

Çoklu DNA Hizalama Yazılımı DNmA^{lign}

Cankut ÇUBUK^a, Alirıza ARIBAŞ^a, Oğuz DİCLE^b

^a Medikal İnfomatik AD, Dokuz Eylül Üniversitesi, İzmir

^b Radyodiagnostik AD, Dokuz Eylül Üniversitesi, İzmir

Multiple DNA Alignment Tool - DNmA^{lign}

Abstract: As necessity of high throughput systems(microarray and next-generation sequencing) develop in genetics, bioinformatics is becoming a hot topic. Comparison of the sequencing data and making sequence-base designs have a high importance for genetic diagnosis and identification in lifesciences & medical genetics. In international terminology of bioinformatics, the software which are used for obtaining homologous and non-homologous regions in DNA sequences are named “Sequence Alignment Tool”. There are lots of sequence analysis software. Open source ones of these software are processing on web-based servers. Uploading, processing data and visualisation of results with these software are time consuming due to server processing load. It may result as failure because of server load.

There are many parameters that determine the popularity of the Sequence Alignment Tools such as user friendly interface, interface design, accurate alignment, system requirements, data processing time and capacity. In this study dataset of “Meleagris Gallopavo” and “Gallus Gallus” were obtained from National Center for Biotechnology Information(<http://www.ncbi.nlm.nih.gov>) used for comparison of sequences with different alignment software. DNmA^{lign} software developed in this study was compared with server based, open source and cost free T-Coffee, ClustalW, Kalign and MUSCLE software. The comparison was done according to the parameters stated above and alignment results. Each parameter scored numerically between 0-10 and software's parametric scores were added to obtain a total score. Although each software uses different algorithm, comparison of sequence alignment results showed %90 similarity. For analysis of data over 2kb, DNmA^{lign} could process better with %100 efficiency whereas the server based ones failed with %30. Server based software provide service to multiple user at the same time and require high speed internet connection. DNmA^{lign} software requires minimum 256mb RAM, Microsoft Windows operating system and does not require internet connection. All the server based software which were compared can construct phylogenetic tree for aligned sequences. DNmA^{lign} differs from others on this step. Instead of constructing phylogenetic trees, DNmA^{lign} permits designing of PCR primers with determining thermodynamic features, reverse, complement, reverse-complement and duplex formations of the selected primer sequences. This way provides the user advantage of designing primers in the same window.

DNmA^{lign} is a server and internet connection independent software which could run locally on a computer by user inputs. In bioinformatics, number of national software are almost nonexistent in spite of huge necessity. Here we started with alignment project but our goal is to develop software with wider subjects and larger groups.

Key Words: Multiple DNA sequence alignment, Open Source, Client Application

Özet: Genetik biliminde yüksek çıktı veren, mikrodizin ve yeni nesil sekanslama gibi gelişen sistemlerin gereksinimi olarak biyoinformatik sıklıkla gündeme gelmektedir. Sekanslama çalışmalarının sonucunda elde edilen verilerin birbiri ile karşılaştırılması, sekans temelli tasarımlar, genetik tanı ve tiplendirme çalışmaları ile yakından ilgili olan tıbbi genetik, ziraat, veterinerlik ve moleküler biyoloji bilim dallarında önemli bir yere sahiptir. DNA sekans verilerindeki benzer olan ve olmayan bölgelerin bulunmasında kullanılan sekans hizalama yazılımları uluslararası terminolojide “Sequence Alignment Tool” olarak yer almaktadır. Günümüzde birçok sekans analizleme yazılımı bulunmaktadır. Bu yazılımlardan sekans hizalama yapanların açık kaynak kodlu olanları web tabanlı olarak sunucu üzerinde çalışmaktadır. Sunucuya veri gönderilmesi,

verinin işlenmesi ve sonucun görüntülenmesi internet bağlantı hızı ve sunucudaki işlem yoğunluğuna bağlı olarak zaman alıcıdır. Sunucunun cevap verememesinden dolayı bu işlemler başarısızlıkla sonuçlanabilmektedir. Bu çalışmamızda, sunucu ve internet bağımsız, bilgisayarda yerel olarak çalışabilen, kullanıcı tarafından girilen sekansları hizalayarak, elde edilen sonuçların tersi, tamamlayıcısı ve ters tamamlayıcısını elde edebilen bir yazılım geliştirdik.

DNA sekans hizalama programlarının popüleritesini belirleyen birden fazla parametre vardır. Kullanıcı arayüzü kolaylığı ve tasarımı, doğru hizalama, veri işleme kapasitesi, sistem gereksinimleri ve işlem süresi bu parametrelerin başında gelmektedir. DNMA^{lign} yazılımımızı National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>) veritabanından elde ettiğimiz “Meleagris Gallopavo” ve “Gallus Gallus ” sekanslarından oluşan veri setini kullanarak sunucu temelli çalışan, açık kaynak kodlu ve ücretsiz olan; T-Coffee, ClustalW, Kalign ve MUSCLE programlarıyla karşılaştırdık. Yaptığımız karşılaştırmalarda her yazılım sekans hizalama sonuçları ve yukarıda belirtilen parametreler bakımından 0-10 arası skorlanarak değerlendirildi. Her yazılım farklı bir algoritma kullanmasına karşın elde edilen sekans hizalama sonuçları %90 oranında benzerlik göstermiştir. 2kb üzeri veri setinin çözülmesinde DNMA^{lign} yazılımı sorunsuz çalışırken, sunucu temelli yazılımlar veriyi işlemede %30 oranında başarısız olmuştur. Sunucu temelli yazılımlar birden fazla kullanıcıya aynı anda hizmet vermektedir ve yüksek özellikte internet bağlantısına ihtiyaç duymaktadır. DNMA^{lign} yazılımı minimum 256 MB RAM donanımına sahip Microsoft Windows işletim sistemli bilgisayarlarda sorunsuz çalışabilmektedir ve internet bağlantısına gereksinim duymamaktadır. Karşılaştırmaya alınan sunucu tabanlı programlar hizalama yapılan sekanslarla filogenetik ağaç inşa edebilmektedir. DNMA^{lign} yazılımı karşılaştırılan yazılımlardan bu basamakta farklılık göstererek filogenetik ağaç inşa etmemektedir ancak hizalama yapılan sekanslar üzerinde genetik çalışmaların temel unsuru olan sentetik oligonükleotid dizilerinin(primer) belirlenmesine kullanılabilir. DNMA^{lign} belirlenen primer sekanslarının termodinamik özelliklerini, ters, tamamlayıcı, ters tamamlayıcı ve dubleks yapılarını analizleyerek primer tasarımı için kullanılan birden fazla basamağın aynı ekranda yapılmasına olanak vermektedir.

Biyoinformatik alanında ulusal yazılımlar yok denilecek kadar az fakat bu yazılımlara duyulan ihtiyaç çok fazladır. Sekans hizalama konusuyla başlamış olduğumuz bu projemizi daha geniş kapsamlı ve daha geniş gruplarla devam ettirmeyi amaçlamaktayız.

Anahtar Kelimeler: Çoklu DNA sekansı hizalama, Açık Kaynak Kodlu yazılımlar, Kullanıcı Uygulaması

Sorumlu Yazarın Adresi

Cankut ÇUBUK Dokuz Eylül Üniversitesi Tıp Fakültesi Sağlık Bilimleri Enstitüsü Medikal İnformatik AD
35340 İnciraltı-İzmir
E-Posta: cankutcubuk@gmail.com