaman ve uzam içinde görülen genetik varyasyona neden olan tarihsel, demografik ve evrimsel süreçlerin tahminlenmesi ve alternatif hipotezlerin sınanması.

Hey J. 2005 Amerika Kıtasının Keşfi!!!



Demographic Term	Population Size Change High t_{upper}	Population Size Change Low t_{upper}	Constant Population Size Low t_{upper}
N_1	7,190	19,200	5,394
N_2	480	830	770
N_A	9,180	9,040	9,640
$s N_A$	9,100	8,970	Not applicable
$(1-s) N_A$	76	70	Not applicable
t (years)	7,130	6,350	44,400 ^a (7,900)
$2N_1m_1 = \theta_1 m_1 / 2$	11.8	10.5	3.2
$2N_2m_2 = \theta_2 m_2 / 2$	0.9	1.4	2.3



D İ R E N T Ü R K İ Y E
D İ R E N E V R İ M