Biyoenformatikte Veri Analizi Uzerine Bir Uygulama

Proje Ekibi: Türkan Arıt

Danışman: Doç. Dr. Çağın Kandemir Çavaş

Amaç ve Kapsam

Projede biyolojik dizilerin birbirleriyle olan ilişkileri göz önünde bulundurularak sınıflandırılmaları amaçlanmıştır.
Bu kapsamda yapılan çalışmalar:

- Biyolojik dizilerin NCBI Genbank veri tabanından elde edilmesi.
- Literatürde yer alan sınıflandırıma algoritmalarının araştırılması ve kodlaması.
- Biyolojik dizilere ait Filogenetik Ağaçların elde edilmesidir.

Hedefler

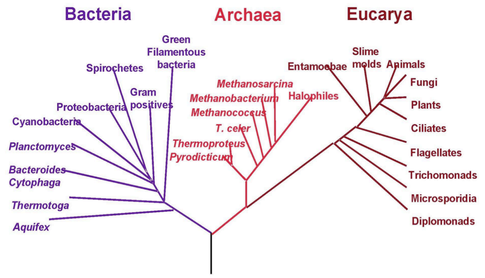
- Moleküler Filogenetik Analiz yöntemlerinin araştırılması ve değerlendirilmesi.
- Gerçek veri setleri üzerinde DNA dizileri arasındaki farklılıkları belirlemek için kullanılan istatistikler ve mesafe ölçütlerinin araştırılması.
- Oluşturulan veri setleri üzerinde etkin Filogenetik Ağaç çıktılarının elde edilmesi.

Veri Seti

Endüstriyel alanda geniş kullanım alanına sahip Lipaz enzimleri biyoteknolojik çalışmalarda sıkça kullanılır. Projede Aktif oldukları sıcaklık aralıklarına göre termofilik ve mezofilik özellik gösteren Lipaz enzimlerinin termal kararlılıklarına göre sınıflandırılması hedeflenmiştir.

Filogenetik Analiz

Filogenetik Analiz, çeşitli organizma grupları arasındaki evrimsel ilişkinin araştırılmasıdır. Moleküler Filogenetik çalışmaları DNA ve protein dizilerinde meydana gelen değişikliklerin hızını ve karakterini belirlemeye yöneliktir. Analizin çıktısı evrimsel dallanma sürecinde organizmaların birbirleriyle olan yakınlık derecelerini gösteren bir Filogenetik Ağaçtır.



Evrimsel Hayat Ağacı

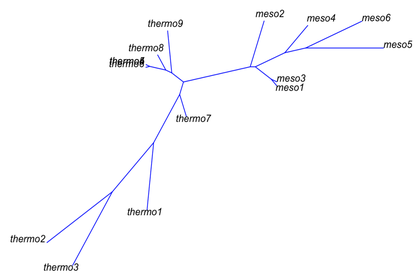
K-mer Natural Vector Yöntemi

Genomik diziler {A,T,G,C} kategorik değerlerini içeren bir liste gibi depolanır. Dolayısıyla kategorik diziler arasındaki benzerlik ilişkilerinin incelenebilmesi için istatistiksel analiz yöntemlerinden faydalanılır. Ardışık en bazın oluşturduğu alt diziler k-mer olarak adlandırılır. Yöntemde alt dizi dağılımlarını temsil eden normalize edilmiş merkezi moment vektörleri elde edilir. Kosinüs benzerliğinden yararlanılarak vektörler arasındaki uzaklıklar hesaplanır.

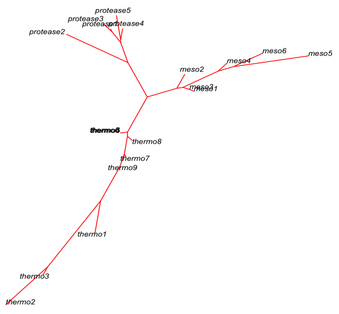
Uygulama

	meso1	meso2	meso3	meso4	meso5	meso6	termo1	termo2	termo3	termo4	termo5	termo6	termo7	termo8	termo9
meso1	0.000	0.158	0.030	0.197	0.331	0.345	0.447	0.742	0.696	0.281	0.281	0.291	0.315	0.301	0.391
meso2		0.000	0.167	0.249	0.389	0.388	0.475	0.668	0.698	0.311	0.311	0.324	0.350	0.331	0.415
meso3			0.000	0.188	0.321	0.343	0.442	0.729	0.687	0.279	0.279	0.285	0.296	0.299	0.389
meso4				0.000	0.275	0.268	0.529	0.780	0.771	0.377	0.377	0.376	0.407	0.393	0.467
meso5					0.000	0.302	0.666	0.745	0.787	0.544	0.544	0.553	0.546	0.573	0.596
meso6						0.000	0.598	0.819	0.789	0.466	0.466	0.484	0.490	0.490	0.534
termo1							0.000	0.391	0.498	0.398	0.398	0.405	0.324	0.410	0.384
termo2								0.000	0.347	0.566	0.566	0.568	0.439	0.559	0.392
termo3									0.000	0.530	0.530	0.533	0.441	0.515	0.450
termo4										0.000	0.000	0.017	0.168	0.080	0.146
termo5											0.000	0.017	0.168	0.080	0.146
termo6												0.000	0.167	0.078	0.146
termo7													0.000	0.138	0.169
termo8														0.000	0.128
termo9															0.000

Veri setinde k-mer Natural Vector Yöntemi ile elde edilen uzaklık matrisi.



Oluşturulan uzaklık matrisinde Neighbor Joining Kümeleme Algoritması kullanılarak elde edilen Filogenetik Ağaç.



Aynı bakterilerden alınan proteaz enzimine ait diziler eklenerek oluşturulan veri setinde k-mer Natural Vector yöntemi kullanılarak elde edilen Filogenetik Ağaç.

Sonuç

Çalışma sonucunda uzaklık tabanlı hizalama içermeyen Filogenetik Analiz yöntemleri uygulamalı olarak çalışılmıştır. Oluşturulan veri seti üzerinde herhangi bir ön işlem adımı uygulanmadan uzaklık tabanlı bir yöntem olan k-mer Natural Vector Yöntemi kullanılarak enzimler üzerinde fonksiyonel bir özelliğe göre başarılı bir kümeleme sonucu elde edilmiştir.

Öneriler

Çalışmanın ilerletilmesine yönelik, elde edilen ağaçların görsel incelemeyle karşılaştırılması yerine, üretilen ağaçların kalitesini değerlendirmek için bootstrap yöntemi kullanılması önerilebilir.

Kaynakça

- Amiri, S., Dinov, D. (2016) “Comparison of genomic data via statistical distribution," Journal of Theoretical Biology 318-327
- Choudri, Supratim. (2014) Bioinformatics for beginners: genes, genomes, molecular evolution, databases and analytical tools. Amsterdam: Elsevier
- Liu LW, Li DB, Bai FL (2012). A relative Lempel-Ziv complexity: Application to comparing biological sequences. Chem Phys Lett 530, 107-112
- Mount DM. (2004). Bioinformatics: Sequence and Genome Analysis (2 bas.). Cold Spring Harbor Laboratory Press. 0-87969-608-7.
- Royter M, Schmidt M, Elend C, Höbenreich H, Schäfer T, Bornscheuer UT, et al. Thermostable lipases from the extreme thermophilic anaerobic bacteria Thermoanaerobacter thermohydrosulfuricus SOL1 and Caldanaerobacter subterraneus subsp. tengcongensis. Extremophiles. 2009;13:769–83.