T.C.
ULUDAĞ ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

DNA,RNA SEKANS HİZALAMA; ÖRÜNTÜ YAKALAMA

Alparslan YÜCE
031690039
Kerem AKIN
031790016
Alihan SULTAN
0316KVU90001

TASARIM DERSİ ARA RAPORU

BURSA 2020

T.C.
ULUDAĞ ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

DNA,RNA SEKANS HİZALAMA; ÖRÜNTÜ YAKALAMA

Alparslan YÜCE
031690039
Kerem AKIN
031790016
Alihan SULTAN
0316KVU90001

Projenin Danışmanı : Doç. Dr. Gıyasettin Özcan
Jüri Üyesi :
Jüri Üyesi :

ÖZET

Bu ara raporda geliştirilecek proje ile ilgili araştırma yapılmıştır. Gerekli algoritmalar araştırılmıştır. Her biri açık şekilde anlatılıp örneklendirilmiştir. Geliştirilecek uygulama için gerekli bilgiler edinilmiştir. Projenin nasıl yapılacağı planlanmıştır. Bu konu ile alakalı yapılan önceki çalışmalar ve bu çalışmalarda elde edilen sonuçlar araştırılmıştır. Sonuçlar karşılaştırılıp algoritmalar hakkında fikir sahibi olunmuştur. Projenin uygulanma aşaması için önemli adımlar atılmıştır. Örnek kodlar incelenmiştir. Yapılacak olan projede amaç; filojeni algoritma uygulamasıdır. Guide trees, evolutionary trees - Neighbor-Joining ve UPGMA, parsimony, Sankoff ve Fitch algoritmalarının kodlanması ve bir arayüzde entegrasyonu hedeflenmiştir.

İÇİNDEKİLER

Sayfa no

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

Sayfa no

1. GİRİŞ

"Filogenetik", bütün organizma grupları arasındaki evrimsel ilişkiyi ata-soy ilişkileri şeklinde ortaya çıkarmayı amaçlar. Organizmaların sahip olduğu moleküler mekanizmalar, tek bir ataya sahip olduklarını göstermektedir. Ortak atadan evrimleşmeleri sayesinde türler, birbirleriyle ilişkilendirilebilirler. Filogenetik sistematiğinin kurucularından Alman biyolog Emil Hans Willi Hennig, bu ilişkilendirmenin; türler arası morfolojik, fizyolojik, genetik, coğrafik ve ekolojik farklılıklar dikkate alınarak gerçekleştirilebileceğini ortaya koymuştur. Saptanan filogenetik ilişkinin, grafiksel olarak gösterimi ise "filogenetik ağaçlar" aracılığıyla olur. Organizmalar arası evrimsel ilişkileri gösteren bu filogeniler, "evrim ağacı" ya da "yaşam ağacı" olarak da bilinmektedir. Yaşam ağacı kavramı, tek atadan köken almış ve dallanarak farklılaşmış türleri tek bir konsept halinde göstermek için, ilk kez İngiliz biyolog Charles Darwin tarafından (1809-1882) evrim teorisi kapsamında kullanılmıştır.

1. 1. Çalışmanın Amacı ve Kapsamı

Bu çalışmanın amacı filojeni algoritmaların kodlanması ve bir arayüzde gösterilmesidir. Kullanılacak olan algoritmalar guide ağaçları, evrimsel ağaçlar (Neighbor-joining ve Upgma), parsimony algoritması, sankoff ve fitch algoritmaları'dır.

1.2. Filojeni Algoritmaları

1.2.1. Guide Ağaçları

Guide ağaçları, sınıflandırma ve regresyon ağaçları oluşturmak için çok amaçlı bir makine öğrenme algoritmasıdır. Wisconsin Üniversitesi, Madison'da Wei-Yin Loh tarafından tasarlanmış ve bakımı yapılmıştır. Kılavuz genelleştirilmiş, tarafsız, etkileşim algılama ve tahmin anlamına gelir.

Kılavuzun geliştirilmesi kısmen ABD Ordusu Araştırma Ofisi, Ulusal Bilim Vakfı, Ulusal Sağlık Enstitüleri, Çalışma İstatistikleri Bürosu ve Eli Lilly'den gelen araştırma hibeleri ile desteklenmektedir. Guide öncülleri üzerindeki çalışma ayrıca IBM ve Pfizer tarafından desteklenmiştir.

Özellikler:
Sınıflandırma ve regresyon ağaçları seçimi.
Bölünmüş değişken seçiminde ihmal edilebilir önyargı.
Önem sıralaması ve önemsiz değişkenlerin belirlenmesi.
Yordayıcı değişken çiftleri arasındaki yerel etkileşimleri tespit etme gücü.
Sıralı (sürekli) ve sırasız (kategorik) yordayıcı değişkenleri kullanma becerisi.
Eksiklik bölünmeleri dahil, eksik değerlerin otomatik olarak işlenmesi.
Yeni (görünmeyen) örnekler için otomatik tahmin.
Ağırlıklı en küçük kareler (Gauss), en küçük kareler medyanı, Poisson, nicelik (medyan dahil), orantılı tehlikeler veya çoklu yanıtlı (örn., Longitudinal) regresyon ağacı modelleri seçimi.
Parçalı sabit, en iyi basit polinom, çoklu veya aşamalı doğrusal regresyon modellerinin seçimi.
Tahmin değişkenleri için rol seçimi (yalnızca bölme, yalnızca düğüm modelleme, ikisi birden veya hiçbiri).
Yalnızca ayırmak için veya kukla 0-1 vektörlerle bölmek ve uydurmak için kategorik değişkenleri kullanma seçeneği (Ancova).

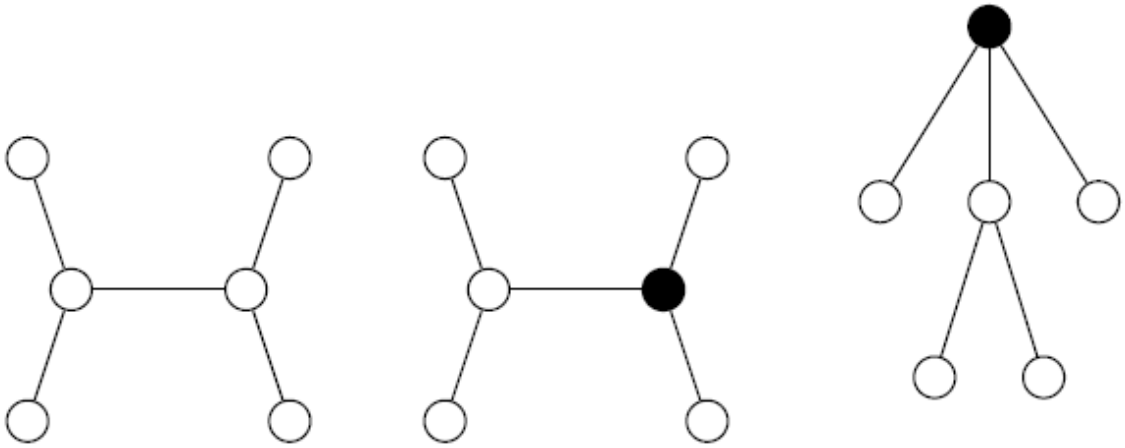
Durdurma kurallarının seçimi: budama yok, çapraz doğrulama ile budama veya test örneğiyle budama yok.
Toplu iş veya etkileşimli çalışma modu seçimi.
Ürünlerin anında üretilmesi ve regresör değişkenleri olarak öngörücü değişkenlerin yetkileri.
PostScript (.ps) formatında ağaç diyagramları için LaTeX (Windows için MikTeX) kaynak kodunun oluşturulması.
Gelecekteki vakaların tahmini için R kaynak kodunun oluşturulması.
Windows, Macintosh ve Linux bilgisayarlar için ücretsiz yürütülebilir dosyalar.

1.2.2. Evrim Ağaçları

Evrim Ağacı (Evolutionary trees) birkaç organizmanın dna dizilerinden oluşmuş bir ağaçtır. Yapraklar mevcut türleri temsil eder. İç köşeler ataları temsil eder. Kenarlar evrimsel adımları temsil eder. Ağaçta temsil edilen mevcut türlerde kök, en eski evrimsel atayı temsil eder.

1.2.2.1. Köklü ve Köksüz Ağaçlar

Köksüz bir ağaçta, kökün konumu (“en eski ata”) bilinmiyor. Aksi takdirde, ağaç köklenir ve kök ile zirvede yönlendirilebilir.



Şekil 1. Köklü ve köksüz ağaçlar

Köksüz ağaçta, her bir taksonun diğeriyle ilişkisi görülür ancak ortak ata tahmin edilemediği için evrimsel yönü yoktur.

Bu evrimsel ilişkiyi belirlemek için kullanılan köklü ağaçlarda ise taksonlar, ortak bir atadan köken alarak yerleştirilir. Bundan dolayı köklü ağaç, köksüz filogenetik ağaçlara oranla daha fazla bilgi sağlamaktadır.

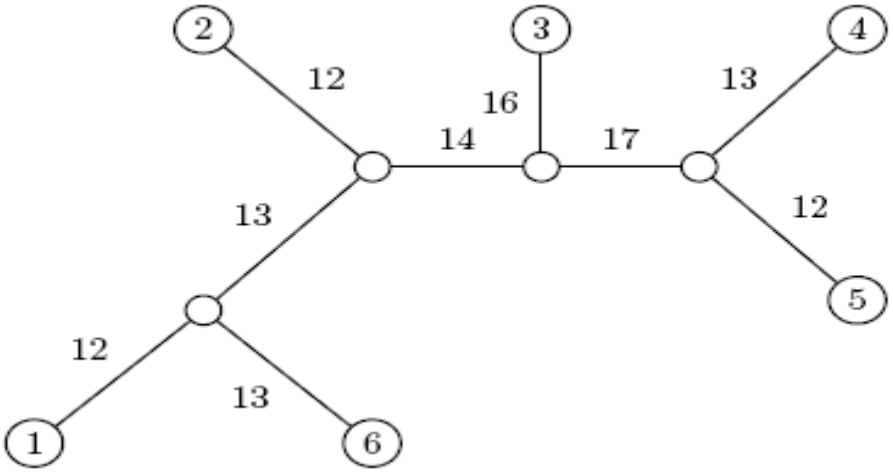
1.2.2.2. Ağaçlarda mesafe

Kenarlar şunları yansıtan ağırlıklara sahip olabilir:

Bir türden diğerine evrimsel yol üzerindeki mutasyonların sayısı.

Bir türün diğerine evrimi için zaman tahmini.

Bir T ağacında genellikle şunları hesaplarız: $d_{i,j}(T)$ = i ve j yaprakları arasındaki yolun uzunluğu ve $d_{i,j}(T)$ = i ve j arasındaki ağaç mesafesi.



Şekil 2. Ağaçlarda mesafe örneği

d1,4 = 12 +13 +14 +17 +13 = 69

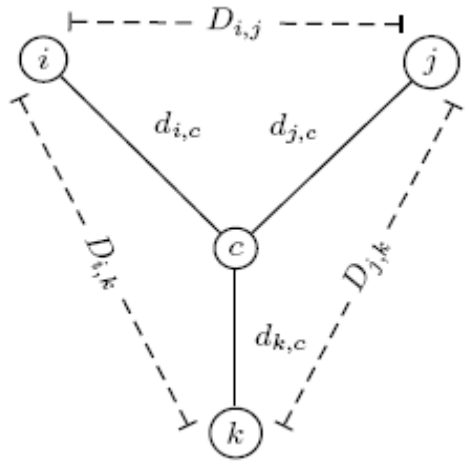
1.2.2.3. Mesafe Matrisi (Fitting)

N tür verildiğinde, nxn mesafeyi Di,j matrisiyle hesaplayabiliriz. Bu genlerin evrimi, bilinmeyen bir ağaçla tanımlanır. Di,j mesafe matrisine en iyi uyan bir ağacı oluşturmak için bir algoritmaya ihtiyaç vardır.

Fitting şu anlama gelir: Di,j = di,j(T). Di,j bilinen türler arasındaki mesafeyi düzenler. di,j(T) ise bilinmeyen bir T ağacının yol uzunluğunu ifade eder.

1.2.2.4. Üç Yapraklı Bir Ağacı Yeniden İnşa Etmek

Herhangi bir 3x3 matrisi için ağacı yeniden inşa etmek kolaydır. Bunun için 3 yaprak i, j, k ve bir c orta tepesine sahip olunur.



Şekil 3. Üç yapraklı bir ağacı yeniden inşa etmek

Gözlem:

d (i,c) + d (j, c) = D(i, j)

d (i,c) + d (k, c) = D(i, k)

d (j,c) +d (k, c) = D(j, k)

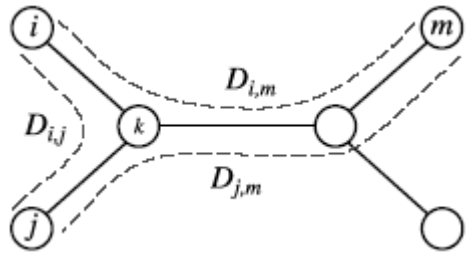
1.2.2.5. Neighbor-Joining Algoritması

Komşu birleştirme algoritmasıdır. Ağacı oluşturmak için komşu yapraklar kullanılır.

K ebeveyni ile i ve j komşu yaprakları bulunur. Daha sonra i ve j yaprakları k'ya sıkıştırılır:

i ve j nin satır ve sütunlarını kaldırılır.

k ye karşılık gelen yeni bir satır ve sütun eklenir, burada k'dan diğer herhangi bir m yapısına olan uzaklık aşağıdaki denklemle yeniden hesaplanabilir:

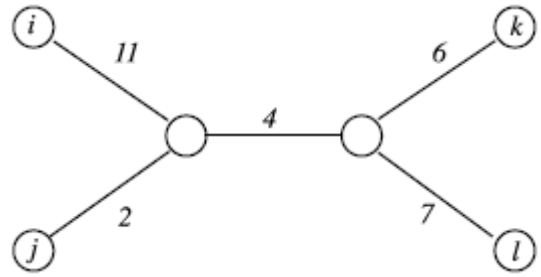


Şekil 4. Komşu birleştirme algoritması

$$D(k,m) = (D(i,m) + Dj,m - D(i,j))/2 \quad (1.1)$$

1.2.2.5.1. Komşu Yaprakları Nasıl Buluruz

Komşu yaprakları bulmak için ağaçtaki en yakın yaprakları seçmek doğru bir çözüm değildir. Ağaçtaki en yakın yaprakların komşu olması gerekmez.



Şekil 5. Komşu yaprakları bulma örneği

Şekil 5. de görüldüğü gibi i ve j komşudur, ancak d (i,j) = 13 > d (j,k) = 12'dir.

Bir çift komşu yaprağı bulmak önemli olmayan bir sorundur.

1987 yılında Naruya Saitou ve Masatoshi Nei, filogenetik ağaç rekonstrüksiyonu (yeniden yapılanma) için bir komşu birleştirme algoritması geliştirdi. Algoritmadaki fikir birbirine yakın fakat diğer yapraklardan uzak bir çift yaprak bulmak, ardından örtülü olarak bir komşu çift yaprak bulmaktır. Bunun avantajları eklemeli ve diğer eklemeli olmayan matrisler için iyi çalışmasıdır ve hatalı moleküler saat varsayımına sahip olmamasıdır.

1.2.2.6. Upgma

Aritmetik Ortalama ile Ağırlıksız Çift Grup Metodudur. Kümelemenin en basit ve en hızlı metodudur. Köklü ağaçlar oluşturan ultrametrik (kökten tüm uçlara eşit uzunlukta dalları olan) bir ağaçlandırma metottur. Verileri uzaklık bakımından algoritmik olarak düzenleyerek taksonları kümeleyen bu metot, bu uzaklığı elde ederken bir formül kullanır. En yakın iki taksonun gruplandırılmasından başlar. Artan uzaklık dikkate alınarak, tüm taksonları gruplamaya dayanır. Kademeli olarak uzaklık arttıkça taksonlar yeni gruplara girmeye ve birbirinden farklılaşmaya başlar. Neighbor Joining olan bu metot, verileri genetik uzaklık bakımından kümeleyerek analiz eder. Kümelendirme metodu, evrimsel saate dayandırılmadığı için köksüz ağaçlar oluşturulur. Köksüz filogenetik ağaç oluşturan en basit metottur.

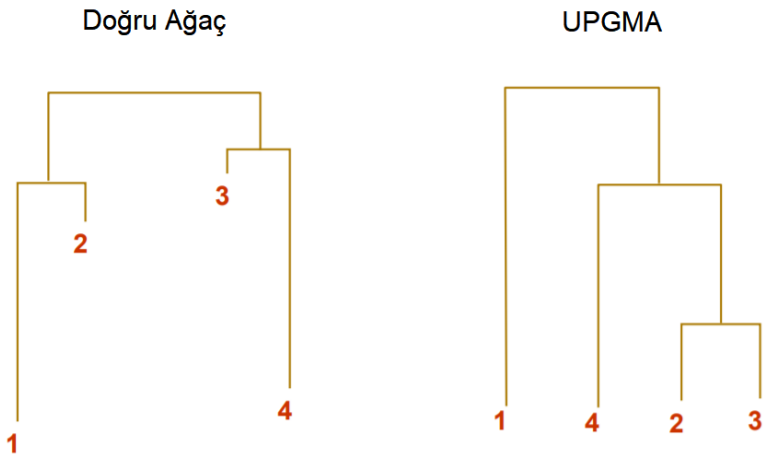
Sonuç olarak algoritmik uzaklık metotları; optimal ağacı bulamaz çünkü alternatif ağaçlar oluşturmaz tek tip ağaç oluşturur, oluşturulan ağaçta herhangi bir dal değiştirilemez, hızlı ve kolay olduğu için başlangıç analizleri için kullanışlıdır. Upgma köklü ve dal uzunlukları eşit ağaçlar oluşturan en hızlı ve basit yöntem iken, bu metodun köksüz ve dal uzunlukları farklı ağaçlar oluşturan formatı ise Neighbor Joining metodudur.

Tablo 1. Kümeleme tabanlı metotların karşılaştırılması

	UPGMA	**NJ**
Esas	Uzaklık, kümeleme	Uzaklık, kümeleme
Ağaç sayısı	Tek	Tek
Formatı	Köklü, kladogram	Köksüz, filogram
Neden bu metot?	Başlangıç analizi yapılacaksa köklü basit bir ağaç elde etmek istenirse	Başlangıç analizi yapılacaksa köksüz basit bir ağaç elde etmek istenirse

Upgma ortalama ikili mesafeyi kullanarak kümeler arasındaki mesafeyi hesaplar. Ağaçtaki her tepe noktasına bir yükseklik atar, etkili bir şekilde moleküler bir saatin olduğunu varsayar ve her tepe noktasına tarih verir.

Upgma'nın zayıf yönü ultrametrik bir ağaç üretmesidir. Kökten herhangi bir yaprağa olan mesafe aynıdır. Bunun nedeni, Upgma'nın sabit bir moleküler saat varsayımıdır. Ağaçtaki yapraklarla temsil edilen tüm türlerin aynı hızda mutasyon biriktirdiği (ve dolayısıyla geliştiği) varsayılır. Bu, Upgma'nın önemli bir tuzağıdır.



Şekil 6. Upgma örneği

1.2.2.6.1. Upgma'da Kümeleme

Ci ve Cj dizilerinin iki ayrık kümesi verildiğinde;

$$d_{i,j} = \frac{1}{|C_i| \cdot |C_j|} \sum_{p \in C_i} \sum_{q \in C_j} d_{p,q}$$

Şekil 7. Upgma kümeleme

Ck, Ci ve Cj'nin birleşimi ise, Ck'den başka bir Cl kümesine olan mesafe:

$$d_{k,l} = \frac{d_{i,l}\left|C_i\right| + d_{j,l}\left|C_j\right|}{\left|C_i\right| + \left|C_j\right|}$$

Şekil 8. Upgma kümeleme 2

1.2.2.6.2. Upgma Algoritması

3 aşamalıdır:

1. Başlatma:

- Her xi kendi Ci kümesine atanır.

- Sıra başına her biri 0 yüksekliğinde bir yaprak tanımlanır.

2. Yineleme:

- (di,j)'nin minimum olacağı şekilde iki Ci ve Cj kümesi bulunur.

- Ck = Ci ∪ Cj olsun.

- Ci'yi Cj'ye bağlayan tepe noktası eklenir ve di (j/2) yüksekliğine yerleştirilir.

- Ci ve Cj silinir.

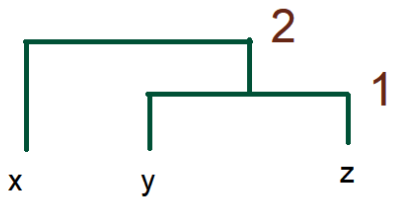
3. Sonlandırma (Termination):

- Tek bir küme kaldığında işlem sonlandırılır.

	v	w	x	y	z
v	0	6	8	8	8
w		0	8	8	8
x			0	4	4
y				0	2
z					0

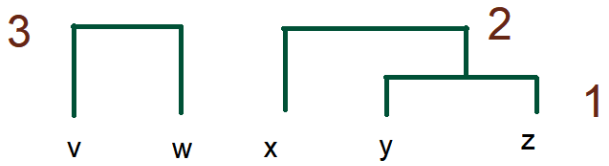
Şekil 9. Upgma algoritması örneği

	v	w	x	yz
v	0	6	8	8
w		0	8	8
x			0	4
yz				0



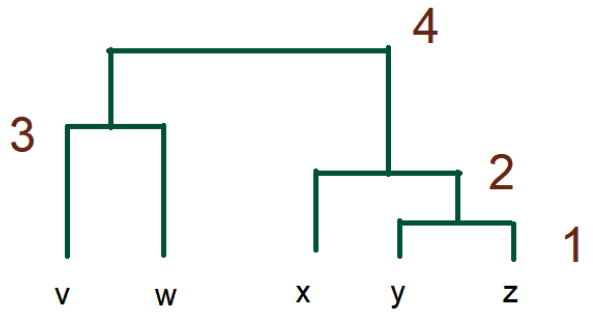
Şekil 10. Upgma algoritması örneği 2

	v	w	xyz
v	0	6	8
w		0	8
xyz			0



Şekil 11. Upgma algoritması örneği 3

	vw	xyz
vw	0	8
xyz		0



Şekil 12. Upgma algoritması örneği 4

1.2.3. Karakter Tabanlı Filogeni

1.2.3.1. Hizalama Matrisi ve Uzaklık Matrisi

N türdeki m nükleotitlerden oluşan bir geni sıralayarak bir n x m hizalama matrisi. Verilen bir hizalama matrisinin mesafe matrisine dönüştürülebileceğini sadece edit/Hamming mesafesini dikkate alarak görülebilir. Bununla birlikte, belirli bir matris benzersiz bir hizalama matrisine dönüştürülemez, çünkü bir hizalama matrisini mesafe matrisine dönüştürme sırasında bilgi kaybolur.

1.2.3.2. Karakter Tabanlı Ağaç Yeniden Yapılandırması

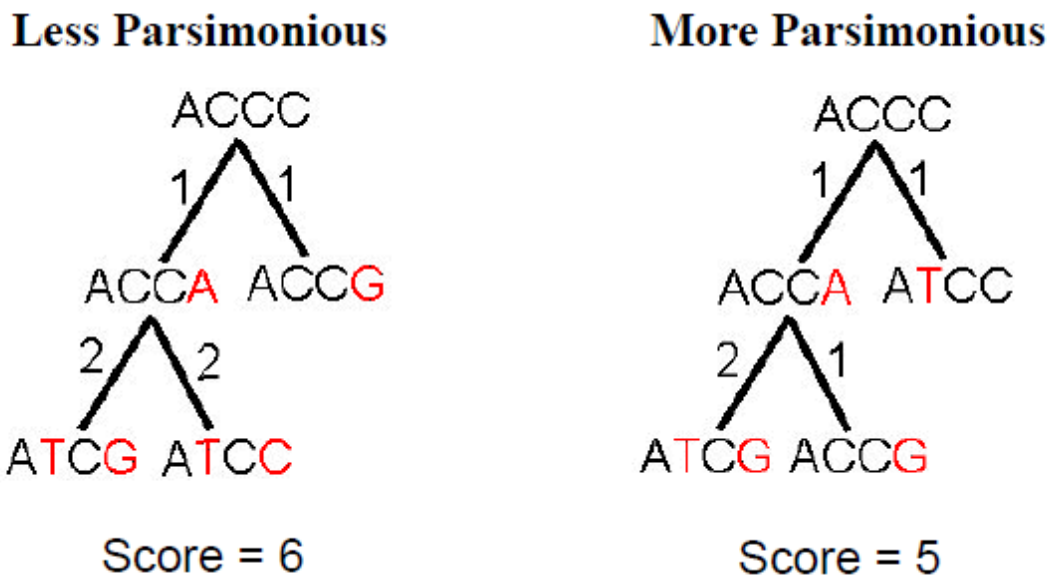
Karakter tabanlı yeniden yapılandırma algoritmaları n x m hizalama matrisini kullanır. Bu daha iyi bir tekniktir. (n = tür sayısı, m = karakter sayısı). Mesafe matrisini kullanmak yerine doğrudan hizalama matrisi kullanılır. Genel olarak amaç dahili düğümlerde n gözlemlenen türleri için karakter dizilerini en iyi şekilde açıklayan karakter dizileri belirlenir. Karakterler nükleotitler olabilir (A, T, C, G). Ağaçtaki bir kenarın uzunluğunu Hamming mesafeye ayarlayarak ağacın parsimony (cimrilik) puanını şu şekilde tanımlayabiliriz: Kenarların uzunluklarının (ağırlıklarının) toplamı.

1.2.4. Küçük Parsimony Problemi

Parsimony: Occam'ın ustura prensibini uygular: Evrimi açıklayan birçok olası ağaç ilişkileri vardır. Veriler için en basit açıklamayı belirlenir. Teknik olmayan terime göre ise basit, aptal olan şey olsun denir. Parsimony bu nedenle gözlemlenen karakter farklılıklarını olası en az mutasyondan kaynaklandı diye varsayar. Olası en düşük cimrilik(parsimony) puanını veren ağacı arar: Ağaçta bulunan tüm mutasyonların maliyetinin toplamı.

1.2.4.1. Ağırlıklı Küçük Parsimony Problemi

Küçük parsimony probleminin daha genel bir versiyonudur. Girdi, k durumlarının her birinin bir diğerine dönüşüm maliyetini açıklayan bir k x k puanlama matrisi içerir.



Şekil 13. Parsimony örneği

Girdi: Her yaprağın bir m karakter dizesiyle etiketlendiği T Ağacı.

Çıktı: T'nin iç köşelerinin parsimony skoru ile etiketlenmesi. Her yaprağın tek bir karakterle etiketlendiğini (labeling) varsayabiliriz, çünkü dizedeki karakterler bağımsızdır. Karışıklığı önlemek için, "en çok cimri ”etiketleme en küçük cimri puanını sağlayacaktır. Küçük Parsimony problemi için puanlama matrisi basitçe hatırlanan Hamming mesafesine göre: dH (v,w) v = w ise 0’ dır. Diğer durumlarda ise 1’dir.

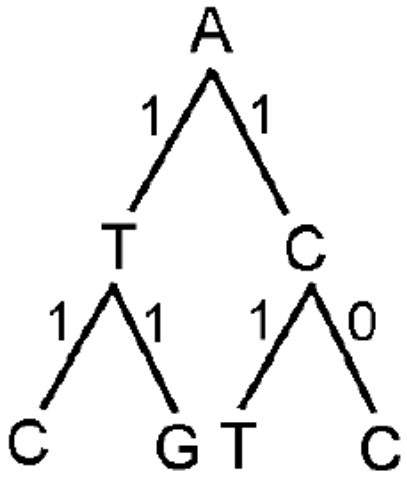
Small Parsimony Problem

	A	T	G	C
A	0	1	1	1
T	1	0	1	1
G	1	1	0	1
C	1	1	1	0

Weighted Parsimony Probler

	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

Şekil 14. Puanlama matrisleri örneği



Small Parsimony Scoring Matrix

	A	T	G	C
A	0	1	1	1
T	1	0	1	1
G	1	1	0	1
C	1	1	1	0

Small Parsimony Score: 5

Şekil 15. Ağırlıklı ve ağırlıksız örneği

Girdi: Her yaprağın k harfinin öğeleriyle etiketlendiği T ağacı alfabe ve bir k x k puanlama matrisi (δi, j).

Çıktı: T'nin iç köşelerinin ağırlıklı cimrilik puanı ile etiketlenmesi.

1.2.5 Maksimum parsimony

Bir gözlemin en az karmaşık olarak açıklanması "parsimoni" şeklinde tanımlanır. Bu nedenle bu ağaçlandırma yöntemi de incelenen diziler ile uyumlu bir ağaç elde etmek için gerekli en az mutasyonların saptanmasını esas alır. Başka bir deyişle "tutumluluk" olarak tanımlanabilme yani, biyolojik değişim süreci boyunca karmaşıklık yerine basit bir açıklama yaparak verilerin yorumlanmasıdır. Maksimum parsimoni yöntemi uygulanırken, dizi pozisyonlarının farklı puanlamaları tercih edilebilir. Örneğin; korunmuş bölgede gerçekleşen bazı mutasyonlar, değişken bölgedeki mutasyonlardan daha çok vurgulanmak istenebilir. Ya da transversiyonlar transisyonlardan daha önemli olarak vurgulanabilir. Maksimum parsimoni ile ağaçların oluşturulmasında 'kesin' ve 'tahmini' yaklaşımlar söz konusu olmaktadır. Çok zaman alıcıdır ve çok sayıda örnek ele alındığında kullanıma uygun değildir.

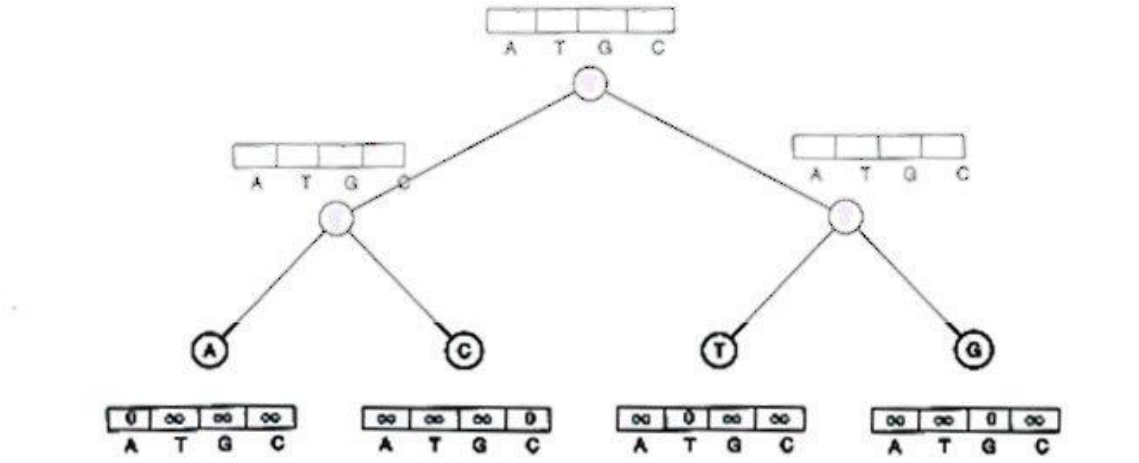
1.2.6. Sankoff ve Fitch algoritmaları

Dinamik programlama: Olası her etiket için bir puan hesaplanır ve her köşe takip edilir. st (v) = köşe v de köklü alt ağacın minimum cimrilik puanı (v, t karakterine sahipse). Her köşedeki puan, çocuklarının puanlarına dayanmaktadır:

$$s_t(\text{parent}) = \min_i \{ s_i(\text{left child}) + \delta_{i,t} \} + \min_j \{ s_j(\text{right child}) + \delta_{j,t} \}$$

Şekil 16. Dinamik programlama formül

Bu nedenle, tüm ağacın cimrilik puanı, kökteki puan olacaktır. Sankoff algoritması: Yapraklarda başlanır. Yaprak söz konusu karaktere sahipse, puan 0'dır. Değilse sonsuz olur.



Şekil 17. Sankoff algoritma örneği

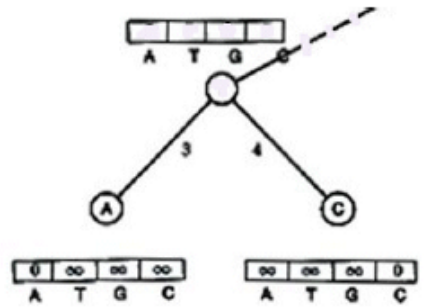
Ardından, tabloda bir seviye yukarı çıkmak için dinamik programlama formülleri uygulanır.

δ	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

$$s_t(v) = \min_i \{ s_i(u) + \delta_{i,t} \} + \min_j \{ s_j(w) + \delta_{j,t} \}$$

$$s_A(v) = 0 + \min_j \{ s_j(w) + \delta_{j,A} \}$$

	$s_i(u)$	$\delta_{i,A}$	sum
A	0	0	0
T	∞	3	∞
G	∞	4	∞
C	∞	9	∞



Şekil 18. Sankoff algoritma örneği 2

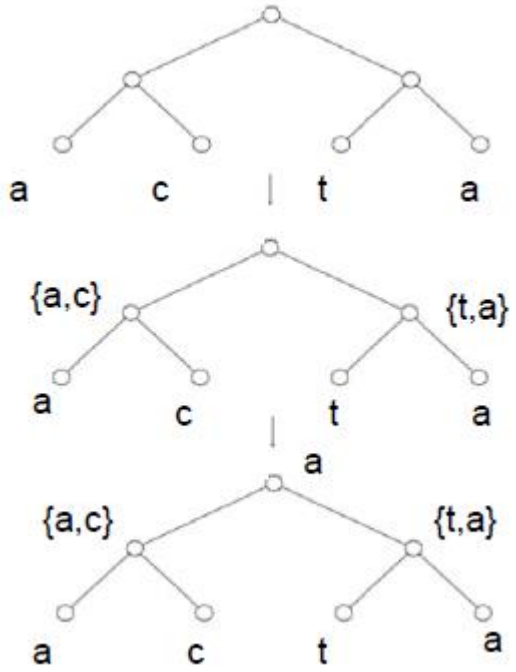
Sağ alt ağaç için tekrarlanır. Kök için tekrarlanır. Kökteki en küçük puan, minimum ağırlıklı cimrilik puanıdır. Bu durumda puan = 9 olur.

1.2.6.1. Ağaçtan Aşağı gezinme

Kök tepe noktasındaki puanlar ağaçta yukarı çıkılarak hesaplanmıştır. Kök tepe noktasındaki puanlar hesaplandıktan sonra, Sankoff algoritması ağaçta aşağı doğru hareket eder ve her bir tepe noktasına optimal bir karakter atar.

1.2.6.2. Fitch Algoritması

Küçük parsimoni problemini çözer. Ağaçtaki her tepe noktasına bir dizi harf atar. Her yaprak, gözlenen karakteriyle etiketlenecektir. Ağaçta yukarı çıkarken, her bir ebeveyn tepe noktası için: İki çocuğun alt karakter kümesi çakışırsa, ebeveynin karakter kümesi, her ikisinde de ortak olan karakter kümesidir. Aksi takdirde, ebeveynin karakter kümesi çocuk karakterlerinden birleşik kümedir.



Şekil 19. Fitch algoritması örneği

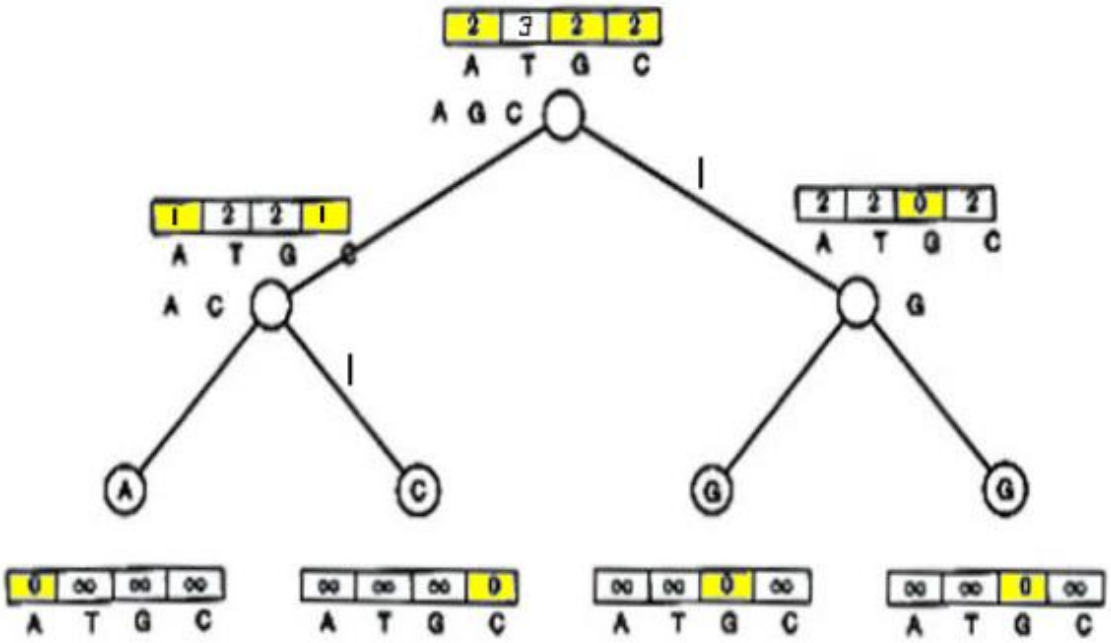
Daha sonra, ağaçtan geçerek kökten yapraklara her bir tepe noktasına etiketler atanır. Köke rastgele bir harf kümesinden bir etiket atanır. Diğer tüm köşeler için, ebeveyn etiketi karakterler içinde ise, ona ebeveynin etiketi atanır. aksi durumda kendi etiket kümesinden rastgele bir karakter seçilir. Fitch ve sankoff için de çalışma zamanı O(nk) dır. Farklı algoritmalar mıdır?

Fitch algoritması için puanlama matrisi yalnızca:

Tablo 2. Fitch algoritma puanlama matrisi

	A	T	G	C
A	0	1	1	1
T	1	0	1	1
G	1	1	0	1
C	1	1	1	0

Sankoff için puanlama matrisi ise:



Şekil 20. Sankoff puanlama matrisi

Sankoff algoritması için, karakter t, köşe v için optimaldir eğer st (v) = min1<i<k si(v) ise. V köşesindeki en uygun harf kümesi S (v) olarak belirtilir. S (sol çocuk) ve S (sağ çocuk) çakışırsa, S (ebeveyn) keşisimleridir. Aksi takdirde S (sol çocuk) ve S (sağ çocuk) birleşimidir. Bu fitch ile aynıdır. Sonuç olarak iki algoritma da aynıdır.

2. KAYNAK ARAŞTIRMASI

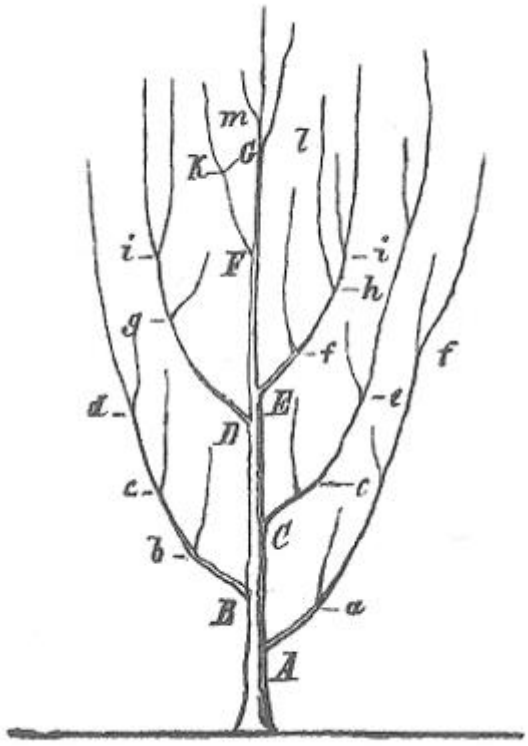
"Filojeni" terimi, Haeckel tarafından 1866'da ortaya atılan Alman Filogenilerinden türetilmiştir ve Darwinci sınıflandırma yaklaşımı, "filojenik" yaklaşım olarak bilinir hale gelmeye başlamıştır.

19. yüzyılın sonlarında, Ernst Haeckel'in özetleme teorisi veya "biyogenetiğin temel yasası" geniş çapta kabul görmüştür. Bu genellikle "birey oluş, soy oluşu (filogeniyi) yeniden özetler" olarak ifade edildi, yani tek bir organizmanın yaşamı boyunca, tohumdan yetişkine gelişimi, ait olduğu türün birbirini izleyen atalarının yetişkin aşamalarını ardışık olarak yansıtmaktadır. Fakat bu teori uzun bir süre boyunca reddedilmiştir. Bunun yerine "ontogenez" gelişir. Haeckel'in düşündüğü gibi, bir türün filogenetik tarihi ontogenezden direk olarak okunamaz, ancak ontogenezden gelen karakterler filogenetik analizler için veri olarak kullanılabilir ve de kullanılabilmiştir. Eğer iki tür ne kadar yakından ilişkili ise; embriyoların paylaştığı apomorfiler de o kadar fazladır. (Apomorfi: Bir dalda belli bir karakterin değişmiş ve türemiş biçimine apomorfi denir.)

Önemli (Kilit) Noktaların Zaman Çizelgesi:

- 14. yüzyılda, lex parsimoniae (parsimony (cimrilik) ilkesi), İngiliz bir filozof, Fransiskan papazı ve ilahiyatçı olan Ockhamlı William tarafından ortaya atılmıştır. Fakat sonradan bu fikirler Aristoteles tarafından benimsenmiştir.
- 1763 yılında Thomas Bayes tarafından "Bayesci Olasılık" (Bayesian Probability) kavramı ortaya atıldı.
- 18. yüzyılda, Pierre Simon (Marquis de Laplace), muhtemelen ML (maximum likelihood) (maksimum olasılık)'ı kullanan ilk kişi oldu.
- 1809 yılında, Jean-Baptiste de Lamarck tarafından yazılan "Philosophie Zoologique" adlı evrim teorisi konulu kitabına göre 17. ve 18. yüzyıllarda Voltaire, Descartes ve Leibniz birçok türün ortaya çıktığını düşündüren gözlemlenen boşlukları hesaba katmak için evrimsel değişiklikleri önermişlerdir. Aynı zamanda M.Ö 6. yüzyılda Anaximander gibi bazı erken Yunan filozofları bu gibi teorileri önerdikleri tahmin edilmektedir.
- 1837'de Darwin tarafından yazılan defterdeki sayfalar bir evrim ağacını göstermektedir.

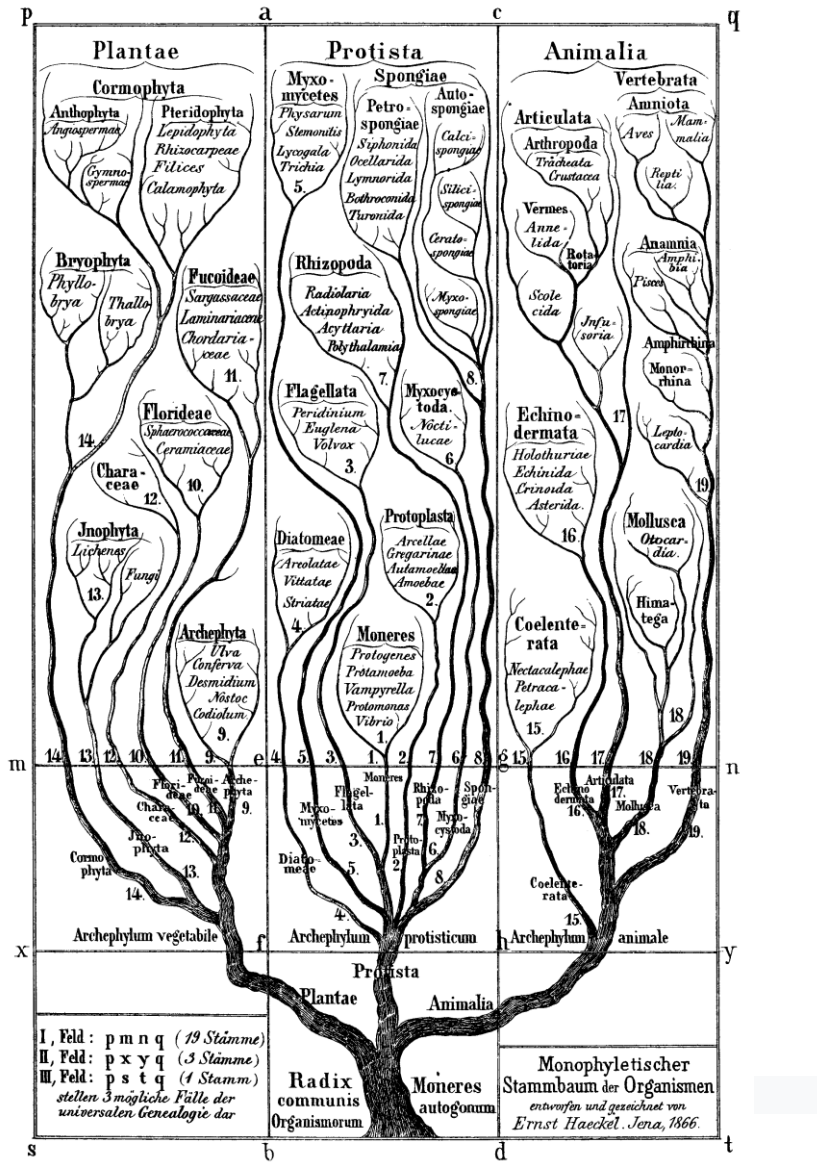
- 1843 yılında Richard Owen tarafından homoloji ve analoji arasındaki ayrımları belirleme çalışmaları yapılmıştır. Şu anda analoji yerine "homoplazi" terimi kullanılmaktadır.
- 1858 yılında, Paleontolog Heinrich Georg Bronn (1800-1862), eski bir türün neslin tükenmesinin ardından, benzer türlerin paleontolojik gelişimini gösteren varsayımsal bir ağaç yayınladı. Bronn, bu tür olaylardan sorumlu bir mekanizma önermedi.



Şekil 21. Heinrich Georg Bronn tarafından oluşturulan dallanma ağacı diyagramı (1858)

- 1858 yılında, Darwin ve Wallace evrim teorisi ve doğal seçilim üzerine beraber çalışmışlar ve aynı zamanda bunların üzerine beraberce tez yazıp yayımlamışlardır.
- 1866 yılında, Ernst Haeckel filogeniye dayalı evrim ağacını ilk kez yayımlamıştır.
- 1893 yılında, Dollo'nun Karakter Durumu Geri Dönülemezlik Yasası (Dollo's Law of Character State Irreversibility) ortaya atıldı.
- 1912 yılında, ML (maximum likelihood) (maksimum olasılık), Ronald Fisher tarafından önerilip, ardından bunun üzerine analizler yapıldı ve popülerleştirildi.

- 1921 yılında, Tillyard “filogenetik” terimini kullanmaya ve kendi sınıflandırma sisteminde arkaik ve özel karakterler arasında ayrım yapmaya başlar. (Arkaik**:** Kullanıldığı çağdan daha eski bir çağa ait olan.)
- 1940 yılında, Lucien Cuénot tarafından “clade” terimi ortaya atıldı.
- 1949 yılında Maurice Quenouille tarafından Jackknife resampling (Jackknife yeniden örnekleme) çalışması yapıldı. (Bu örnekleme çalışması 1946 yılında Mahalanobis tarafından öngörülmüş ve 1958 yılında Tukey tarafından uzatılmıştır.)
- 1950 yılı, Willi Hennig’in klasik biçimlendirmesi. (Willi Hennig’s classic formalization)
- 1952 yılı, William Wagner’in zemin planı sapma yöntemi (William Wagner’s ground plan divergence method



Şekil 22. Haeckel tarafından önerilen filogenetik ağaç (1866)

- 1953 yılında “klagogenez” (cladogenesis) icat edildi.
- 1960 yılında Cain ve Harrison tarafından “kladistik” (cladistic) terimi ortaya atıldı.
- 1963 yılında, Edwards ve Cavalli-Sforza filogenetik için ML (maximum likelihood) (maksimum olasılık)’ı ilk kez kullanmayı denemişlerdir.
- 1965 yıllarında,
    - Camin-Sokal parsimony (cimrilik), ilk parsimony (optimizasyon) kriteri ve hem Camin hem de Sokal tarafından kladistik analiz için kullanılan ilk bilgisayar programı ve algoritması.

    - Klik analizi (clique analysis) olarak da adlandırılan karakter uyumluluk yöntemi, Camin, Sokal ve E. O. Wilson tarafından bağımsız olarak sunulmuştur.

- 1966 yıllarında,
    - Hennig'in İngilizce çevirisi
    - "cladistics" ve "cladogram" terimlerinin ortaya atılması
- 1969 yıllarında,
    - James Farris'in dinamik ve ardışık ağırlıklandırması (dynamic and successive weighting)
    - Kluge ve Farris'in Wagner parsimony'u (cimriliği)
    - Kluge ve Farris'in CI tutarlılık indeksi (consistency index)
    - Le Quesne'nin klik analizi (clique analysis) için ikili uyumluluğunun tanıtımı
- 1970 yılında, Farris, Wagner parsimony'u genişletmiş ve genelleştirmiştir.
- 1971 yıllarında,
    - Neyman tarafından ML (maksimum olasılık)'ın filogenetiğe ilk kez başarılı uygulaması yapıldı. (protein dizileri için)
    - Fitch tarafından ortaya atılan Fitch parsimony.
    - Robinson, Moore ve arkadaşları tarafından bağımsız olarak geliştirilen ilk dal dal değiştirme arama stratejisi (first branch-swapping search strategy) olan NNI (en yakın komşu değişimi) (nearest neighbour interchange).
    - Kidd ve Sgaramella-Zonta'nın ME (minimum evolution) (minimum evrim)'i (Bunun ikili mesafe yöntemi olup olmadığı veya Edwards ve Cavalli-Sforza'nın ML'yi (maksimum olasılık) minimum evrim olarak adlandırması nedeniyle ML ile ilişkili olup olmadığı belirsizdir.)
- 1972, Adams konsensüsü, Adams
- 1976, rütbeler için önek sistemi, Farris
- 1977, Dollo parsimony, Farris
- 1979
    - Nelson fikir birliği, Nelson

    - MAST (maksimum anlaşma alt ağacı) ((GAS) en büyük anlaşma alt ağacı), bir konsensüs yöntemi, Gordon
    - bootstrap, Bradley Efron, öncül konsept
    - 1980, PHYLIP, filogenetik analiz için ilk yazılım paketi, Felsenstein
- 1981
    - çoğunluk fikir birliği, Margush ve MacMorris
    - katı fikir birliği, Sokal ve Rohlf
    - hesaplama açısından verimli ilk makine öğrenimi algoritması, Felsenstein
- 1982
    - PHYSIS, Mikevich ve Farris
    - dal ve sınır, Hendy ve Penny
- 1985
    - kombine fenotipik ve genotipik kanıtlara dayanan ilk kladistik ökaryot analizi Diana Lipscomb
    - *Cladistics'in* ilk sayısı
    - bootstrap'in ilk filogenetik uygulaması, Felsenstein
    - jackknife'ın ilk filogenetik uygulaması, Scott Lanyon
- 1986, MacClade, Maddison ve Maddison
- 1987, komşu birleştirme yöntemi Saitou ve Nei
- 1988, Hennig86 (sürüm 1.5), Farris
    - Bremer desteği (çürüme indeksi), Bremer
- 1989
    - RI (tutma indeksi), RCI (yeniden ölçeklendirilmiş tutarlılık indeksi), Farris
    - HER (homoplazi fazlalık oranı), Archie
- 1990
    - birleştirilebilir bileşenler (yarı katı) fikir birliği, Bremer
    - SPR (alt ağaç budama ve yeniden greftleme), TBR (ağacın ikiye bölünmesi ve yeniden bağlanması), Swofford ve Olsen
- 1991
    - DDI (veri kararlılık indeksi), Goloboff

    - Ökaryotların yalnızca fenotipik kanıtlara dayanan ilk kladistik analizi, Lipscomb
- 1993, Goloboff ağırlıklandırması
- 1994, azalan fikir birliği: Köklü ağaçlar için RCC (azaltılmış kladistik fikir birliği), Wilkinson
- 1995, köksüz ağaçlar için azaltılmış fikir birliği RPC'si (azaltılmış bölüm fikir birliği), Wilkinson
- 1996, Li, Mau, ve Rannala ve Yang tarafından bağımsız olarak geliştirilen ve tümü MCMC (Markov zinciri-Monte Carlo) kullanılarak BI (Bayesian Inference) için ilk çalışma yöntemleri
- 1998, TNT (Yeni Teknolojiyi Kullanan Ağaç Analizi), Goloboff, Farris ve Nixon
- 1999, Winclada, Nixon
- 2003, simetrik yeniden örnekleme, Goloboff
- 2004,2005, benzerlik ölçüsü (Kolmogorov karmaşıklığına bir yaklaşım kullanarak) veya NCD (normalize kompresyon mesafesi), Li ve diğerleri, Cilibrasi ve Vitanyi.

3. MATERYAL VE YÖNTEM

3.1. Neighbor Joining Algoritması Yöntemi

K ebeveyni ile i ve j komşu yaprakları bulunur. Daha sonra i ve j yaprakları k'ya sıkıştırılır:

i ve j nin satır ve sütunlarını kaldırılır.

k ye karşılık gelen yeni bir satır ve sütun eklenir, burada k'dan diğer herhangi bir m yapısına olan uzaklık aşağıdaki denklemle yeniden hesaplanabilir.

$$D(k,m) = (D(i, m) + Dj,m - D(i, j))/2 \quad (1.1)$$

3.2 Upgma Algoritması Yöntemi

3 aşamalıdır:

1. Başlatma:

- Her xi kendi Ci kümesine atanır.
- Sıra başına her biri 0 yüksekliğinde bir yaprak tanımlanır.

2. Yineleme:

- (di,j)'nin minimum olacağı şekilde iki Ci ve Cj kümesi bulunur.
- $Ck = Ci \cup Cj$ olsun.
- Ci'yi Cj'ye bağlayan tepe noktası eklenir ve di (j/2) yüksekliğine yerleştirilir.
- Ci ve Cj silinir.

3. Sonlandırma (Termination):

- Tek bir küme kaldığında işlem sonlandırılır.

3.3. Parsimoni yöntemi

Girdi: Her yaprağın bir m karakter dizesiyle etiketlendiği T Ağacı.

Çıktı: T'nin iç köşelerinin parsimony skoru ile etiketlenmesi. Her yaprağın tek bir karakterle etiketlendiğini (labeling) varsayabiliriz, çünkü dizedeki karakterler bağımsızdır. Karışıklığı önlemek için, "en çok cimri "etiketleme en küçük cimri puanını sağlayacaktır. Küçük Parsimony problemi için puanlama matrisi basitçe hatırlanan Hamming mesafesine göre: dH (v,w) v = w ise 0' dır. Diğer durumlarda ise 1'dir.

3.4. Sankoff Algoritması Yöntemi

Sankoff algoritması için, karakter t, köşe v için optimaldir eğer st (v) = min1<i<k si(v) ise. V köşesindeki en uygun harf kümesi S (v) olarak belirtilir. S (sol çocuk) ve S (sağ çocuk) çakışırsa, S (ebeveyn) keşisimleridir. Aksi takdirde S (sol çocuk) ve S (sağ çocuk) birleşimidir. Bu fitch ile aynıdır.

3.5. Guide Ağaçları Yöntemleri

Sınıflandırma ve regresyon ağaçları seçimi.

Bölünmüş değişken seçiminde ihmal edilebilir önyargı.

Önem sıralaması ve önemsiz değişkenlerin belirlenmesi.

Yordayıcı değişken çiftleri arasındaki yerel etkileşimleri tespit etme gücü.

Sıralı (sürekli) ve sırasız (kategorik) yordayıcı değişkenleri kullanma becerisi.

Eksiklik bölünmeleri dahil, eksik değerlerin otomatik olarak işlenmesi.

Yeni (görünmeyen) örnekler için otomatik tahmin.

Ağırlıklı en küçük kareler (Gauss), en küçük kareler medyanı, Poisson, nicelik (medyan dahil), orantılı tehlikeler veya çoklu yanıtlı (örn., Longitudinal) regresyon ağacı modelleri seçimi.

Parçalı sabit, en iyi basit polinom, çoklu veya aşamalı doğrusal regresyon modellerinin seçimi.

Tahmin değişkenleri için rol seçimi (yalnızca bölme, yalnızca düğüm modelleme, ikisi birden veya hiçbiri).

Yalnızca ayırmak için veya kukla 0-1 vektörlerle bölmek ve uydurmak için kategorik değişkenleri kullanma seçeneği (Ancova).

Durdurma kurallarının seçimi: budama yok, çapraz doğrulama ile budama veya test örneğiyle budama yok.

Toplu iş veya etkileşimli çalışma modu seçimi.

Ürünlerin anında üretilmesi ve regresör değişkenleri olarak öngörücü değişkenlerin yetkileri.

4. ARAŞTIRMA SONUÇLARI

Yapılan araştırmanın sonucuna göre sankoff ve fitch algoritmaları aynıdır. İkisi de dinamik programlama kullanır. Fitch ve sankoff algoritmaları için de çalışma zamanı O(nk) dır. Aynı sonucu verirler fakat çalışma şekilleri farklıdır. Fitch algoritması küçük parsimoni problemini çözer. Sankoff algoritması ağaçta aşağı doğru hareket eder ve her bir tepe noktasına optimal bir karakter atar.

Parsimoni; küçük parsimoni, ağırlıklı küçük parsimoni ve büyük parsimoni problemleri şeklinde ayrılır. Parsimoni olası en düşük cimrilik(parsimony) puanını veren ağacı arar. Maksimum parsimoni yöntemi uygulanırken, dizi pozisyonlarının farklı puanlamaları tercih edilebilir. Maksimum parsimoni ile ağaçların oluşturulmasında 'kesin' ve 'tahmini' yaklaşımlar söz konusu olmaktadır. Çok zaman alıcıdır ve çok sayıda örnek ele alındığında kullanıma uygun değildir. Ağırlıklı küçük parsimoni problemi, küçük parsimoni probleminin daha genel bir versiyonudur.

Evrim Ağacı (Evolutionary trees) birkaç organizmanın dna dizilerinden oluşmuş bir ağaçtır. Neighbor-Joining algoritması komşu birleştirme algoritmasıdır. Ağacı oluşturmak için komşu yapraklar kullanılır. Upgma aritmetik ortalama ile ağırlıksız çift grup metodudur. Kümelemenin en basit ve en hızlı metodudur. Verileri uzaklık bakımından algoritmik olarak düzenleyerek taksonları kümeleyen bu metot, bu uzaklığı elde ederken bir formül kullanır. En yakın iki taksonun gruplandırılmasından başlar. Artan uzaklık dikkate alınarak, tüm taksonları gruplamaya dayanır. Upgma köklü ve dal uzunlukları eşit ağaçlar oluşturan en hızlı ve basit yöntem iken, bu metodun köksüz ve dal uzunlukları farklı ağaçlar oluşturan formatı ise Neighbor Joining metodudur.

Upgma ortalama ikili mesafeyi kullanarak kümeler arasındaki mesafeyi hesaplar. Ağaçtaki her tepe noktasına bir yükseklik atar, etkili bir şekilde moleküler bir saatin olduğunu varsayar ve her tepe noktasına tarih verir. Upgma'nın zayıf yönü ultrametrik bir ağaç üretmesidir. Kökten herhangi bir yaprağa olan mesafe aynıdır. Bunun nedeni, Upgma'nın sabit bir moleküler saat varsayımıdır. Ağaçtaki yapraklarla temsil edilen tüm türlerin aynı hızda mutasyon biriktirdiği (ve dolayısıyla geliştiği) varsayılır. Bu, Upgma'nın önemli bir tuzağıdır. Guide ağaçları, sınıflandırma ve regresyon ağaçları oluşturmak için çok amaçlı bir makine öğrenme algoritmasıdır.

5. KAYNAKLAR

URL 1. https://biotechgo.org/tr/?view=article&id=247:lo2&catid=136

URL 2. http://pages.stat.wisc.edu/~loh/guide.html#:~:text=GUIDE%20is%20a%20multi%2Dpurpose,Unbiased%2C%20Interaction%20Detection%20and%20Estimation.

URL 3. http://compeau.cbd.cmu.edu/wp-content/uploads/2016/08/Ch10_MolEvo.pdf

URL 4. https://en.wikipedia.org/wiki/Maximum_parsimony_(phylogenetics)

URL 5. https://almob.biomedcentral.com/articles/10.1186/1748-7188-7-9#:~:text=The%20parsimony%20approach%20seeks%20a,the%20weights%20on%20the%20edges

URL 6. http://pages.stat.wisc.edu/~loh/treeprogs/guide/eqr.pdf

URL 7. https://medium.com/@sddkal/dizi-hizalama-algoritmalar%C4%B1-946de50f24d8

URL 8. https://biotechgo.org/tr/?view=article&id=247:lo2&catid=136#blast-kullanarak-yerel-hizalama-tabanl%C4%B1-arama

URL 9. https://dergipark.org.tr/tr/download/article-file/627347

URL 10. https://en.wikipedia.org/wiki/Phylogenetics#History

6. TEŞEKKÜR

Bu çalışmanın gerçekleştirilmesinde, bir dönem boyunca değerli bilgilerini bizlerle paylaşan, kullandığı her kelimenin hayatımıza kattığı önemi asla unutmayacağımız saygıdeğer danışman hocamız; Doç. Dr. Gıyasettin ÖZCAN'a sonsuz teşekkürlerimizi sunarız.

Alparslan YÜCE
Kerem AKIN
Alihan SULTAN

BURSA 2020

7. ÖZGEÇMİŞ

Adı Soyadı	: Alparslan YÜCE
Doğum Yeri ve Tarihi	: Amasya 05.06.1998
Lisans	: Uludağ Üniversitesi Bilgisayar Mühendisliği

Adı Soyadı	: Alihan SULTAN
Doğum Yeri ve Tarihi	: Bursa 15.06.1995
Lisans	: Uludağ Üniversitesi Bilgisayar Mühendisliği

Adı Soyadı	: Kerem AKIN
Doğum Yeri ve Tarihi	: Sivas 21.08.1999
Lisans	: Uludağ Üniversitesi Bilgisayar Mühendisliği