

Genetik Algoritmaya Çoklu Dizi Hizalama

Konu

1.15 Yapay Zeka ve Sinir Ağları
1.14 Biyomedikal Sinyal İşleme

Ömer Sinan Saraç¹, Özgür Alan¹, Erkin Bahçeci¹, Kemal Leblebicioğlu²
Orta Doğu Teknik Üniversitesi

¹Bilgisayar Mühendisliği Bölümü

²Elektrik ve Elektronik Mühendisliği Bölümü

06531, Ankara, Türkiye

sarac@ceng.metu.edu.tr, alan@ceng.metu.edu.tr, erkinb@ceng.metu.edu.tr, kleb@metu.edu.tr

Haberleşme için:

Ömer Sinan Saraç
ODTÜ Bilgisayar Müh.
Araş. Görv.
ODTÜ-Ankara
Tel.:0312 210 5533
Fax.:0312 210 1259
sarac@ceng.metu.edu.tr

Özetçe

Çoklu dizi hizalama (ÇDH) başta bioinformatik olmak üzere bir çok alanda önemli uygulama alanları olan bir hesaplama problemidir. ÇDH hizalanacak dizi sayısı ile üssel tanımlı hesaplama karmaşıklığına sahiptir. Bu da günümüz hesaplama gücüyle ve pratikte geçerli büyük boyutlu verilerle hesaplamayı imkansız kılar. Yaklaşık çözümler bulan yöntemler ise yerel optimuma takılma ya da veriye özel doğru parametreleri ayarlama zorlukları yaşamaktadır. Genetik algoritma yöntemleri son dönemlerde oldukça yaygınlaşmış optimizasyon yöntemleridir. Bu çalışmada genetik algoritma metodları kullanılarak ÇDH probleminin çözülmesi ve şimdiye kadar yapılmış olan çalışmalarda güçlüklere aşılanmasını sağlayacak ve uzmanlara üzerinde yorum yapılmak üzere alternatif iyi hizalamalar sunacak bir yöntemin geliştirilmesi amaçlanmıştır.

Çoklu Dizi Hizalama

Çoklu dizi hizalama (ÇDH), biyoinformatikte temel bir araçtır. ÇDH, dizi oluşturma, moleküler modelleme, veritabanı aramaları, filogenetik ağaç oluşturulması gibi konularda başlıca araçtır. İkili dizi hizalamadan (İDH) çok daha fazla bilgi verir. İDH'da dinamik programlama tekniğiyle polinom zamanda optimum sonuç bulunabilmesine rağmen, var olan ÇDH teknikleri optimum hizalamayı ancak dizi sayısı ile üssel ilişkili bir sürede bulabilir. Bu da pratikte gerekli olan büyük boyutlu verilerde problemin çözümünü, günümüz hesaplama gücüyle imkansız kılar. Ayrıca bazı buluşsal (heuristic) yöntemlerle ortaya çıkan sonuçların yerel optimuma takılma ihtimalleri çok yüksektir. Buna ek olarak, bulunan hizalamalar uzmanların bakış açılarına göre de farklı değerler alabilmektedirler. Bu yüzden yerel optimaya takılmadan farklı iyi hizalamalar verebilecek yöntemler gittikçe önem kazanmaktadır.

İlgili Çalışmalar

ÇDH'da en çok kullanılan yöntem aşamalı hizalama yöntemidir. Bu yöntemi kullanan araçların en yaygını CLUSTALW'dir [1]. Önce eldeki dizilerin ikili benzerlikleri kullanılarak tahmini

bir filogenetik ağaç oluşturulur. Daha sonra bu bilgiden yararlanılarak, aşamalı olarak diziler hizalanır. Bu yöntemin en önemli dezavantajları, yerel minimuma takılması ve problemi doğru şekilde modelleyecek parametrelerin belirlenmesinin zor olmasıdır.

Bunun yanısıra olasılıklı (stochastic) yöntemler kullanan yaklaşımlar da olmuştur. Çok yavaş yöntemlerdir ve daha çok çözüme yakın noktalardan başlayarak iyileştirme yapmak amacıyla kullanılırlar.

ÇDH problemlerini çözmek için kullanılan diğer bir ana yaklaşımsa son dönemlerde oldukça yaygın olarak kullanılan genetik algoritmalarıdır. En çok bilinen örneklerden biri SAGA (Sequence Alignment by Genetic Algorithm)'dır [2]. ClustalW'ya yakın hizalamalar bulmuştur ama daha başarılı değildir.

Yöntem

Temel olarak genetik algoritma kullanılmıştır. Genetik algoritmalarda problem çözümleri kromozom olarak kodlanır. Bu kodlama genetik algoritmanın etkin çalışabilmesi için büyük önem taşır. Bu çalışmada kullanılan kodlamada hizalama için gerekecek boşlukların pozisyonu tutulmaktadır [3]. Buradaki pozisyon boşluğun, verilen orjinal dizide kaçınıcı bazdan sonra geldiğini belirtir. Bir bazdan sonra birden çok boşluk bulunabilir. Bu durumdaki boşluklar yani grup boşluklar aynı sayı ile ifade edilir. Bütün dizilerin hizalaması, yani problemin olası bir çözümü, tek bir kromozomda kodlanmıştır. Aşağıda üç diziden oluşan bir hizalama ve ona karşılık gelen kodlama (kromozom) gösterilmiştir:

```
AAT---T-GCCTCG-GCAATC
-ATG--T--CCT---GC---C
AATGTTT-----CGGCCAATC
```

3	11	4	3	3	3	0	3	4	4	7	9	9	7	7	9	7	7	7	7	7
---	----	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Bu gösterimin tercih edilmesinin sebebi çaprazlama ve mutasyonlar sonucu oluşan bireylerin her zaman geçerli bir hizalama belirtmesidir. Bu kodlamaya özel çaprazlama ve mutasyon işlemleri tanımlanmıştır. Dört çeşit çaprazlama tekniği kullanılmıştır.

1. **Dizi bazlı tek noktadan çaprazlama:** İki birey arasında, rastgele seçilen ve belli bir dizinin kodlamasının başlangıcına karşılık gelen bir noktadan itibaren boşluk pozisyonlarının yer değiştirmesi.
2. **Dizi bazlı eş dağılımlı çaprazlama:** İki birey arasında, belli bir olasılıkla her bir dizinin kodlamasının yer değiştirmesi.
3. **Tek noktadan çaprazlama:** İki birey arasında, kromozonun herhangi bir noktasından itibaren boşluk pozisyonlarının yer değiştirmesi.
4. **Çok noktadan çaprazlama:** İki birey arasında, birden çok noktanın seçilip oluşan aralıkların her ikisinden birinin (bir atlayarak) yer değiştirmesi.

Çeşitli mutasyon işlemleri tanımlanmıştır. Bunlar arasında, boşlukların bir araya getirilmesi, boşlukların tek başlarına ya da grup olarak sona taşınmaları, bir pozisyon sağa ya da sola kaydırılmaları, ya da yerlerinin tümünden değiştirilmesi sayılabilir. Belirtilmesi gereken önemli bir nokta da, sezgisel (heuristic) yöntemler içeren mutasyonlar yapıldığıdır. Hizalamada daha düşük skorlara sebep olan dizi kodlamaları daha fazla mutasyona uğrama olasılığına sahiptir.

Kromozomların uygunluk (fitness) değerleri benzer çalışmalardan farklı olarak bütün dizilerin hizalamasına bakılarak hesaplanır. L dizi uzunluğu, N dizi sayısını gösterirse bu hesaplamanın maliyeti $L*N$ dir. Benzeri çalışmalarda ikili hizalamalar sonucu oluşturulan uygunluk değerleri (SP-score) kullanılmıştır. Bunun da maliyeti $L*N^2$ dir.

Sonuçlar

Bu çalışmanın asıl amacı hızlı sonuçlar üretmektense, aşamalı hizalama yöntemlerinden farklı ve daha iyi hizalamalar bulabilecek bir yöntem geliştirmektir. Aşamalı hizalama yöntemleri yerel minimumlara takılabilir ve farklı hizalama alternatifleri sunmaz. Bu çalışmada bu güçlükleri aşan yani, yerel minimumdan kurtulabilen ve birden fazla alternatif hizalama sunabilen bir yöntem geliştirilmiştir. Örnek bir program çıktısı, ClustalW çıktısı ile karşılaştırmalı olarak **tablo 1**'de verilmiştir.

Baz dizileri	S1> ATTTGTGGCCTGCATTTGTGGCCTGCA S2> ATGTGTGCCCTGCAATGTGTGCCCTGCA S3> TTTGTGGCCTGGCCTGCA S4> GTTTGTGGCCTGCAGTTTGTGGACTGCA S5> ATTTGTGGCTGCAATTTGTGGCTGCA S6> ATCTGTGGCCTGCAATCTGTGGCCTGCA
Program çıktısı	S1 ATTTGTGGCCTGCA-TT-TGTG-GCC-TGCA S2 ATGTGTGCCCTGCAAT---GTGTGCCCTGCA S3 -TTTGTG-----GCC-T-----G-GCC-TGCA S4 GTTTGTGGCCTGCAGTT-TGTG-GAC-TGCA S5 ATTTGTGGC-TGCAAT-TTGTG-G-C-TGCA S6 ATCTGTGGCCTGCAATC-TGTG-GCC-TGCA
ClustalW çıktısı	S1 ATTTGTGGCCTGCA-TTTGTGGCCTGCA S2 ATGTGTGCCCTGCAATGTGTGCCCTGCA S3 -TTTGTGGCCTG-----GCCTGCA S4 GTTTGTGGCCTGCAGTTTGTGGACTGCA S5 ATTTGTGGC-TGCAATTTGTGGC-TGCA S6 ATCTGTGGCCTGCAATCTGTGGCCTGCA

Tablo 1. Karşılaştırmalı sonuçlar

Kaynakça

1. Thompson, J.D, Higgins, D.G, and Gibson, T.J., "CLUSTALW: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position Specific Gap Penalties and Weight Matrix Choice". *Nucleic Acid Research*, 22:4673-4680, 1994.
2. Notredame, C. and Higgins, D.G., "Saga: Sequence Alignment by Genetic Algorithm." *Nucleic Acids Research*, Vol. 24, No. 8, pp. 1515-1524, 1996.
3. Karadimitriou, K. and Kraft, D. H., "Genetic Algorithms and the Multiple Sequence Alignment problem in Biology", *Proceedings of Annual Molecular Biology and Biotechnology Conference*, Baton Rouge, LA, Feb. 1996.