BİYOLOJİK VERİ TABANLARINA GİRİŞ

Bu bölümde, Biyoenformatiğin, Moleküler Biyolojik araştırma alanlarında sıklıkla kullanılan uygulama alanlarından biri olan veri tabanlarını ve bu veri tabanlarında nasıl arama yapılması gerektiği ile ilgili çeşitli ipuçlarını açıklayacağız. Ancak bunlardan önce Biyoenformatik kavramından kısaca bahsetmemiz gerekmektedir.

Biyoenformatik, Bilgisayar mühendisliği, İstatistik ve uygulamalı Matematik alanlarındaki yaklaşımların ve yöntemlerin, biyolojik veri analizi için uygulanması olarak tanımlanabilir. Gelişen teknoloji ve Moleküler Biyoloji alanındaki ilerleyişle birlikte, araştırmacıların elinde büyük boyutlarda, deneysel veri birikmiştir ve artarak birikmeye devam etmektedir. Deneysel verinin hızlanarak çoğalması ile bu verilerin analizi, anlamlandırılması, ileri araştırmalar ve uygulamalar için hipotezler geliştirilmesi, Biyoenformatik disiplininin çalışma alanlarındandır. Daha önce bahsettiğimiz gibi, Biyoenformatik, yeni gelişen disiplinler-arası bir araştırma alanıdır. Biyoenformatik alanının, temel amacı, Moleküler Biyolojik verilerin idaresi ve analizine yönelik, veritabanlarının oluşturulmasını, algoritmaların geliştirilmesini, hesaplamalı ve istatistiki yöntemlerin ve yaklaşımların oluşturulmasını kapsamaktadır.

Gelişen teknoloji ile beraber, moleküler biyoloji ve genetik alanında çalışan her biyologun bir dereceye kadar biyoenformatiksel yaklaşımları bilmesi gerekmektedir.

İnternetten kolayca erişilebilen, sıklıkla kullanılan, 3 tane genom veri tabanı vardır. Bunların haricinde, kullanılmakta olan ve farklı amaçlara hizmet eden birçok veri tabanına da ulaşmak mümkündür.

- **The National Center for Biotechnology Information**

  http://www.ncbi.nlm.nih.gov/

- **UCSC Genome Bioinformatics Site**

  http://genome.ucsc.edu/

- **The Ensembl Project**

  http://www.ensembl.org/index.html

Veri tabanlarının Karşılaştırılması

NCBI

- En popüler olan ve sıklıkla kullanılan biyolojik veri tabanıdır.
- Barındırdığı bilgi yoğunluğu ve çeşidi bakımından en az Ensembl kadar zengin bir veri tabanıdır.
- Kullanıcıya sunulan belirtim tabloları ve bu tabloların görsel sunumu Ensembl'a ve UCSC'ye göre daha azdır.
- Entrez uygulaması ile NCBI'ın içinde barındırdığı bilgiler birleştirilmiştir.
- Genomik bilgiler ile ilgili olan diğer veri tabanlarına veya uygulamalarına yönlendiren bağlantıları barındırmaktadır.
- Genom dizilerinin yanında protein yapı tahminlerini de içeren bir veri tabanıdır.

UCSC

- İlk genom taramasını sağlayan, veri tabanıdır
- Genomların basit bir dizi olarak görselleştirilmesi sayesinde, sıklıkla tercih edilen bir veri tabanıdır
- Her seviyeden kullanıcının rahatlıkla kullanabileceği, birçok belirtim seçeneği vardır ve bu belirtimleri grafiksel olarak incelemek mümkündür.
- BLAST uygulamasından daha hızlı ve daha iyi sonuç veren, BLAT dizi hizalamasına göre arama uygulamasını, içinde barındırır.
- USCS veri tabanı içindeki organizmaların, referans dizilimlerine ve oluşturulmakta olan dizilimlere ulaşılabilir.
- Genlerin ve dizilerin lokasyonlarını görsel olarak incelemek mümkündür.
- Diğer veri tabanlarına geçiş sağlayan bağlantıları barındırır

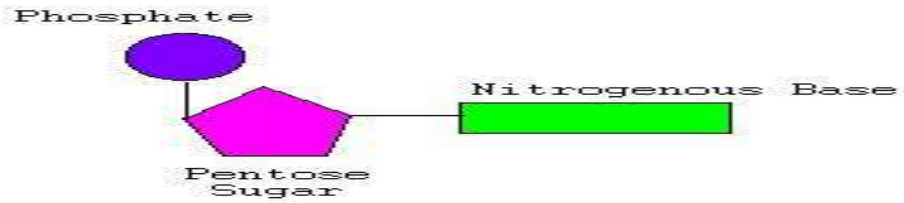
Ensembl

- İçerdiği organizma sayısına göre ve içerdiği bilgiye göre geniş kapsamlı bir veri tabanıdır. Organizmaların genetik özelliklerinin yanı sıra, birçok uygulamayı da içinde barındırır.
- Kullanılması göreceli olarak zordur, belli bir tecrübe gerektirir, diğer veri tabanlarına göre göreceli olarak karmaşıktır.
- UCSC deki gibi direk olarak eklenebilecek kullanıcı odaklı belirtim seçenekleri azdır fakat ileri seviyedeki kullanıcıların görselleştirebilecekleri pek çok seçenek vardır, hatta kullanıcıların kendi özel belirtimlerini genom üstüne eklemesi mümkündür.
- Sıklıkla güncellenir ve yeni özellikler sıklıkla eklenmektedir.
- İçinde barındırdığı bilgi yoğunluğu ve bilgi çeşitleri bakımından zengin bir veri tabanıdır.
- Genomik bilgiler ile ilgili olan diğer veri tabanlarına veya uygulamalara yönlendiren bağlantıları barındırmaktadır.

Önemli Terimler:

Nükleotit

Nükleotit, bir fosfat, beş karbonlu bir şeker (pentoz) ve bir azotlu organik bazdan oluşan bir kimyasal bileşiktir. Nükleik asitlerin (DNA ve RNA) yapı taşlarını oluştururlar.



Gen

Protein veya fonksiyonel RNA üretilmesinden sorumlu olan ve çeşitli uzunluklarda olabilen DNA dizilerine denilmektedir. İnsan genomunun %5'inin şu ana kadar genlerden oluştuğu bilinmektedir. Geri kalan kısım ise, görevleri henüz açıklanamamış olan, kodlama özelliği olmayan DNA olarak adlandırılır.

Genom

Bir organizmanın sahip olduğu kalıtımsal materyalinin tümünü ifade eder (kodlanan DNA + kodlanmayan DNA). Ökaryot organizmalar için, kromozomlarında bulunan bütün DNA sekansı, bakteriler için genomik DNA'da bulunan bütün DNA dizisi, virüsler için barındırdığı bütün DNA ve RNA, ilgili organizmanın genomu olarak ifade edilir.

Genomiks

Organizmaların, genomlarına odaklanan araştırma alanıdır. Organizmanın sahip olduğu bütün genleri veya foksyonel birimleri, birbirleriyle etkileşim halinde incelenmesini içeren bir araştırma alanıdır.

STS (işaretlenmiş dizi bölgeleri)

İlgili organizmanın genomunda yeri ve dizi özellikleri bilinen yaklaşık olarak 200-500 baz çiftinden oluşan DNA dizilerine denmektedir. Dizi ve lokasyon bilgileri bilindiğinden dolayı genetik haritalama işlemlerinde, belirteç olarak kullanılmaktadır.

EST (İfade edilmiş dizi etiketleri)

Bu diziler, bilinen bir genin ifade edilen, küçük kısımlarını belirtmektedirler. İlgili genlerin cDNA kütüphanelerinden elde edilirler, dolayısıyla protein kodlama görevi olan DNA'nın kısımlarını belirtirler. EST dizileri, bilinmeyen genlerin belirlenmesi ve bu genlerin genomdaki yerlerinin tespit edilmesi gibi işlemler için sıklıkla kullanılmaktadır.

Proteome:

Proteom, genom tarafından ifade edilen bütün protein ürünlerini ifade etmektedir. Bir organizmanın proteomu veya hücrenin veya dokunun proteomu olarak ifade edilebilir. Daha ayrıntılı belirtmek gerekirse, belli şartlar ve uyarıcılar altında hücre tarafından ifade edilen bütün proteinleri ifade eder.

Proteomiks:

Belirli bir genom tarafından ifade edilen proteinleri bütünsel ve birbiriyle etkileşimlerini dahil ederek inceleyen araştırma alanıdır. Bu alan, protein düzeyinde, gen ifadesi örüntüleri protein ve genom ilişkileri, protein-protein etkileşimlerini, protein modifikasyonlarını vs. incelemektedir.

NCBI Alt Veri Tabanları

- **PubMed**

PubMed, NCBI bünyesinde bulunan, yaklaşık 20 milyon atıflık biyomedikal literatürü barındıran, NCBI'ın makale, kitap vb. ile ilgili bilgileri barındırdığı alt veri tabanıdır. PubMed linkinden, anahtar kelimeye göre bilimsel dergi, makale, kitap vb.aramalar yapılabilmektedir. Dahası, NCBI da yapılan özgün aramalarda ve çıkan sonuçlarda NCBI, PubMed linkine erişim sağlayan bağlantıları sunmaktadır. Örnek vermemiz gerekirse, NCBI bünyesinde gen arama işlemi yapılırken, NCBI'ın sağladığı bağlantılarla, kolay bir şekilde ilgili gen ile ilgili yayınlara ulaşılması mümkündür.

- **OMIM (Online Mendelian Inheritance in Man)**

OMIM bilinen insan genleri ve bu genlerin ilişkilendirildiği hastalık fenotip bilgisini barındıran NCBI alt veri tabanıdır. Sıklıkla güncellenmektedir. İnsan genleri ve bu genlerin ilişkilendirildiği hastalıklar, hastalıkların özellikleri ve bu hastalıkların moleküler mekanizmaları ile ilgili birçok özet bilgiyi ve ilgili referansları barındırır.

- **Nücleotide (Nükleotit Veri Tabanı)**

Nükleotit veri tabanı GenBank, Refseq, TPA ve PDB gibi çeşitli kaynaklardan toplanmış ve düzenlenmiş, DNA ve RNA dizi bilgilerini barındıran veri tabanıdır. Bu amaçla genom, gen, transkript dizi bilgisini barındırır.

- **GSS Bölümü(Genome Survey Sequence)**

Bu bölümde barındırılan diziler, EST'lere benzemektedir. GSS dizileri, karakterize edilmemiş, kısa parçalar halinde olan, genomik DNA parçalarının dizi bilgilerini belirtmektedir.

- **Protein**

İlgili gen ürünlerinin, ifade ettiği proteinlerin dizi bilgisini belirten bölümdür.

- **Unigene**

Unigene bölümü, ilgili genlerin ve ifade edilen psödogenlerin (yalancı genler) ürünlerini, ifade edildikleri gen ismi altında toplayan bölümdür. Başka bir deyişle, bir geninin birden

fazla transkripti olabilir, Unigene bölümü bunun gibi birden fazla RNA ürünü olan genleri tek bir gen ismi adı altında belirtmektedir.

❖ **Genome**

Bütünsel genom dizi bilgilerini barındırır.

❖ **Structure**

İlgili genlerin, belirttiği RNA ürünlerinin ifade ettiği proteinler ile ilgili yapı bilgilerini içermektedir.

❖ **Taxonomy**

Taksonomi alt veri tabanı, genetik veri tabanlarında, en az bir protein veya nükleotit dizi bilgisi bulunan organizmaların Latince isimlerini ve evrimsel sınıflandırılmaları ile ilgili detaylı bilgileri barındırmaktadır.

❖ **SNP (tek nükleotit varyasyonları)**

İnsan genomunda ve diğer genomlarda, en sık rastlanan çeşitlilik, tek nükleotit polimorfizmleridir (SNP). Yaklaşık olarak, insan genomunda, her 100-300 bazda bir SNP'lere rastlanmaktadır. SNP'lerin sık olması ve diğer çeşitliliklere göre, kolay tanımlanmalarından dolayı, genom boyutunda ilişkilendirilme çalışmalarında sıklıkla kullanılmaktadırlar.

NCBI SNP veri tabanı, insan ve diğer organizmaların genomunda bulunan SNP'ler için kaynak görevi görmektedir. Bu amaçla belirlenen, varlığı ve ilişkisi doğrulanan SNP'ler ile ilgili detaylı bilgileri barındırmaktadır.

❖ **HomoloGene**

Ökaryot organizmalarda, gen olduğu kanıtlanmış ve belirtimi yapılmış genlerin, diğer organizmalardaki homolog eşlerini bulmaya yarayan, homolog arama sistemini ve homolog gen gruplarını içermektedir.

❖ **RefSeq**

Veri tabanlarında birçok dizi birden fazla kez belirtilmiş ve gösterilmiştir. Sekans bilgileri için gereksiz fazlalığı olan gösterimleri engellemek ve bu belirtimleri düzenlemek için NCBI, RefSeq ikincil alt veri tabanını oluşturmuştur. Bu amaçla RefSeq bölümü, genomik DNA,

RNA ve protein dizi bilgileri için, geniş kapsamlı, düzenlenmiş ve gerekli olan sekans bilgilerini tekrar düzenlemiştir. Bir başka deyişle, Refseq bölümü, her bir DNA, RNA ve protein dizisi için doğruluğu kanıtlanmış ve kabul edilmiş sekans bilgilerini içerir.

Accession number (Erişim Numarası)

Erişim numarası veri tabanı araştırmalarında, sıklıkla kullanılan bir ifade şeklidir. Erişim numarası, her bir sekans kaydını belirten özgün gösterim şeklidir. Bir başka değişle veri tabanındaki her bir DNA dizisi, RNA dizisi ve protein dizisi için özgün erişim numaraları vardır. Eğer sorgulatılmak istenen nükleik asit dizisinin veya protein dizisinin erişim numarası elimizde mevcut ise, NCBI veri tabanında bu numara ile de arama yapabilmemiz mümkündür. Erişim numaraları harflerden ve numaralardan oluşur ve başındaki harflerin özelliğine göre, hangi çeşit molekül olduğu bilgisini de içinde barındırır.

Erişimi numarası	**Molekül çeşidi**	**Açıklama**
AC_123456	Genomik	Analizi tamamlanmış, alternatif genomik DNA dizisini ifade eder.”A” harfi, alternatif kurulumu veya belirtimi ifade eder.
AP_123456	Protein	Analizi tamamlanmış, alternatif protein dizi bilgisini belirtir.
NC_123456	Genomik	Bütünsel genomik molekülleri ifade eder (kromozomlar, organel DNA’sı, plasmidler)
NG_123456	Genomik	Analizi ve belirtimi tamamlanmamış genomik bölgeleri ifade eder.
NM_123456	mRNA	Genlerin transkript ürünlerini ifade etmektedir (haberci RNA). **Veri tabanlarında bulunan, mRNA’lar, aslında ilgili RNA’ların cDNA’ya çevrilmiş hallerini belirtmektedir.**
NP_123456	Protein	Protein dizilerini ifade eder.
NR_123456	RNA	Protein ifadesi olmayan RNA dizilerini belirtir (yapısal RNA’lar, ifade edilmiş pseudogenler vs.)

NT_123456	Genomik	Bakteri yapay kromozomuna(BAC) klonlanmış veya genom boyutunda rastgele dizileme metoduyla dizi bilgisi çıkartılmış genom dizilerini ifade eder.
NW_123456	Genomik	Bakteri yapay kromozomuna(BAC) klonlanmış veya genom boyutunda rastgele dizileme metoduyla dizi bilgisi çıkartılmış genom dizilerini ifade eder.
XM_123456	mRNA	Genomik kontig sekansına göre, genom belirtim sürecinde, modellenmiş RNA dizilerini ifade eder.(protein ifade etme özelliği olan RNA ürünleri için)
XP_123456	Protein	Genomik kontig sekansına göre, genom belirtim sürecinde, modellenmiş Protein dizilerini ifade eder.
XR_123456	RNA	Genomik kontig sekansına göre, genom belirtim sürecinde, modellenmiş RNA dizilerini ifade eder.(protein ifade etme özelliği **olmayan** RNA ürünleri için)
YP_123456	Protein	Protein ürünü mevcut olan ancak, ilgili transkript ile ilgili bilgisi olmayan, Protein dizilerini ifade eder. (Birincil olarak, bakteri, virüs ve mitokondri için kullanılır)
ZP_12345678	Protein	Hesaplamalı ve tahminsel yöntemlerle belirlenmiş proteinleri ifade eder.

BLAST

Nükleotit ve protein dizilerinin, aynı organizma ile veya farklı organizmalar ile karşılaştırılması, moleküler biyoloji araştırmalarında, çeşitli amaçlar için sıklıkla kullanılmaktadır. Araştırmacılar dizi benzerlikleri ve dizilerin karşılaştırılması ile,yeni bulunmuş ve dizi bilgisi çıkartılmış genlerin görevlerini tahmin edebilmektedirler. Dahası, gen ailelerinin belirlenmesi, organizmalar arasındaki evrimsel ilişkilerin ortaya çıkartılması gibi birçok alanda dizi benzerliklerinden yararlanılmaktadır.

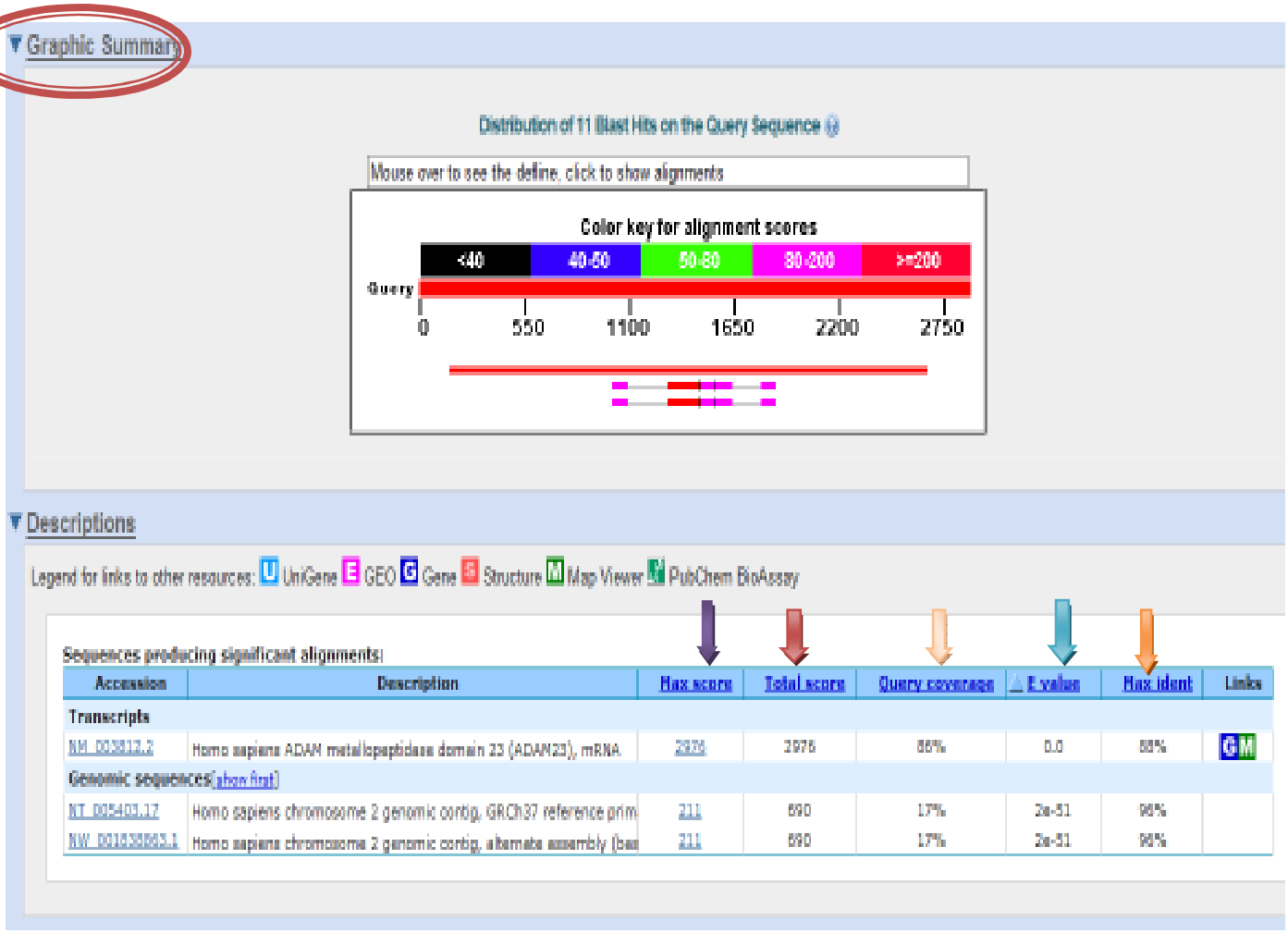
BLAST uygulaması sorgulatılmak istenen protein veya nükleik asit dizisini, benzerlik kıstaslarına ve kendi içinde barındırdığı algoritmaya göre, veri tabanı içinde arayan bir dizi karşılaştırma programıdır. BLAST, sorgulatılan diziyi veri tabanı içindeki diğer dizilerle karşılaştırabildiği gibi kullanıcı tanımlı dizileri ikili olarak da karşılaştırabilmektedir.

Çeşitli amaçlar için BLAST seçenekleri mevcuttur;

BLAST Türü	**Birinci sorgulama**	**İkinci sorgulama**
Blastn, megablast, tblastx	**Nükleotit**	**Nükleotit**
Blastx	**Nükleotit**	**Protein**
Tblastn	**Protein**	**Nükleotit**
Blastp	**Protein**	**Protein**

Sorgulama işlemi için özel olarak ayrılmış olan alana, dizi bilgisi yazılarak, dizi bilgisi kopyalanıp yapıştırılarak ve erişim numaraları veya gen kimlik numaraları kullanılarak da yapılabilmektedir.

- Aşağıda ADAM23 geninin 1. RNA ürünü için gerçekleştirilen BLAST sorgulama sonuç sayfası örnek olarak verilmiştir.



Grafiksel Özet (Graphical Summary) başlığı altında, sorgulatılan dizinin, BLAST arama sonuçları ile eşleştirilmiş şekli grafiksel olarak kullanıcıya sunulmuştur. Sorgulamaya verdiğimiz dizinin, benzerlik gösterdiği diğer diziler, en yüksek benzerlik gösterenden en az benzerlik gösterene doğru, yukarıdan aşağıya doğru sıralanmış bir şekilde kullanıcıya sunulmaktadır.

BLAST arama sonuçlarının karşılaştırılmasında önemli olan bazı parametreler vardır. Bu parametrelere ve bu parametreler arasındaki ilişki incelenerek, sonuçların güvenirliliği veya sorulan bilimsel soruya göre, arama sonuçlarını seçmek mümkündür. BLAST arama sonuçlarının karşılaştırılmasında kullanılan değişkenler;

- **Maksimum Skor (Maximum Score)**
- **Toplam Skor (Total Score)**
- **Sorgulama Kapsamı (Query Coverage)**
- **E-Değeri (E-Value)**
- **Maksimum Benzerlik (Maximum Identity)**

Bu parametrelerin hepsi sonuçların değerlendirilmesi için kullanılmaktadır. Ancak bunlar arasındaki en önemli parametre E-değeridir. E-değeri, yaptığımız hizalamaların şans eseri olma ihtimalinin hesaplanması ile sonuçların istatistiksel önemini değerlendirmemizi sağlayan bir parametredir. Bu durumda E-değerimiz 0 'a eşit ise, sorguladığımız dizi, çıkan sonuç ile bire bir eşleşmiş demektir ve bu eşleşmede şans faktörü 0 'dır. Sorgulama kapsamı, sorgulatılan dizi ile diğer dizilerin uzunluk bazında eşleşme oranını belirtmektedir. Maksimum benzerlik ise, sorgulatılan dizi ile diğer diziler arasındaki dizi benzerliğinin yüzde olarak oranını belirtmektedir.

Çıkan sonuçların değerlendirilmesinde, En düşük E-değerine sahip olan, maksimum benzerliği ve sorgulama kapsamı en yüksek olan sonuçlara öncelik verilir. Ancak, BLAST sonuçlarının değerlendirilmesi ve seçimi, sorulan bilimsel biyolojik soruya göre değişiklik gösterebilmektedir.

Referanslar:

- **The National Center for Biotechnology Information**

http://www.ncbi.nlm.nih.gov/

- **The NCBI Handbook**

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook

- **The Genome User's Guide**
  Nature Genetics Supplement September 2003 issue