Doğrusal Olmayan Gömme Teknikleri Altında Gen Dizilerinin Evrimsel İlişkileri
Evolutionary Relationships Between Gene Sequences via Nonlinear Embedding

Tunca Doğan [1], Bilge Karaçalı [2]

1. Biyoteknoloji ve Biyomühendislik Bölümü,
İzmir Yüksek Teknoloji Enstitüsü
tuncadogan@iyte.edu.tr

2. Elektrik ve Elektronik Mühendisliği Bölümü,
İzmir Yüksek Teknoloji Enstitüsü
bilgekaracali@iyte.edu.tr

Özetçe

Bu çalışmada, çoklu dizi hizalamasını takiben çıkarımsanmış gen dizi ikilileri arasındaki evrimsel uzaklıklara uygulanan doğrusal olmayan gömme tekniğinin hata analizi sunulmuştur. Bu amaçla, sentetik gen dizileri oluşturulmuş ve üç ayrı evrimsel yol boyunca, her dizi ikilisi arasındaki gerçek evrimsel uzaklık kayıt altına alınaraktan rastlantısal yer değiştirme mutasyonuna tabi tutulmuşlardır. Düşük boyutlu bir vektör uzayına yapılan doğrusal olmayan gömme işlemi öncesi ve sonrası elde bulunan çıkarımsanmış evrimsel uzaklıklar ile gerçek evrimsel uzaklıklar arasındaki farklılıklar karşılaştırılmıştır. Sonuçlar, doğrusal olmayan gömme işleminin çıkarımsanan evrimsel uzaklıklarda yer alan hata payını düşürme konusundaki başarısını göstermiştir. Sonuç olarak, evrimsel uzaklıklar üzerine uygulanan doğrusal olmayan gömme işleminin, gen dizileri arasındaki evrimsel ilişkileri ortaya çıkarmak adına güvenilir çıkarımlarda bulunabileceği belirtilmiştir.

Abstract

We present an error analysis on the application of non-linear embedding on pairwise evolutionary distances inferred over a collection of genetic sequences following multiple sequence alignment. To this end, we have generated gene sequences evolved by random substitutions along three different evolutionary pathways with known evolutionary distances between every sequence pair. We have compared the discrepancy between the inferred evolutionary distances to the true distances before and after non-linear embedding into a low dimensional vector space. The results indicate that non-linear embedding achieves significant reduction in error in the estimated evolutionary distances. Consequently, non-linear embedding of evolutionary distances can provide more reliable inferences on the evolutionary relationships between genetic sequences.

978-1-4244-6382-4/10/$26.00 ©2010 IEEE

1. Giriş

Biyolojik organizmaların gen ve protein dizileri arasındaki evrimsel ilişkilerin keşfi, sistemlerini meydana getiren moleküler ve fizyolojik mekanizmaları ortaya çıkarmak açısından önem taşır. Bu çalışmalar eldeki dizilerin arasındaki benzerlikleri dikkate alarak, ortak ataların ortaya çıkarılması ve tüm bu dizilerin atalar-torunlar şeklinde birbirleri ile ilişkilendirildikleri soy ağaçlarının oluşturulmasıyla gerçekleştirilir. Bir diğer deyişle filogenetik yöntemler, bu gen ve protein dizileri arasındaki en muhtemel hizalamayı dikkate alarak moleküler yer değişimlerini saptamayı ve bu şekilde dizilerin ortak geçmişlerini ortaya çıkarmayı hedefler. Bu yöntemler, dizi uzaklıklarının hesaplanıp, bu uzaklıkları evrimsel uzaklıklarla ilişkilendiren matematiksel bir modelin uygulanması ve sonuçta evrimsel uzaklıkların çıkarsanmasına dayanır.

Pratikte, gen ve protein dizilerinin çok uzun olması ve özellikle geçmiş zamana ait dizilerin çoğunlukla elde bulunmaması, sistemin karmaşıklığını arttırır ve kavranmasını güçleştirir. Bu nedenden dolayı bu tip araştırmalar, istatistiksel anlamda gerçekleştirilir. Bu uygulamalar gerek doğruluk gerekse hız gibi özellikleri dolayısı ile işlemsel ortamda yürütülür. Konvansiyonel anlamda kullanılan bu yöntemlerden bazıları Jukes-Cantor modeli [1], Kimura'nın iki parametreli modeli [2] ve genel zaman tersinir model'dir [3]. Bu klasik yöntemler bilinmeyen evrimsel uzaklıklar hakkında bazı tahminlerde bulunsa da, bu tahminler bazen sonuçların güvenebilirliğini düşüren hatalara tabi olabilmektedir.

Bu tip çalışmalarda sağlıklı sonuçlar elde edilmesinin ve bu sonuçların kavranabilir görsel ifadelerle açıklanmasının karşısındaki en büyük engellerden birisi de çok boyutluluk sorunudur. Analize girdi olarak verilen bu dizilerin üzerinde kodlanmış her bir karakter, analize katılmış bir yeni boyutu ifade eder ve bu girdiler kimi zaman onbinlerce karakter uzunluğunda olduğunda (gen ve protein dizilerinde olduğu gibi) boyut sorunu çalışmadan sağlıklı sonuçlar almayı hayli güçleştirebilir. Bu sorunu aşmak için doğrusal ve doğrusal olmayan gömme teknikleri kullanılmaktadır. Bu teknikler birçok farklı bilim dalında, yapılan analizlerin sonuçlarını

makul boyutlarda ifade etmek için uygulanmıştır. Klasik çok boyutlu ölçeklendirme [4] ve esas bileşen analizi [5] bu tekniklerin ilk ve en sık kullanılan örnekleridir. Analiz edilecek veriler bir düzlem üzerinde dağılım gösterdiği zaman başarılı sonuçlar üretmesine rağmen, bu yöntemlerin eğri yüzeylerde başarısız olduğu literatürde belirtilmiştir [6]. Eğri yüzeylerdeki dağılımları yakalayabilmek için, bu klasik yöntemlerin uyarlanması ile birçok doğrusal olmayan gömme tekniği geliştirilmiş ve farklı alanlarda uygulanmıştır. Bu yöntemlerden bazıları, doğrusal olmayan haritalama [7], yerel doğrusal gömme [8] ve stokastik yakınlık gömmesi [9] yöntemleridir.

Doğrusal olmayan gömme yöntemlerinin protein ve gen dizileri arasındaki benzerlikleri ortaya çıkarmak adına, çoklu dizi hizalamasını takiben dizi verilerinden çıkarımsanan evrimsel uzaklıklara uygulanması daha önce çalışılmıştır. Örneğin bir çalışmada, stokastik yakınlık gömmesi tekniği kullanılarak, birbiri ile yakın ilişkiler içinde olan bazı protein dizileri anlamlı kümeler içinde gruplanmaya çalışılmıştır [10]. Farklı analiz parametrelerinin denenip bu parametrelerin optimize edilmeye çalışıldığı araştırmada, bazı protein ailelerinin fonksiyonel olarak birbirine yakın üyelerinin evrimsel anlamda uzak proteinlerden ayrılıp iki boyutlu uzayda gruplanması başarılmış, bu şekilde doğrusal olmayan gömme tekniklerinin protein ve gen dizilerini kümeleyebilme becerisi ortaya koyulmuştur [10].

İzometrik özellik haritalama (ISOMAP), doğrusal olmayan gömme yöntemlerinden biridir. Bu yöntem klasik ölçeklendirme içine gömülmüş olarak, komşuluk diyagramı üzerindeki jeodezik uzaklıkları dikkate alır [6]. Bu özellik, yönteme eğri yüzeyler üzerine dağılmış verileri bile başarılı şekilde yakalama becerisini kazandırır. Analize girdi olarak verilen çıkarımsanmış evrimsel uzaklıklar, komşu (yakın) noktalar arasındaki jeodezik uzaklıkları hesaplamak için iyi bir yaklaşım olarak yer alır [6]. Uzak noktalar içinse, bu komşu noktalar arasından birçok küçük atlama üst üste eklenerek jeodezik uzaklıklar hesaplanabilir.

Bu bildiride, sentetik gen dizilerinin, ISOMAP algoritması kullanılarak uygulanmış olan doğrusal olmayan gömme sonrası elde edilen evrimsel uzaklıklarının hata analizi sunulmuştur. Bu raporun sıradaki kısmında, farklı evrimsel yollar izleyen sentetik dizilerin üretilmesi ve hata analizinin gerçekleştirilmesi için kullanılan parametre açıklanmış, takip eden bölümde ise sonuçlar ortaya konulmuş ve bu sonuçların yorumlanması gerçekleştirilmiştir.

2. Yöntem

Çalışmada ilk önce, işlemsel ortamda hazırlanmış olan sentetik gen dizilerinin aralarındaki evrimsel uzaklıklar, konvansiyonel Jukes-Cantor modeli [1] kullanılarak çıkarımsanmıştır. Daha sonra bu veriler üzerinde ISOMAP yöntemi uygulanarak, doğrusal olmayan gömme sonrası evrimsel uzaklık değerleri elde edilmiştir. Bu iki grup evrimsel uzaklık değeri, sentetik diziler arasındaki gerçek evrimsel mesafelerle karşılaştırılarak hata analizi ortaya koyulmuştur. Hata miktarları grafikler üzerinde karşılaştırılarak, ISOMAP yöntemi ile yapılan doğrusal olmayan gömme işleminin, çıkarımsanmış evrimsel uzaklıklar içinde yer alan hata miktarlarını anlamlı olarak düşürmeyi başarıp başaramadığı diğer bir deyişle, hata düşürme kapasitesi incelenmiştir. Sentetik gen dizilerinin oluşturulması dahil tüm işlemler ve analizler MATLAB® yazılımı kullanılarak yapılmıştır.

2.1. Sentetik gen dizilerinin oluşturulması

Sentetik gen dizilerinin hazırlanmasında üç ayrı evrimsel yol izlenmiş, analizler de bu üç yol için ayrı ayrı gerçekleştirilmiştir. Evrimsel yollar, orijinal sentetik gen dizisinin bazı bölgelerinde rastlantısal bazal yer değiştirmeler yerleştirilmesi ile yeni diziler oluşturulmasına dayanır. Bu operasyon işlemsel ortamda simule edilmiştir.

Her bir dizi kendisinden bir önce gelen dizinin (önceki neslin) en yakın akrabasıdır. Bu şekilde, nesiller ilerledikçe oluşan yeni diziler, orijinal (ilk) diziden gittikçe farklılaşırlar. Her bir evrimsel yolda 501'er (orijinal diziyle beraber) dizi bulunur. Her bir dizi 1000 nükleotidden oluşur. Analizde kullanılan yer değiştirme hızı %2'dir. Diğer bir deyişle, var olan bir diziden yeni bir dizi oluşturulurken, dizilerde yer alan ve her birini bir nükleotidin kapladığı 1000 bölgeden her biri için, eski dizideki nükleotidin aynen yeni diziye yerleştirilme ihtimali %98, rastgele olarak 4 farklı nükleotidden birinin yerleştirilme ihtimali ise %2'dir. Evrimsel yol 1 (EY1), Evrimsel yol 2 (EY2) ve Evrimsel yol 3'ün (EY3) oluşturulmasında izlenen yollar Şekil 1'de verilmiştir.

Hata değerleri EY1, EY2 ve EY3 için ayrı ayrı elde edilmiştir. Bununla beraber hata analizi, her bir evrimsel yol için, 500 dizi içinde yer alan tüm dizi ikilileri için toplu şekilde değil, aralarında belirli nesil sayıları bulunan diziler için ayrı ayrı ortaya konmuştur (10, 20, 50, 100 ve 200 nesil farklılıklar için). Bunun nedeni, aradaki nesil sayısının değişimiyle beraber, hata oranının da şiddetli şekilde değişmesidir.

2.2. Gen dizileri arasındaki çıkarımsanmış evrimsel uzaklıkların hesaplanması

Sentetik gen dizileri arasında yer alan evrimsel uzaklıkların çıkarımsanmasında kullanılan konvansiyonel Jukes-Cantor modeli, stokastik bir yaklaşımdır. Bu ve benzeri konvansiyonel modeller, iki dizi arasındaki dizi uzaklığını bazı parametreler ve basit bir formulasyon yardımı ile yine bu iki dizi arasındaki evrimsel uzaklığa çevirirler. Jukes-Cantor modeli nükleotid yer değiştirme hızını 4 çeşit nükleotidin tüm farklı ikili kombinasyonları için eşit kabul eder. Bunun yanında nükleotid frekansları ve dizi üzerindeki bölgeler arası yer değiştirme hızları da eşit kabul edilmiştir. Jukes-Cantor modelinin fomulasyonu 1 numaralı denklemde verilmiştir.

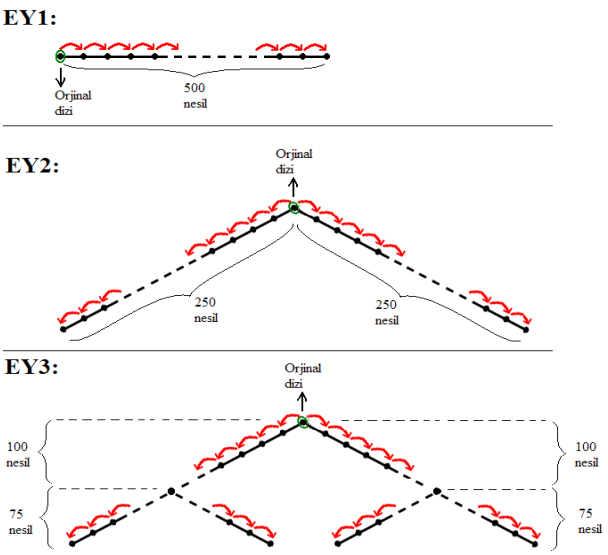
$$d = -3/4 \log[1 - (4/3)D] \quad (1)$$

Denklemde yer alan "d" evrimsel uzaklığı, "D" ise dizi uzaklığını temsil eder. Çoklu dizi hizalaması sonucu ortaya çıkan diziler üzerinde, her bir bölge için ayrı ayrı yapılan ikili karşılaştırmalar sonrasında, ikili dizi uzaklıkları ortaya çıkar. Bu hesaplamada herhangi bir bölgede, iki dizide de aynı nükleotid bulunuyorsa bir eşleşme, farklı nükleotidler varsa eşleşmeme kayıt edilir. İşlem sonunda toplam eşleşmeme sayısı iki dizide de boşluk olmayan toplam bölge sayısına oranlanır. Denklemin derivasyonu referans bölümünde verilen 1 numaralı kaynaktan incelenebilir.

2.3. Gen dizileri arasındaki çıkarımsanmış evrimsel uzaklıkların ISOMAP yöntemi ile işlenmesi

ISOMAP yöntemi kendisine girdi olarak verilen bir uzaklık matrisini alıp, matriste aralarındaki mesafeler verilen noktaları çok boyutlu vektör uzayında sabit noktalara atar. Bu şekilde, nümerik uzaklık matrislerini de, geometrik olarak temsil etmiş olur. Yöntemden alınan çıktıda verinin gömüleceği vektör uzayı boyut sayısı isteğe göre seçilir. Genelde, işlem birçok farklı boyut sayısı için ardı sıra tekrarlanır ve işlem sonunda araştırmacının bu boyut sayılarından birini seçip doğrusal olmayan gömmenin bu boyuttaki bir vektör uzayına yapılması sağlanır. Boyut sayısının sağlıklı olarak seçilebilmesi için boyut sayısına karşılık artık varyans değerlerinin çizildiği bir grafik verilir (artık varyans grafiği). Bu grafikte, artan boyut sayısı ile beraber düşen artık varyans değerinin incelenmesi, varyanstaki düşüşün sona erip minimuma yaklaştığı en düşük boyut sayısının seçimini sağlar. Mümkün olduğu kadar düşük boyuttaki ifadenin seçilmesinin nedeni, veri setinin içindeki yapıyı bozmadan, sistemi maksimum derecede basitleştirmek, hatta imkan varsa görsel geometrik bir temsili sağlayabilecek olan 1, 2 veya 3 boyutta çalışmaktadır.

Bir önceki adımda, Jukes-Cantor modeli ile çıkarımsanmış olan evrimsel uzaklıklar, ISOMAP yöntemine girdi olarak verilerek doğrusal olmayan gömme tekniği ile çok boyutlu bir vektör uzayında ifade edilmeleri sağlanır. ISOMAP yöntemi, her gen dizisini, diziler arası ikili evrimsel uzaklıkları dikkate alarak, çok boyutlu uzayda yer alan, koordinatları belirli bir nokta ile ifade eder. Her bir gen dizisine tekabül eden bu noktalar arasındaki metrik mesafeler, yöntemin sonuç olarak verdiği evrimsel uzaklıklar olarak kaydedilir.



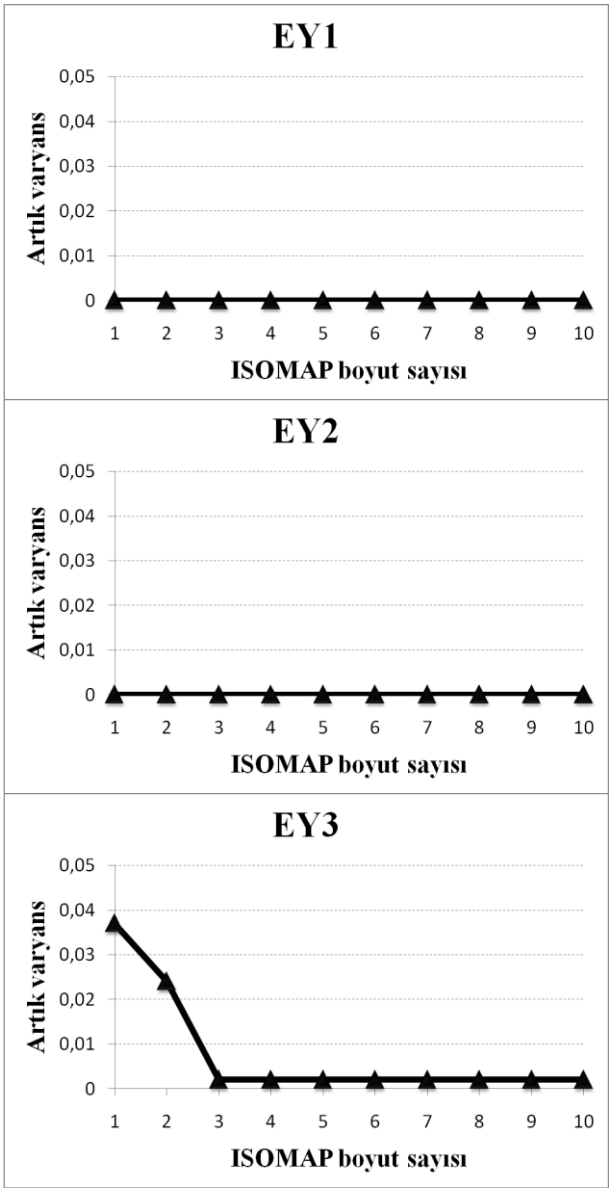
Şekil 1: EY1, 2 ve 3'ün oluşturulmasında izlenen yollar.

3. Sonuçlar ve tartışma

Hata düşürme kapasitesinin ortaya konması için kullanılan parametre, ortalama mutlak hata değerleridir. Bu parametre, bir yöntemden çıktı olarak alınan bir dizi ikilisi arasındaki çıkarımsanmış evrimsel uzaklık değeri ile bu iki dizi arasındaki gerçek evrimsel uzaklığın birbirlerinden çıkarılması ile ortaya çıkan farkın mutlak değer içine alınması, aynı işlemin tüm dizi ikilileri için tekrarlanması ve sonunda elde edilen tüm değerlerin ortalamasının alınmasıyla elde edilir. Sonuçta biri konvansiyonel yönteme ve diğeri de ISOMAP'e ait olmak üzere iki hata değeri oluşur.

ISOMAP işlemi sonunda, doğrusal olmayan gömme işleminin yapılacağı vektör uzayının boyut sayısını seçmek için artık varyans grafikleri (Şekil 2) incelenmiştir. Artık varyanstaki düşüşün sona erdiği boyutlar olan, EY1 ve EY2 için 1 boyutlu, EY3 içinse 3 boyutlu vektör uzaylar seçilmiştir.

Sonuçların incelenmesinin ve birbirleri ile karşılaştırılmasının görsel anlamda ve kolay anlaşılabilir olması için grafik gösterimler hazırlanmış ve bu bölümün devamında sunulmuştur. Bu kısımda verilen her bir şekil, ayrı bir evrimsel yol (EY1, EY2 ve EY3) için yapılmış analizi temsil eder. Her şekilde, x-ekseninde nesil uzaklıkları (10, 20, 50, 100 ve 200), y ekseninde ise ilgili nesil uzaklıklarına ilişkin ortaya çıkan hata değeri (konvansiyonel Jukes-Cantor modeli ve ISOMAP yöntemi için iki ayrı bar) verilmiştir.

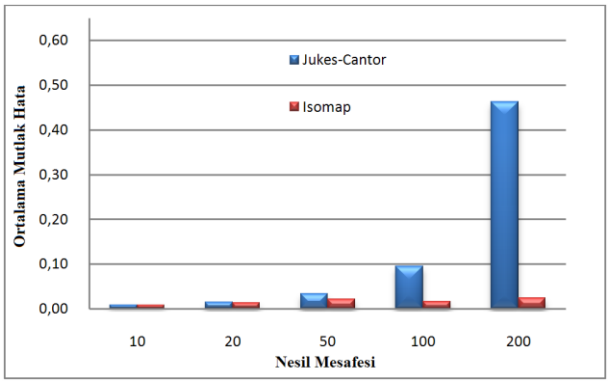


Şekil 2: EY1, 2 ve 3 için artık varyans grafiği.

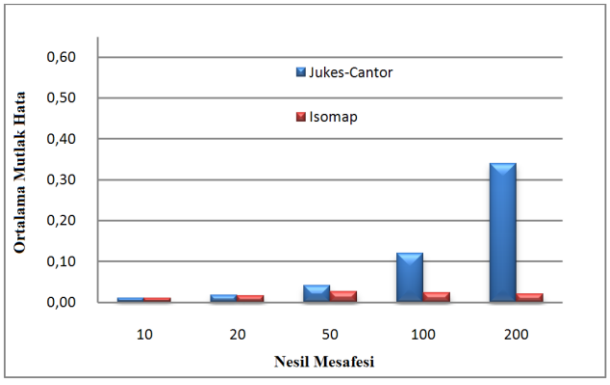
Şekil 3, 4 ve 5'de görüldüğü üzere, tüm analize tabi tutulmuş evrimsel yollar için, 50 nesil farka kadar (0 – 50 arası), konvansiyonel model ve ISOMAP çıktıları birbirlerine çok yakın hata değerleri ortaya koymuştur. Sonuç olarak, ISOMAP yöntemi 50 nesile kadar olan mesafelerde, konvansiyonel yöntemin çıkarımsadığı evrimsel uzaklıklarda yer alan hatayı belirgin olarak düşürmemiştir. Bunun nedeni ise, bu analiz şartlarında, 50 nesile kadar olan mesafelerde konvansiyonel modelin zaten oldukça düşük ve kabul edilebilir ölçüde hatalar vermiş olmasıdır.

Yine aynı şekillerde görülmektedir ki, 100 ve 200 nesil mesafelerde konvansiyonel modelin ürettiği hata bir anda yükselmektedir. Bunun yanında ISOMAP sonunda çıkan değerler içerisinde kalan hata, çok düşük mesafelerde gözlenen hata ile yakın değerlere sahiptir. Bu da, ISOMAP için 100 ve 200 nesil (sırasıyla 0,65 ve 0,75 değerlerinde dizi uzaklıklarına ve 1,5 ve 3,0 değerlerinde çıkarımsanmış evrimsel uzaklıklara karşılık gelir) mesafelerde ortaya çıkan çok belirgin bir hata indirgeme kapasitesini işaret etmektedir.

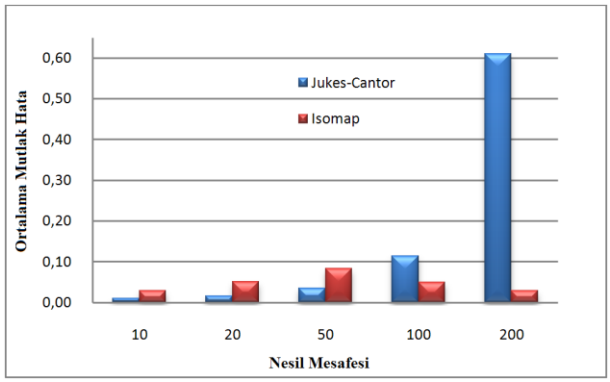
ISOMAP yönteminin yüksek nesil mesafelerindeki hataları indirgemede ki bu başarısı, algoritmanın birbirinden çok uzakta yer alan iki nokta arasındaki metrik mesafeyi hesaplarken, girdi olarak kendisine verilen değeri dikkate almak yerine, birbirine çok yakın mesafedeki noktalar arasında küçük atlamalar yaparak bu iki uzak noktanın birinden diğerine ulaşmasına ve oluşan bu mesafeyi dikkate almasına dayanır.



Şekil 3: EY1 için, konvansiyonel model (Jukes-Cantor) ile ISOMAP arasında ortalama mutlak hata karşılaştırması.



Şekil 4: EY2 için, konvansiyonel model (Jukes-Cantor) ile ISOMAP arasında ortalama mutlak hata karşılaştırması.



Şekil 5: EY3 için, konvansiyonel model (Jukes-Cantor) ile ISOMAP arasında ortalama mutlak hata karşılaştırması.

Bir başka deyişle kısa mesafeler arasındaki uzaklığı hesaplarken ortaya çıkan hata değeri küçük, uzun mesafeleri hesaplarken ortaya çıkan hata değeri büyüktür, ISOMAP uzun mesafeler arasındaki uzaklıkları hesaplarken, kısa ve güvenilir mesafeleri birbirine ekleyerek bir uzaklık belirlediği için hatayı küçük değerlerde tutmayı başarmaktadır.

Elde edilen bu sonuçlar, doğrusal olmayan gömme yöntemi ISOMAP'in gen ve protein dizilerinin evrimsel sınıflaması konusunda, evrimsel mesafelerin tahmininindeki hataları belirgin şekilde düşürme kapasitesine sahip olduğunu ortaya koymuştur. Algoritmanın ve parametrelerin en uygun şekilde seçimi ile yöntemin diziler arası evrimsel ilişkileri ortaya çıkarma konusunda verimli olarak kullanılma potansiyeline sahip olduğunu düşünülmektedir.

4. Kaynakça

[1] T. H. Jukes, C. R. Cantor, "Evolution of protein molecules". Pp. 21-123 in H. N. Munro, ed. *Mammalian protein metabolism.* Academic Press, New York, 1969.

[2] M. A. Kimura, "Simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences". *Journal of Molecular Evolution* 16:111-120, 1980.

[3] S. Tavaré, "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences". *American Mathematical Society: Lectures on Mathematics in the Life Sciences* 17: 57–86.

[4] I. Borg, P. Groenen, "Modern Multidimensional Scaling: theory and applications", *Springer New York,* 2005.

[5] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space". *Philosophical Magazine 2* (6): 559–572, 1901.

[6] J. B. Tenenbaum, V. de Silva, J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction". *Science, 290,* 2319–2323, 2000.

[7] J. W. Sammon, "A nonlinear mapping for data structure analysis". *IEEE Trans. Computers* 1969, 18, 401-409.

[8] S. Roweis, L. Saul, "Nonlinear dimensionality reduction by LLE". *Science*, 290, 2323-2326, 2000.

[9] D. K. Agrafiotis, H. Xu, "A self-organizing principle for learning nonlinear manifolds". Proc. *Natl. Acad. Sci. U.S.A.*, 99, 15869- 15872, 2002.

[10] M. A. Farnum, H. Xu, D. K. Agrafiotis, "Exploring the nonlinear geometry of protein homology". *Protein Science,* 12:1604–1612, 2003.