

Biyolojik Veritabanları ve Biyoinformatik Analiz Araçları

Aslı Yazğan

Özet: Günümüzde bilgi üretimi, bu bilgilerin yorumlanmasından daha hızlı bir şekilde ilerlemektedir. Bu nedenle bilişim alanında öncelikli araştırma konularından birisi büyük verilerin yönetimi olmuştur. Biyoloji ve medikal alanında veri analizleri, moleküler biyoloji ve biyomedikal tekniklerinin gelişmesiyle gerekli hale gelmiştir. İşlenmemiş biyolojik ve medikal verilerden anlamlı bilgiler çıkarabilmek zorlu bir işidir. Bu çalışma biyoinformatik alanının önemini, araştırma konularını açıklamayı ve en çok kullanılan biyolojik veritabanları ve internet üzerinden kullanılan analiz araçlarının genel tanıtımını yapmayı amaçlamaktadır.

Anahtar Sözcükler: Biyoinformatik, biyolojik veritabanları, biyoinformatik analiz araçları, genom analizi, protein analizi

Biological Databases and Bioinformatic Analysis Tools

Abstract: Due to advances in biomedical techniques and molecular genomic, the rate of accumulation of medical and biological data is much faster than the rate of data interpretation. Extracting useful information from raw biomedical data is a challenging task so that the data management is a core issue in medical domain. This paper introduces the importance of bioinformatics field and its subjects, discusses widely used bioinformatic databases and analysis tools available online.

Keywords: Bioinformatics, biological databases, bioinformatics analysis tools, genomic analysis, protein analysis

1. Giriş

Moleküler biyoloji ve Genombilim alanındaki ilerlemeler, dünyada özellikle sağlık sektöründe devrim niteliğinde değişikliklere yol açmaktadır. Bu durum, biyolojik ve medikal araştırmaların yapılaş şekli kökünden değiştirmektedir. Hastalıklarla ilişkili olabilecek yeni genlerin, genler arasındaki ilişkilerin, genom yapılarının belirlenmesi gibi çalışmaların, gelecekteki koruma, tanı, tedavi ve ilaç keşfi araştırmalarını doğrudan etkileyeceği öngörülmektedir [21].

İnsan genetik kodunun ve DNA dizisinin çözümlenmesi amacıyla 1987 yılında *İnsan Genom Projesi* Amerika tarafından başlatılmış, 2001 yılında insan genomunun büyük bir bölümü açıklanmıştır. İnsan Genom Projesinin ilk sonuçlarına göre insan genomu, %99,9'u tüm insanlarda aynı olan yaklaşık 3.10^9 nükleotidden oluşmaktadır ve yaklaşık 20.500 gen içermektedir [1]. Genomik bilgi akışının yoğunluğu ve boyutu, bu verilerin yönetiminin sistematik olarak gerçekleştirilmesi gerekliliğini ortaya çıkarmıştır. Ayrıca İnsan Hakları Evrensel

Bildirgesinde bu bilgilerin ücretsiz ve kısıtsız bir şekilde herkesin erişimine açık olacağı bildirilmiştir [10]. Dolayısıyla, bu verilerin depolanması ve korunması için bilgisayar teknolojilerinden, paylaşımı için de internet teknolojilerinden faydalanmak kaçınılmaz olmuştur. Başta Amerika ve Avrupa olmak üzere Japonya, Çin, Hindistan gibi ülkeler bu alanda çalışmalar yapmaktadırlar. Tüm bu gelişmelerin sonucu olarak internet üzerinden erişilebilen biyolojik veritabanlarının sayısında büyük bir artış gözlenmiştir.

Üretilen ve depolanan verilerin analiz edilmesi de önemli diğer bir boyuttur. Veri boyutunun ve yoğunluğunun çok olması, analizlerde bilişimsel analiz yöntemleri kullanılmasını gerekli hale getirmiştir. Bu da disiplinler arası yeni bir alan olan biyoinformatiğin doğuşuna sebep olmuştur.

Bu çalışmada öncelikle biyoinformatik alanı ve araştırma konuları kısaca açıklanmıştır. Sonrasında en çok kullanılan, internet üzerinden erişilebilen biyolojik veritabanları ve analiz araçlarını Avrupa ve Amerika tabanlı kaynaklar üzerinden tanıtılmıştır.

2. Biyoinformatik

İnsan Genom projesinin çalışmaları, biyoinformatik alanının gelişmesine önemli bir katkı sağlamıştır. Biyoinformatiğin gelişmesinden önce genlerin kromozom üzerinde belirlenmesinin tek yolu; canlı üzerinde, “*in vivo*”, davranışlarının çalışılması ile ya da “*in vitro*” izole edilen DNA'nın test tüpünde incelenmesi ile mümkündür. Bioinformatik sayesinde DNA ve protein bilgisi belirli harflerden oluşan, nümerik olmayan diziler şeklinde ifade edilmiş ve bu çok boyutlu verilerin analizinde bilişimsel yöntemler kullanılmıştır. Gen dizisi bilgisi üzerinde bilgisayarlar yardımı ile araştırmalar yapılabilir ve gen grupları arasındaki potansiyel ilişkiler ortaya çıkarılabilir hale gelmiştir [16].

Genetik bilgiler üzerindeki analizler, hastalık-gen arasındaki ilişkilerin ortaya çıkarılması, hastalıklarının anlaşılması, hastalıkların ilerlemesindeki yolların saptanması ve buna göre tedavilerin geliştirilebilmesi amacıyla büyük önem arz etmektedir.

Biyoinformatik araştırmalarının üç temel amacı vardır [11] :

- Varolan biyolojik verileri araştırmacıların ulaşabileceği biçimde organize etmek ve yeni bilgiler üretildikçe veriler üzerine ekleme yapabilmeye imkan vermek
- Var olan verilerin analizleri için araçlar geliştirmek
- Analiz araçlarını kullanarak, verileri biyolojik olarak yorumlayabilmek

Yukarıda belirtilen amaçlar doğrultusunda, biyolojik veritabanları ve biyoinformatik analiz araçları geliştirilmektedir. Bu çalışma kapsamında en çok kullanılan Avrupa ve Amerika tabanlı biyoinformatik araştırma kaynakları tanıtılacaktır.

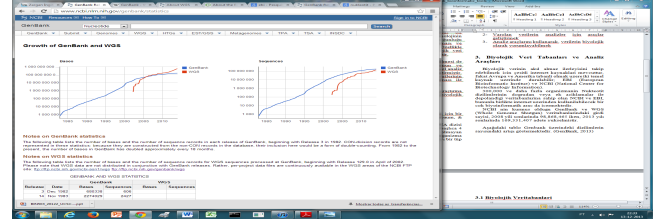
3. Biyolojik Veri Tabanları

Araştırmacıların var olan biyolojik verilere kolayca ulaşabilmeleri ve elde ettikleri yeni sonuçları kolayca paylaşabilmeleri için bazı kuruluşlar internet üzerinden ücretsiz erişilebilecek veritabanları kurmuşlardır. Günümüzde en aktif olarak bilinen kuruluşlar, Avrupa tabanlı EBI (European Bioinformatic Institute), Amerika tabanlı NCBI (National Center for Biotechnology Information) ve Japonya tabanlı Japanese Institute

of Genetics'dir. Bu kuruluşların başlattığı sırasıyla ENA (European Nükleotid Archieve), GenBank ve DDBJ (DNA Data Bank of Japan) en çok bilinen ve kullanılan elektronik veritabanlarıdır. Bu üç kuruluş nükleotid dizi bilgilerinin toplanması ve paylaşılmasında birbirleri ile ortak çalışırlar. Aynı zamanda bu kuruluşlar internet üzerinden kullanılacak birçok biyoinformatik aracını da araştırmacıların hizmetine sunmaktadırlar. Bu çalışmada, NCBI ve EBI kuruluşlarının sunduğu biyoinformaik araçları tanıtılacaktır.

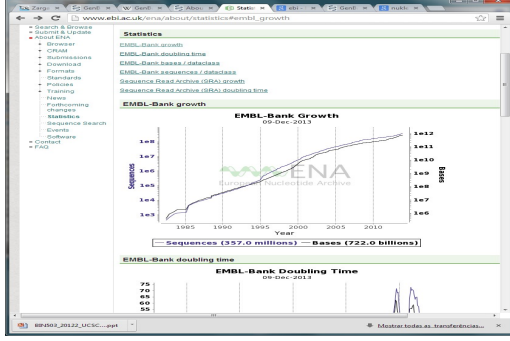
NCBI'nin kurmuş olduğu GenBank ve WGS (Whole Genome Shotgun) nükleotid dizilimlerinin doğrudan veya ek açıklamalar (annotation) ile depolandığı veritabanlarıdır. Bu veritabanlarındaki kayıt sayısı, 2008 yılı sonlarında 98,868,465 iken, 2013 yılı sonlarında 169,331,407 adete yükselmiştir. GenBank veritabanındaki tüm gen dizilimleri ek açıklamalar ile desteklenen dizilimlerdir. WGS veritabanındaki dizilimler tamamlanmamış dizilimlerdir. Bu dizilimler ek açıklama içermek zorunda değildirler.

Şekil 1'deki grafiklerde Genbank ve WGS üzerindeki dizilimlerin ve bazların sayısındaki artış görülmektedir [9].



Şekil : GenBank ve WGS veritabanlarının gelişimi

EBI'nin kurmuş olduğu ENA (European Nucleotide Archive) veritabanındaki dizilim ve bazların sayısındaki artış Şekil 2'deki grafikteki gibi görülmektedir.



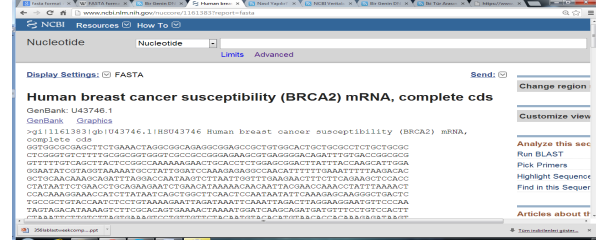
Şekil : ENA veritabanının gelişimi

Grafiklerde de görüldüğü gibi hızla artan biyolojik veriler içinden belli bir amaca yönelik olanları ayırmak zorlaşmaktadır. Bu nedenle veriler farklı amaçlar için kategorize edilerek çeşitli veritabanları oluşturulmuştur. Wikipedia biyolojik veritabanları listesi başlığı altında 19 kategori belirlemiştir [23]. Nükleotid dizilimlerin yanında, proteinler aileleri, yapıları ve dizilimleri ile ilgili, genetik varyasyonlarla, biyolojik yollarla (pathways) ilgili birçok biyolojik veritabanı mevcuttur. Bu çalışmada biyolojik veritabanları, nükleotid dizilim veritabanları ve protein veritabanları olmak üzere iki farklı kategoride incelenmektedir.

3.1 Nükleotid Dizilim Veritabanları

Nükleotid verisi üzerine tasarlanmış veritabanlarından olan Avrupa tabanlı Ensembl ile Amerika tabanlı GenBank bu bölümde tanıtılmıştır. Bu veritabanları genom bilgisini grafiksel öğelerle destekleyen ve bu bilgi üzerinde kaliteli, tutarlı ek açıklamalar sunan, devamlı güncellenen ve internet üzerinden ulaşılabilen veritabanlarıdır [17]. Bu iki veritabanı da genom üzerine çalışan araştırmacılar için önemli başvuru kaynaklarıdır.

Bu veritabanları hayvan, bitki, mantar, bakteri gibi birçok canlı âlemi için nükleotid dizilimlerini metin formatında depolar. Nükleotid dizilimlerinin metin tabanlı gösterilmesi “FASTA” formatıyla belirlidir. FASTA formatında her harf bir nükleotidi simgeler. FASTA formatı sayesinde dizilimler karşılaştırılabilir ve dizilimlerin benzerlikleri araştırılabilir. FASTA formatının ilk satırı açıklama satırıdır. Açıklama satırı “>” işareti ile başlar, ardından dizilimin id numarası, adı, kromozom numarası gibi bilgileri içerir. Bu bilgilerin ardından nükleotid dizilim görülür. Tüm dizilimler FASTA formatında, bir metin dosyası olarak indirilebilir.

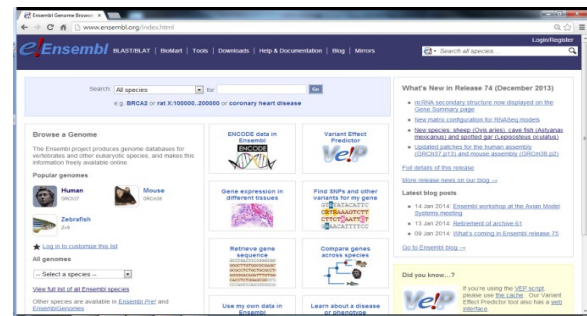


Şekil : Fasta Formatı

3.1.1 Ensembl

Avrupa Biyoinformatik Enstitüsünün 1999’da başlattığı bir proje olan Ensembl, başta insan, fare, sıçan ve zebrafish olmak üzere birçok canlı âlemi için genom bilgisi içeren, devamlı güncellenen ve <http://ensembl.org> internet adresinden ulaşılabilen bir biyolojik veritabanıdır. Gen bilgisi üzerinde detaylı açıklamalar içerir. Tüm dizilimler fasta formatında görüntülenebilir ve indirilebilir. Organizmaların kromozom dizisi ve genom verisi grafik olarak detaylı görüntülenebilir. Ayrıca farklı organizmaların kromozomları arasında karşılaştırma yapılabilir. Grafik üzerindeki veriler araştırmacının istediğine göre gizlenip tekrar görüntülenebilir.

Ensembl veritabanı bir gen adı, bir hastalık adı ve ya belirli bir tür için kromozom üzerinde bir lokasyon belirlenerek de arama yapılabilir. Bir hastalık adı ile arandığında o hastalıkla ilgili olabilecek muhtemel genleri ve transkriptleri listeler. Bir gen adı ile arandığında farklı türlerdeki ilgili geni ve transkriptlerini listeler. Belirli bir lokasyon ile arandığında ise o lokasyon için dizilim görüntülenir.



Şekil : Ensembl Arama Sayfası

Belirli bir genin dizilimi görüntülediğinde gen ile ilgili kısa bir bilgi, genin kromozom üzerindeki yeri, transkript sayısı gibi genel bilgiler görüntülenir. “Show transcript table” butonuna basıldığında transkriptler ile ilgili tablo görüntülenir.

Gen: BRCA2 ENSG00000139618

Description: breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]

Location: Chromosome 13:32,889,611-32,973,805 forward strand

Structure: chromosome GRCh37:CM000675.1:32889611-32973805.1

Transcripts: This gene has 8 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CDS incomplete	CCDS
BRCA2-201	ENST00000444453	10964	ENSP00000429927	3418	Protein coding	-	CCDS93344
BRCA2-001	ENST00000380152	10930	ENSP00000394957	3418	Protein coding	-	CCDS93344
BRCA2-003	ENST00000530883	2011	ENSP00000435692	481	Protein coding	3	-
BRCA2-002	ENST00000430884	642	ENSP00000433698	180	Nonsense mediated decay	5	-
BRCA2-005	ENST00000528762	495	ENSP00000433168	64	Nonsense mediated decay	5	-
BRCA2-006	ENST00000533770	523	No protein product	-	Retained intron	-	-

Summary

Name: BRCA2 (HGNC Symbol)

Synonyms: BRCC2, FACD, FAD, FAD1, FANCD, FANCD1 [View all Ensembl genes linked to the name]

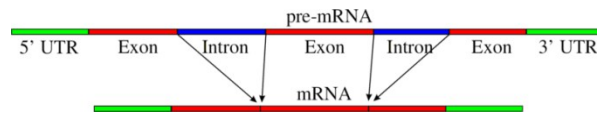
CCDS: This gene is a member of the Human CCDS set: CCDS93344

LRG: LRG_293 provides a stable genomic reference framework for describing sequence variations for this gene.

Şekil : Ensembl'da aranan gen için genel bilgiler ve transkript tablosu

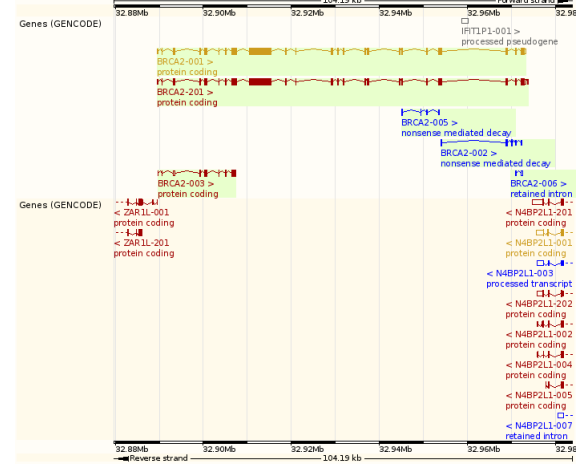
Transkript tablosu bir gen için tüm kesim varyan (*splice variants*) ları kodlanmamış transkriptlerle birlikte gösterir. Eğer transkript CCDS (Consensus CoDing Sequence) listesinde yer alıyor ise CCDS numarası da tabloda listelenir. CCDS; temel protein kodlama bölgelerinin belirlenmesi üzerine çalışmalar yapan EBI, NCBI, UCLA (the University of California at Santa Cruz) ve WTSI (Wellcome Trust Sanger Institute) enstitülerinin bir araya gelmesiyle oluşmuş bir kuruluştur. CSDD' de yer alan transkriptler gözden geçirilmiş ve yüksek kalitede olan transkriptlerdir.

Gen için olan genel bilgiler ve özet bilgiler dışında sayfanın alt kısmına doğru, gen ve farklı transkriptler hakkındaki ek açıklamalar içeren bir grafik arayüz görüntülenir. Bu grafik arayüzde genin protein kodlayan kısımları ve kodlamayan kısımları görülmektedir. Protein ve enzimler üretilirken, DNA üzerindeki genlerin harf dizilimleri örnek alınarak bu genlere karşılık gelen mRNA kopya dizilimleri çıkarılır. Kopyalanan bu mRNA yapılırken, genin harf dizilimi baştan sonra tümüyle okunmaz. Bir kısmı okunup kopyası çıkartıldıktan sonra bir bölüm okunmadan atlanıp başka bir bölüme geçilir ve oradan okunmaya devam edilir. DNA'nın okunmadan atlanan bu bölümlerine Intron denir ve intronlar protein kodlamasına katılmazlar. Kodlanan diğer kısımlara ise ekson adı verilir [22].



Şekil : Ekson ve Intron

Şekil 7 de görülen grafikte kutular eksonlar, çizgiler intronlardır. Dolu kutular kodlanmış dizilerdir. Boş veya tamamlanmamış kutular ise dizilenmemiş bölgeleri (UTR) ifade eder.



Şekil : Ensembl transkriptlerin grafiksel gösterimi

Transkriptlerin farklı renklerle ifade edilmesi, transkriptlerle ilgili ek açıklamaların (annotations) elde edilmesi hakkında bilgi verir. Ek açıklamalar manuel olarak uzmanlar tarafından analizler yapılarak elde edilebilir. WTSI enstitüsünde sürdürülen *Vega/Havana projesi* sonucu elde edilen transkriptler bu şekilde olanlardır. Manuel yapılan analizler uzun sürmebilir ve yavaş ilerleyebilir. Bu nedenle, Ensembl üzerinde transkriptler için "*otomatik açıklama ekleme sistemi*" (*Automated ensembl annotation pipe*) kurulmuştur. Otomatik açıklama ekleme sistemi, bilim çevrelerinin oluşturduğu erişime açık mRNA ve protein dizilimi veritabanları üzerinde çalışmakta olan bir sistemdir [5].

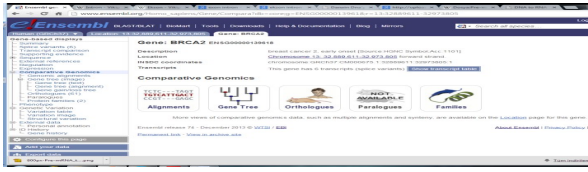
Kırmızı transkripler *otomatik açıklama ekleme sistemi* ile elde edilmiş veya manuel olarak yapılan *VEGA/Havana projesi* sonucu elde edilmiş transkriptlerdir. Ensembl'daki tüm bu transkriptler deney sonuçları ile desteklenmektedir. *Otomatik açıklama ekleme sistemi* ile elde edilmiş transkriptlerin ismi 2 ile başlarken *VEGA/Havana projesi* sonucu oluşturulan transkript isimleri 0 ile başlar. Örneğin; BRCA2-201 *otomatik açıklama ekleme sistemi* ile elde edilmiş transkript iken BRCA2-001 *VEGA/Havana projesi* sonucu elde edilen bir transkripttir.

Altın renkli olan transkriptler hem *VEGA/Havana projesi* hem de *otomatik açıklama ekleme sistemi* ile aynı şekilde elde edilen transkriptlerdir. Mavi, pembe veya gri olan transkripler ise kodlanmamış (non-coding) transkriptlerdir.

Sonuç olarak; Ensembl veritabanında yapılan arama sonucu gen ve transkriptler hakkında ilgili elde edilen

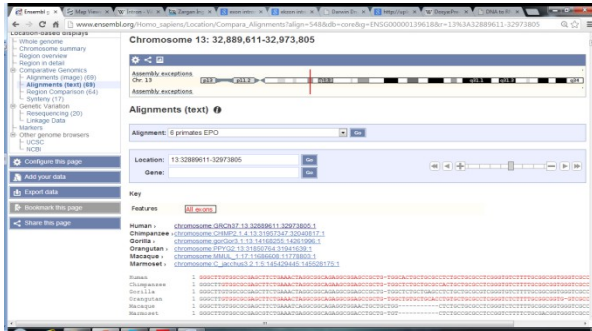
veriler arasında, gözden geçirilmiş ve yüksek kaliteli plan transkriptler, grafikte altın renkli olan transkriptler ve transkript tablosunda CSDD kimlik numarasına sahip olan transkriptlerdir.

Ensembl ile transkriptlerin detaylı incelenmesi dışında, farklı türlerdeki aynı genin birbirleri ile karşılaştırılması, genin farklı organizmalardaki genler ile benzerliklerinin ağaç gösterimi, bu gen ile ortolog ve paralog olan genlerin görüntülenmesi de mümkündür. Soldaki menüden “Comparative Genomics” seçeneği seçilerek bu işlemler gerçekleştirilebilir.



Şekil : Comparative Genomics seçenekleri

Örneğin, Şekil 9’da altı farklı tür için aynı genin dizilimleri listelenmiştir. Bu veri FASTA formatında indirilebilir.

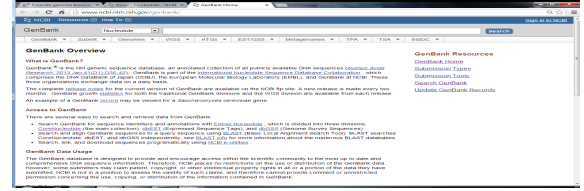


Şekil : Farklı türlerin dizilemesi

Ensembl veritabanı üzerinde çalışan farklı amaçlar için özelleştirilmiş analiz araçları mevcuttur. Bu araçlardan olan BioMart ve BLAST bu çalışmanın “Analiz Araçları” kısmında anlatılacaktır.

3.1.2 GenBank

GenBank, Ensembl’a benzer olup Amerika kaynaklı bir veritabanıdır. NCBI’nin yaklaşık 30 yıl önce oluşturduğu GenBank, <http://www.ncbi.nlm.nih.gov/genbank> internet adresinden ulaşılabilen, belirli bir gen için DNA dizilimi arama ve analiz etmek için günümüzdeki güvenilir kaynaklarından biridir.



Şekil : GenBank Anasayfa

GenBank bir gen ile ilgili sorgulandığında, Ensembl’dan farklı olarak o gen hakkındaki genel bilgiler dışında, gen ile ilgili yayınlanmış makaleler ve genin dizilim güncellemeleri ile ilgili bilgiler de elde edilir. Herhangi bir GenBank verisi Şekil 11 ‘deki gibi görüntülenir. LOCUS alanı gen ile ilgili birden fazla veri içerir. İlk bilgi Locus ismidir ve dizilimlerin benzerliklerine göre gruplandırılması için kullanılır. Örneğin şekildeki gen için locus adı SCU49845’dir ve ilk üç karakter genin ait olduğu organizmayı simgeler. İkinci bilgi dizilim uzunluğudur. Bu dizilimde kaç tane nükleotid baz çifti (base pair) bulunduğunu gösterir. Üçüncü bilgi molekül tipidir. Dördüncü bilgi GenBank kategori bilgisidir. GenBank içerisinde primat dizilimleri (PRI), kemirgen dizilimleri (ROD), viral dizilimleri(VRL) gibi 18 adet kategori bulunur. Beşinci bilgi, genin en son düzenlendiği tarihi işaret eder. DEFINITION kısmında dizilimin kısa tanımı, hangi organizmadan olduğu, gen adı, bilinen fonksiyonları belirtilir. Eğer dizilimin protein kodlayan bölgesi (CDS) var ise, “complete cds” olarak bu bölümde ifade edilir. ACCESSION dizilimin GenBank üzerindeki kimlik numarasıdır. Bir dizilim güncellense bile kimlik numarası kesinlikle değişmez. VERSION dizilimin farklı versiyonlarını ifade eden numaradır. KEYWORDS dizilimi ifade eden anahtar kelimelerdir. Eğer dizilim hiçbir anahtar kelimeye sahip değilse, bu alanda nokta işareti görülür. SOURCE dizilimin hangi organizmadan olduğu bilgisidir. REFERENCE dizilim hakkında yayınlanmış makalelerin başlıkları, yazarları, yayınlandığı dergisi ve pubmed kimlik numarası ile en eskiden yeniye doğru sıralandığı bölümdür.

GenBank Flat File Format

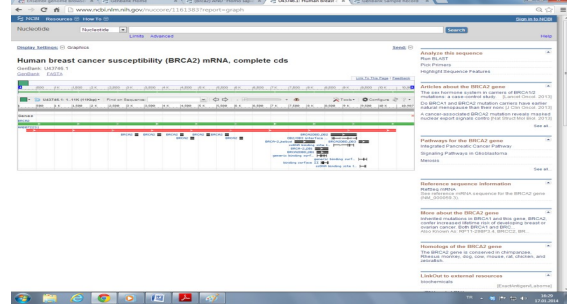
Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the [Alphabetical Quicklinks Table](#) or [Resource Guide](#)

```
LOCUS      SC049845      5028 bp      DNA      PLN      21-JUN-1999
DEFINITION Saccharomyces cerevisiae TCP1-beta gene, partial cds, and AxL2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION  U49845
VERSION    U49845.1  GI:1293613
KEYWORDS   .
SOURCE     Saccharomyces cerevisiae (baker's yeast)
ORGANISM   Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycotales; Saccharomycetaceae; Saccharomycos.
REFERENCE  1 (bases 1 to 5028)
AUTHORS    Torpey,L.E., Gibbs,F.E., Nelson,J. and Lawrence,C.W.
TITLE      Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
JOURNAL    Yeast 10 (11), 1503-1509 (1994)
FURMD     7871890
REFERENCE  2 (bases 1 to 5028)
AUTHORS    Roemer,T., Madden,K., Chang,J. and Snyder,M.
TITLE      Selection of axial growth sites in yeast requires AxL2p, a novel
            plasma membrane glycoprotein
JOURNAL    Genes Dev. 10 (7), 777-793 (1996)
FURMD     8846915
REFERENCE  3 (bases 1 to 5028)
AUTHORS    Roemer,T.
TITLE      Direct Submission
JOURNAL    Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES   Location/Qualifiers
            source             1..5028
                                /organism="Saccharomyces cerevisiae"
                                /db_xref="taxon:4932"
                                /chromosome="IX"
                                /map="9"
            CDS                1..206
                                /codon_start=3
                                /product="TCP1-beta"
                                /protein_id="AA98665.1"
                                /db_xref="GI:1293614"
                                /translation="SLSIMIGISTGGLDLNNGTIADMRQLGIVESYKLRKRVVSSAEAA
                                REVLELVNIRIRAREPTANRQNM"
            gene                687..3158
                                /gene="AXL2"
            CDS                687..3158
                                /gene="AXL2"
                                /note="plasma membrane glycoprotein"
                                /codon_start=1
                                /function="required for axial budding pattern of S.
                                cerevisiae"
                                /product="AxL2p"
                                /protein_id="AA98666.1"
                                /db_xref="GI:1293615"
                                /translation="MFDLQISLILLTATISLLELVVATPYEAYPIGQYPPVAVNESF
                                TPQINDYKSSVDKTAQITMCFELFGLFSSSTFEGEFSDDLSDANTLLVFN
                                VLEGGDSADSTLANTQVFNWTFSLSSDFNLALAKNIGYTWKRALKLFPE
                                VNFVTFRSMFTNSESIVYVRSQGLVWFLMFLFEGELETGCTGYSALAE
                                TSYSFVIATDIEGFSAVEFELVIGARQLTTSIQNSLILNVITDGNVYDPLNIV
                                FLDDIISDQSLGSLNLLAFWALNATIGQSVFDELLGNSNPANFVSIYDTTV
                                DVITPFRVYVYDPLFASISLFINMATSSENFSTYLFQPTVYVWVLEKTFSSQ
                                DHWVVFQSNLTLAGEVFRNFKLSLGLKAKQSQSQELVFNIGMDSKITHSNESA
                                NATSTKSHRSTSTSTSTSTYTKIISTSAATSSAALPAANKTSSHNKRAVATA
                                COVALFLGIVVALICILPWRGRENFGSELPRASISGFLNPAKFPQKNTLAS
                                NFFDDASVDYDTSIARLLAALMTLKLONHSAETSDISVDEKRLDLSGMNTYDQFQ
                                SQSKELLARFVQPFESFFVFQNRGSSVYMSSEFANRWRVTONLFPVSDIVDVS
                                VSGKTYVTEFLFLKARFKEKSTSDPTKSSLQPNESIISFVRSYVTFSTVYK
                                HNRHLQIQDSQSGONGITPTMTSTSSDDFVFKDGENFCWVHSMFDRRPSKRL
                                VDFSNKSNVQVQKVDIGRIFEMLS"
            gene                complement (3300..4037)
                                /gene="REV7"
            CDS                complement (3300..4037)
                                /gene="REV7"
                                /codon_start=1
                                /product="Rev7p"
                                /protein_id="AA98667.1"
                                /db_xref="GI:1293616"
                                /translation="MNRWVEKLVKCYINILFYRNVYFQSFDTTYQSNLQ
                                FVFINRFPALDYIEKLLDVLKSLTVYVYFICINKNKDLCKYVLDVDFSELQHD
                                KDIQITTEYVDFEFSLLKLNMLSEKFRWDTITFEAVVNALELGLKLDNR
                                RVSLDEKAEIERSDNVVKQDEENLFPNNFQPKIKLTELVSQVGLIHQSEK
                                LIGSDKILNIVSYQEEGSEIFGSLF"
ORIGIN     1 gatootcaat atacaacggt attocacoc attcagttaga tctacaacac ggaacattg
            61 ccagatgag acagtaggt atctogcaga gtacaagct aaacagcaga gtatgcaagt
            121 ctgcactoga agcactgaa gtctactaaa gytgagataa catcactcgt gaaagacaaa
            181 gaacocgcaa tagacaacat atgtacataa tttagatat accctgaaaa taataaacog
            241 ccacactcgt attattataa ttagaacagc aagcaaaaa ttatcocaata tataattcaa
```

Şekil :Örnek GenBank kayıtu

FEATURES kısmı gen, gen ürünleri ve genin protein, RNA kodlayan bölümleri gibi biyolojik olarak önemli bölgeleri hakkında bilgiler içerir. ORIGIN kısmında da dizilim görüntülenir.

GenBank'de bir gen dizilimi arandığında ilk ulaşılan sayfa yukarıda açıklanan sayfadır ve buna GeneBank görünümü denir. Bunun dışında bir gen dizilimi, FASTA formatında ve Grafik olarak da görüntülenebilir.



Şekil : GenBank Grafik görünümü

Sonuç olarak, Ensembl ve GenBank internet üzerinden herkesin erişimine açık, biyoinformatik alanında çalışan araştırmacılar tarafından çok kullanılan, bir çok tür için dizilim verisi içeren, ayrıca bu veriler ile ilgili olabildiğince fazla kaynaktan toplanmış ek bilgileri hem metinsel hemde grafiksel olarak sunan kapsamlı veritabanlarıdır.

Bu veritabanlarındaki veriler kullanılarak farklı amaçlar için özelleşmiş analiz araçları geliştirilmiştir. Bu çalışmanın “Biyoinformatik Analiz Araçları” bölümünde bu araçlar detaylı incelenmiştir.

3.2 Protein Veritabanları

Proteinler, aminoasitlerin birbirine bağlanması ile oluşurlar. Her proteininin kendine has özelliklerinin olmasını sağlayan özel aminoasit dizilimleri vardır. Her aminoasit bir harf ile ifade edilir. Protein dizilimleri de FASTA formatı ile gösterilir. Belli özellikleri, fonksiyonları ve yapıları benzer olan proteinler, protein aileleri oluştururlar. Protein dizilimlerinin birbirleri ile benzerlikleri analiz edilerek, belli bir proteinin hangi protein ailesine ait olduğu tahmin edilebilir.

Bu çalışmada protein ile ilgili olan veritabanları üç kategoride incelenmiştir. Protein dizilimleri için UniProt, protein aileleri için InterPro ve yapısal protein veritabanları için PDB ve PDBe veritabanları tanıtılmıştır.

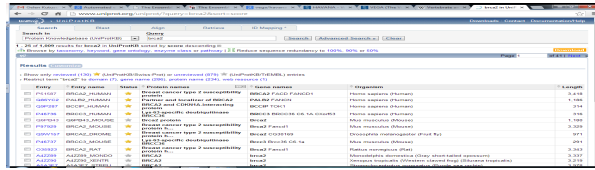
3.2.1 Protein Dizilimleri: UniProt

EBI'in sayfasından erişilebilen Uniprot, EBI, SIB (Swiss Institute of Bioinformatics) ve PIR (Protein Information Databases) topluluklarının Aralık 2003'de birleşmesi ile oluşmuştur. UniProt araştırmacılara protein dizilimi ve fonksiyonları ile ilgili kaliteli bilgi sunmayı amaçlamaktadır. UniProt üç alt veritabanı sunar.

UniProtKB: Bu veritabanında iki tür protein bilgisine erişilebilir.

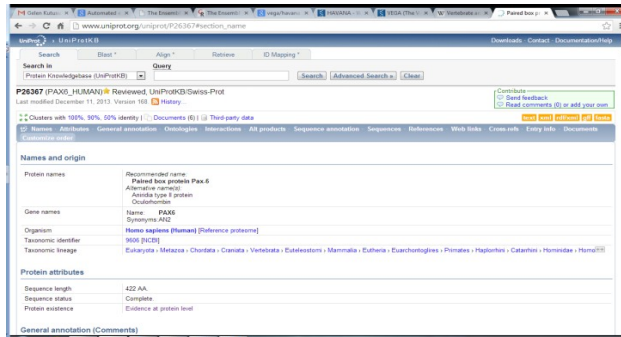
- Protein hakkında manüel olarak ek açıklamalar eklenen ve gözden geçirilmiş olan Swiss-Prot bölümü
- Protein hakkında ek açıklamaların otomatik olarak eklendiği ve gözden geçirilmemiş olan TrEMBL bölümü

Manüel olan açıklamalar, deney sonuçları ile desteklenmiş veya bilgisayar ile tahmin edilmiş veriler içerir. Bu kayıtlar uzman bir biyolog ekip tarafından devamlı güncellenir [4]. UniProtKB'de herhangi bir protein arandığında her iki türden kayıtlar da listelenir. Her protein için bir kimlik numarası, kayıt adı, durum bilgisi (status), protein isimleri, gen isimleri, organizma ve uzunluk bilgileri yer alır. Listede *Status* kısmında sarı yıldız bulunan proteinler Swiss-Prot bölümünden, gri yıldız bulunan proteinler ise TrEMBL bölümünden olanlardır.



Şekil : UniProtKB protein arama sonucu listesi

Listeden herhangi bir protein seçildiğinde, bu protein ile ilgili genel bilgiler, bilinen fonksiyonları, herhangi bir hastalıkla ilişkisi, başka proteinler ile etkileşimi, bu protein dizilimi hakkında ek açıklamalar (natural varyansları, ikicil yapısı vb.), protein diziliminin harflerle ifadesi ve dizilimin FASTA formatı görüntülenebilir. Bunların yanında, bu protein dizilimi hakkında yayınlanmış makaleler, farklı veritabanlarında bu proteine erişim bağlantıları, protein hakkındaki bu bilgilerin ilk yayınlana ve düzenlenme tarihleri ve protein ile ilgili farklı dokümanlara bağlantılar da görüntülenir.



Şekil : Örnek UniProtKB kayıtu

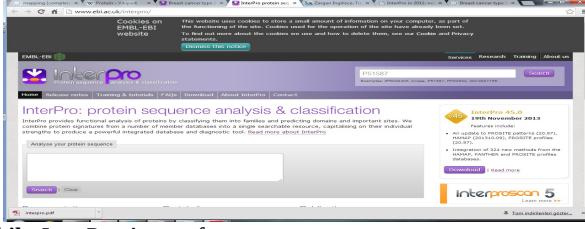
UniParc: Erişime açık olan temel protein veritabanlarındaki tüm protein dizilimlerini ek açıklama içermeden listeleyen veritabanıdır. Dizilimleri tamamen aynı olan proteinleri, farklı organizmalardan da olsa da tek bir kayıt olarak veritabanında tutar [20]. Yeni veya düzenlenmiş protein dizilimlerini günlük olarak günceller. Eski kayıtları da silmez. Böylelikle protein diziliminin zaman içindeki değişimi izlenebilir.

UniRef: Protein dizilimleri kümeleri içerir denebilir. Bir protein dizilimine benzeyen diğer protein dizilimleri aranırken, veritabanında yer alan fazla miktardaki dizilimler nedeniyle aramayı ve sonuçların yorumlanmasını zorlaştırır. Bu nedenle proteinler benzerliklerine göre kümelere ayrılır, böylelikle hem arama yapılacak veri seti boyutu azalır, hem de örneklem eğilimini (sampling bias) de azaltır [19]. UniProtKB ve UniParc veritabanlarındaki girdiler kümelendirilerek üç çeşit alt veritabanı oluşturulmuştur: UniRef100, UniRef90 ve UniRef50. UniRef100, aynı türden olmasa da %100 aynı olan dizilimleri ve dizi parçalarını tek bir UniRef girdisi olarak birleştirir. UniRef100 gereksiz veri içermeyen en kapsamlı protein dizilim veritabanıdır. UniRef90 ve UniRef50, sırasıyla %90 ve %50 benzer olan dizilimleri birleştirerek bir UniRef girdisi oluşturur [24]. Bir protein dizilimine benzeyen diğer protein dizilimleri aranırken, UniRef90 ve UniRef50 veritabanlarında daha hızlı arama gerçekleştirilir.

3.2.2 Protein Aileleri: InterPro

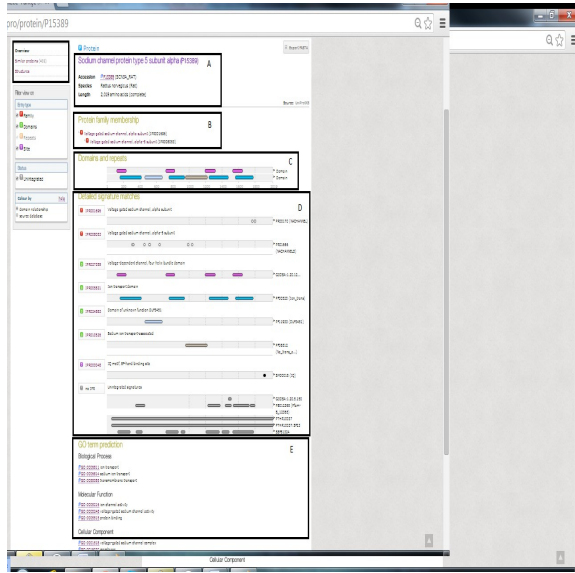
Avrupa tabanlı, EBI sitesinden erişilebilen InterPro; proteinlerin işlevlerini anlamak için onları belirli protein aileleri ile ilişkilendirmek ve protein içerisindeki işlevsel bölgelerin varlığını tahmin edebilmek için kullanılan bir veritabanıdır. Pfam, Smart, Prosite gibi protein aileleri ve protein bölgeleri (domain) ile ilgili veritabanlarını kullanır. Yeni bir dizilimin hangi bilinen protein ailesinden olduğunun anlaşılmasına yardımcı olur.

Şekil 15'de InterPro'nun ana sayfası görülmektedir. Arama kutusuna, UniProt'dan elde edilmiş bir protein kimlik numarası veya direkt FASTA formatında bir protein dizilimi girilerek arama yapılabilir.



Şekil : InterPro Anasayfa

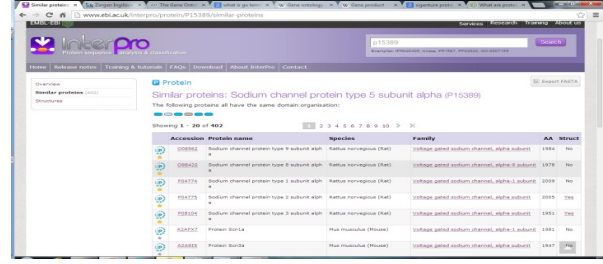
Aranan proteinin hakkındaki veriler üç bölümde sunulur: Genel açıklama, benzer proteinler ve yapı. Genel açıklama kısmı protein hakkında birçok veri sunar ve beş alt bölüme ayrılmıştır.



Şekil : InterPro Genel Açıklama Sayfası

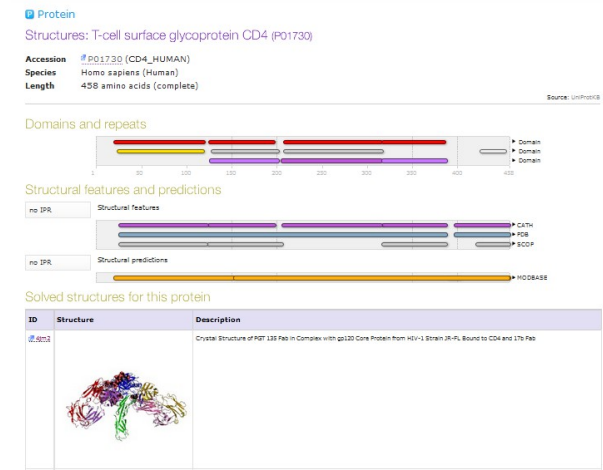
Şekil 16 'da görülen A bölümünde, proteinin UniProtKB veritabanındaki kimlik numarası, adı, hangi organizmadan olduğu ve uzunluğu görüntülenmektedir. B bölümünde, protein için protein ailesi tahminleri hiyerarşik bir yapıda listelenir. Üzerilerine tıklanarak o protein ailesi hakkındaki bilgilere erişilebilir. C bölümünde, proteinin fonksiyonel bölgeleri (domains) ve tekrar bölgeleri (repeats), fare ile üzerine gelindiğinde bu bölgeler hakkında detaylı bilgiler görüntülenir. D bölümünde, "protein signatures" denilen, proteinin hangi aileye ait olduğunu tahmin etmek için kullanılan modeller listelenir. E bölümünde protein için "Go Terms" olarak adlandırılan, proteinin özelliklerini ifade eden terim tahminleri listelenir.

Benzer Proteinler sayfasında, UniProt veritabanında bulunan aynı fonksiyonel bölge (domains) organizasyonuna sahip proteinler listelenir.



Şekil : InterPro Benzer Proteinler sayfası

Yapı bölümünde, protein ile ilgili deneysel olarak çözümlenmiş ve tahmin edilen tüm yapısal bilgiler görüntülenir. Proteinin 3D grafik görüntüsü var ise, PDB yapısal veritabanından çekilerek bu sayfada görüntülenebilir.



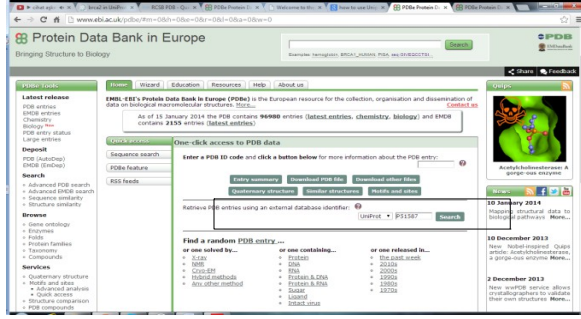
Şekil : InterPro Yapı Sayfası

2009 yılında, InterPro veritabanı üzerinde detaylı sorgular yapılabilmesi için BioMart analiz aracı eklenmiştir [7].

3.2.3 Yapısal Veritabanları: PDBe

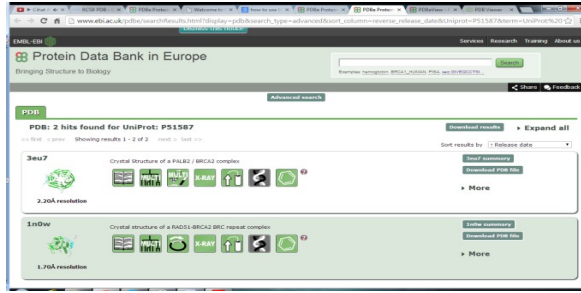
Avrupa tabanlı Protein DataBank in Europe (PDBe) ile Amerika tabanlı Protein DataBank (PDB) ortak çalışan kuruluşlar olup, protein, RNA ve DNA gibi moleküllerin 3D yapılarını arşivlerler [18]. Bu veritabanları, her yapı için, detaylı yapı analizleri, şematik diyagramlar ve farklı moleküllerle etkileşim bilgileri içerirler. PDB veritabanı 2000 yılında yaklaşık 13 bin yapıyı içerirken, 2013 yılı sonları itibariyle yaklaşık 97 bin yapıya ulaşmıştır [15].

UniProt protein veritabanında incelenen bir proteinin yapısı hakkında bilgi edinmek için, PDBe veritabanı UniProt'dan elde edilen kimlik numarası ile sorgulanabilir. Şekil 19'da gösterilen alandan, UniProt veritabanı seçilip aranan proteinin kimlik numarası girilerek arama yapılabilir.



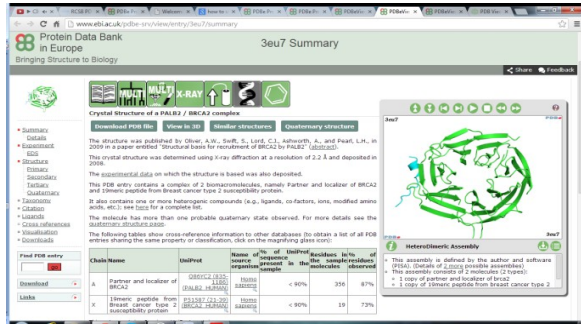
Şekil : PDBe Ana Sayfa

Arama sonucunda listelenen kayıtlar çeşitli simgeler ile etiketlenmiştir. Bu etiketler yardımıyla, kayıt ile ilgili var olan makalelerin, deneysel verilerin varlığı, hangi organizmadan, hangi yöntemle elde edildiğini, bu kayıttan protein, DNA/RNA içerip içermediği ifade edilmiştir.



Şekil .PDBe Arama Sonucu

Herhangi bir kayıttın sayfasına gidildiğinde daha detaylı bilgilere erişilebilir ve protein 3D olarak görüntülenebilir.



Şekil :PDBe Kayıt Sayfası

Bu bölümde, nükleotid veritabanları olarak Ensembl ve GenBank, protein veritabanları olarak da UniProt, InterPro ve PDBe incelenmiştir. Bu veritabanlarının genel içeriği, arayüzleri ve işlevleri ile ilgili açıklamalar yapılmıştır.

4. Biyoinformatik Analiz Araçları

Biyolojik veritabanları sayesinde nükleotidler ve proteinler hakkında yalın veya ek açıklamalar ile desteklenmiş kapsamlı verilere kolaylıkla ulaşılabilir. Fakat bu işlenmemiş verilerden anlamlı bilgiler çıkarabilmek zorlu bir iştir. İstenmeyen verileri filtrelemek, veriler arası bağlantılar kurmak, verilerin birbirleri ile benzerliklerini anlayabilmek ve daha detaylı araştırmalar yapabilmek için analiz araçları geliştirilmiştir.

Biyolojik veritabanları üzerinde detaylı sorgular oluşturabilmek amacıyla BioMart analiz aracı geliştirilmiştir. Bir veri madenciliği aracı olarak BioMart bu bölümde tanıtılacaktır.

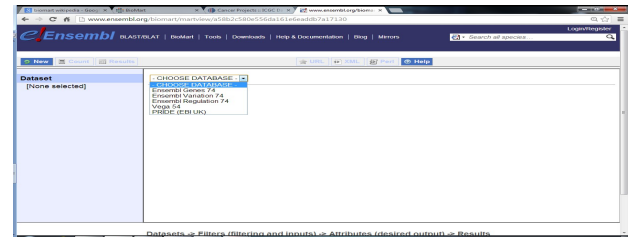
Biyoinformatikte dizi hizalama ve dizi karşılaştırma en çok kullanılan analiz yöntemlerindedir. DNA, RNA ve protein dizilerini düzenleyerek benzer bölgelerin tespit edilmesine dizi hizalama; var olan bir nükleotid dizisine benzer dizilerin tespit edilmesine dizi karşılaştırma denir. Dizi hizalama aracı olarak, CLUSTAL; dizi karşılaştırma aracı olarak ise BLAST bu bölümde tanıtılacaktır.

4.1 Veri Madenciliği: BioMart

Biomart biyolojik veritabanları üzerinde detaylı sorgular oluşturabilmeye olanak sağlayan, açık kaynak kodlu bir biyoinformatik analiz aracıdır. 4 kıtada, 11 ülkede 46 adet veritabanına sahiptir [2] .

BioMart ilk önce bir nükleotid veritabanı olan Ensembl veritabanı üzerinde çalışmak üzere kurulmuştur. Daha sonra bir çok veritabanı için geliştirilmiştir. 2009 yılında bir protein veritabanı olan InterPro üzerinde de detaylı sorgular oluşturabilmek amacıyla BioMart kullanılmaya başlanmıştır [7]. Bu bölümde Ensembl ve InterPro üzerinde çalışan BioMart aracı üzerinde durulacaktır.

Biomart'ın arayüzü oldukça basit tasarlanmıştır. Öncelikle sorgu oluşturulmak istenen veritabanı seçilir. BioMart'ın anasayfasında Ensembl'a ait olan Genes, Variations, Regulations, Vega ve EBI'nin PRIDE altveritabanları listelenir. Tablo 1'de bu veritabanlarının içerikleri hakkında kısa bilgiler verilmiştir.



Şekil : BioMart Veritabanı Seçim Sayfası

Veritabanı
Ensembl Genes 74

Veri tanımı

79 türden ek açıklamalar zenginleştirilmiş gen bilgileri, protein domainleri, farklı tür karşılaştırmaları (ortolog paralog) varyasyonlar, regülasyonlar, gen ontolojisi verileri içerir.

Ensembl Variation 74

21 tür için varyasyon verisi içerir.

Ensembl Regulation 74

İnsan ve Fare için regülasyon verisi içerir.

Vega 54

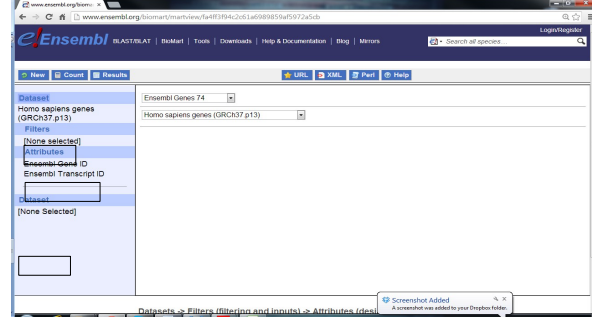
Manuel olarak yapılabildiği VEGA/Havana projesi sonuçları elde edilmiş insan, fare ve zebrafish gen verileri içerir.

PRIDE (EBI UK)

PRIDE veritabanından çekilen Proteomics verisi içerir.

Tablo : Ensembl BioMart'da Listelenen Veritabanları

İstenen veritabanı seçildikten sonra hangi tür ile ilgili arama yapmak isteniyor ise o verisi seçilir. Daha sonra "Filters" kısmından sorgu için kullanılacak filtreler belirlenir. "Attributes" kısmından ise sorgu sonuç tablosunda gözükmesi istenen özellikler seçilir.

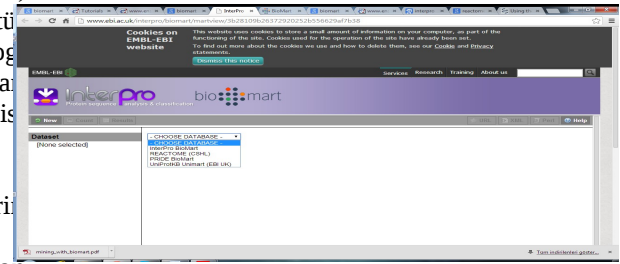


Şekil : BioMart Anasayfa

Belirtilen veriler ile başka bir veri setinden veriler birleştirilmek (çapraz sorgu) istenir ise, "Dataset" kısmında farklı bir veri seti eklenebilir. Yeni seçilen veri seti için de filtreler ve tabloda gözükmesi istenen özellikler belirlenebilir. Sorgu tablosunu ekranda görüntülemeye önce sorgu sonucunu elde edilen kayıt sayısı "Count" butonuna basılarak ekranda görüntülenebilir. "Results" butonu ile de sorgu sonucu ekranda görüntülenir. Yeni bir sorgu için "New" butonu kullanılır.

InterPro BioMart, Ensembl BioMart ile çok benzerdir. Anasayfadan öncelikle sorgu yapılmak

istenen veritabanı seçilir. InterPro BioMart, REACTOME, PRIDE BioMart ve UniProtKB Unimart



Şekil : InterPro Anasayfa

Tablo 2'de bu veritabanlarının içerikleri hakkında kısa bilgiler verilmiştir.

Veritabanı
InterPro BioMart

Veri tanımı

Protein aileleri hakkında veriler içerir.

REACTOM (CSHL)

Biyolojik yolak (pathway) verisi içerir.

PRIDE BioMart

PRIDE veritabanından çekilen Proteomics verisi içerir.

UniProtKB Unimart

UniProt veritabanında protein verisi içerir.

Tablo : InterPro BioMart'ta Listelenen Veritabanları

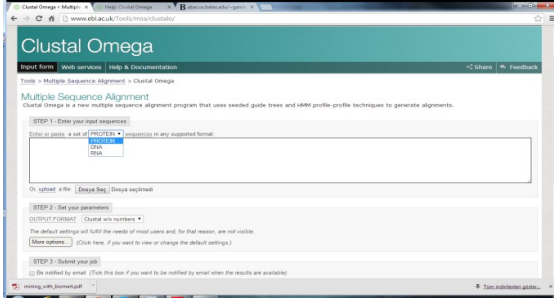
İstenilen veritabanı seçildikten sonra yukarıda anlatıldığı gibi filtreler ve diğer özellikler aynı şekilde belirlenebilir.

BioMart, biyolojik veritabanları üzerinde kapsamlı sorgular oluşturulabilmek ve detaylı araştırmalar yapılabilmek için değerli bir veri madenciliği aracıdır.

4.2 Dizi Hizalama: CLUSTAL

Biyoinformatikte dizi hizalama, DNA, RNA ve protein dizilerini düzenleyerek benzer bölgelerin tespit edilmesidir. Bu bölgelerin benzer olması, diziler arasında işlevsel, yapısal veya evrimsel bir ilişki olduğu anlamına gelir. İki nükleotid dizisi hizalanabileceği gibi birden fazla dizi de hizalanabilir. Belirli iki veya daha fazla diziyi hizalamak için EMBOSS (European Molecular Biology Open Software Suite) adlı açık kaynak kodlu bir program paketi mevcuttur. Bu paket geliştirilerek Cluster Omega adlı bir web servisi

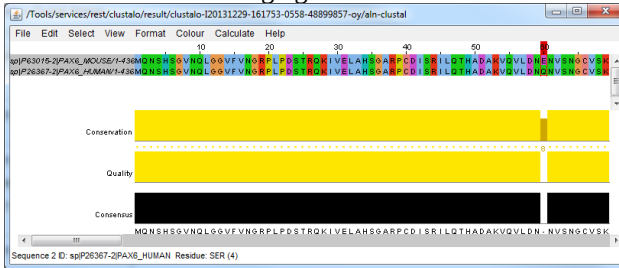
EBI altında başlatılmıştır. İnternet üzerinden erişilebilen bu site yardımıyla, iki veya daha fazla dizilim belirli bir formatta yüklenerek dizi hizalaması yapılabilir.



Şekil : Clustal Omega Anasayfa

Clustal ile protein, DNA veya RNA dizilimleri hizalanabilir. Öncelikle hizalama yapılacak dizi türü seçilir. Daha sonra kaç tane dizilim hizalanmak isteniyor ise, o dizilimler girdi kutusuna yapıştırılır. Dizilimler daha önceden belirlenmiş formatta olmalıdırlar. Clustal FASTA formatını destekler. Hizalama yapılacak diziler sisteme girildikten sonra, hizalama parametreleri değiştirilebilir. Varsayılan parametreler, çoğu kullanıcının isteyeceği tarzda düzenlenmiştir. Son olarak "Submit" butonu ile hizalama gerçekleştirilir.

Hizalama sona erdiğinde, renklerle desteklenmiş bir hizalama görüntüsü elde edilir. Hizalama sonucu üzerinde detaylı inceleme yapabilmek için "Result Summary" kısmından JalView adı verilen bir java programı ile de görüntülenebilir. Örneğin, Şekil 26'da insan ve faredeki pax6 proteinin dizi hizalaması JalView görünümünde görüntülenmektedir. Şekil dikkatli incelendiğinde 60 ile işaretlenen aminoasidin iki dizilimde farklı olduğu görüntülenmektedir.



Şekil : Clustal Hizalama Sonucu

Bu örnekte yalnız iki dizilim hizalanmıştır. Clustal ile yalnız iki değil daha fazla dizilim aynı anda hizalanabilir.

4.3 Dizi Karşılaştırma: BLAST

BLAST (Basic Local Alignment Search Tool) çok yaygın olarak kullanılan bir biyoinformatik araçtır. Var olan bir nükleotid veya protein dizilimine benzer diğer dizilimleri bulmak için kullanılır. Bir protein dizilimindeki aminoasit farklılıklarını veya bir DNA dizilimindeki nükleotid farklılıklarını karşılaştırmak için geliştirilmiş bir algoritma kullanır.

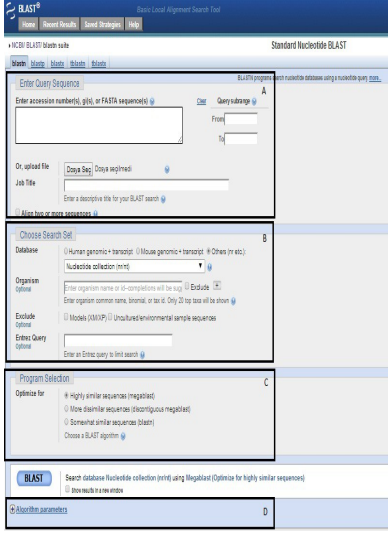
Bir dizilim, veritabanındaki diğer dizilimler ile ikili olarak karşılaştırılır ve iki dizilimin benzerlikleri istatistiksel olarak hesaplanır. BLAST sorgu dizilimini belli uzunluktaki kelimelere ayırır ve bu kelimeleri tüm veritabanında bulunan dizilimler içinde arar. Sorgu kelimesi ile aynı olan bölgeler üzerinde hizalama genişletilir ve bu genişletme belli bir puanlama sistemi ile puanlandırılır. Bu puanlama sistemi sonucunda sorgu dizilimine en benzer dizilimler listelenir [3].

BLAST ile yalnızca bir tür üzerinde arama yapılabilir. Bunun yanında farklı sorgu türleri ile farklı veritabanlarında da arama yapılabilir. Bu şekilde BLAST beş farklı arama olanağı sağlar. Aşağıdaki tablo BLAST ile yapılacak farklı aramaları listeler.

BLAST	Açıklama
blastn	Nükleotid sorgu dizilimi ile nükleotid veritabanı içinde arama yapılır.
blastp	Protein sorgu dizilimi ile protein veritabanı içinde arama yapılır.
blastx	İfadelemiş nükleotid (<i>translated nucleotide</i>) sorgu dizilimi ile protein veritabanı içinde arama yapılır.
tblastx	Protein sorgu dizilimi ile İfadelemiş nükleotid veritabanında arama yapılır.
tblastx	İfadelemiş nükleotid sorgu dizilimi ile İfadelemiş nükleotid veritabanında arama yapılır.

Tablo : BLAST Arama Seçenekleri

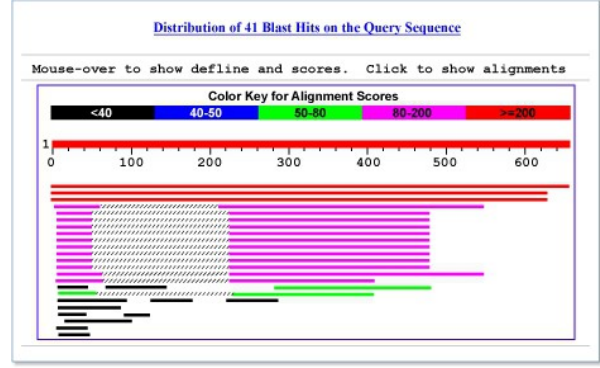
Farklı BLAST seçenekleri için arama arayüzleri birbirine çok benzer olup, blastn üzerinden açıklamalara devam edilmiştir. Blastn seçildikten sonra gelen arayüz dört kısımdan oluşmaktadır: Sorgu dizilimi girişi, arama seti seçimi, hizalama algoritması seçimi ve algoritma parametreleri seçimi.



Şekil : BLAST Anasayfa

Şekil 27’de gösterilen A bölümünde sorgu dizilimi FASTA formatında veya NCBI kimlik numarası ile girilebilir. B bölümünde arama yapılacak veri seti seçilir. İnsan geni+transkript, fare geni+transkript ve diğer seçeneklerden birisi seçilebilir. Gen+transkript veri setleri sadece NCBI’da yer alan dizilimleri içerir ve gen ile birlikte mRNA dizilimlerini de içerir. Bu veri setleri insan ve fare dizilimi aramalarını kolaylaştırır. İnsan ve fare dışındaki organizmalar için farklı veri setleri açılır menüden seçilebilir ve ya türün adı “Organism” alanında belirtilebilir. C bölümünde arama ve hizalama yapma algoritması seçilebilir. Algoritma parametreleri de “Algoritma parametreleri” kısmından değiştirilebilir. *Megablast* varsayılan algoritma olarak gelir ve büyük kelime uzunluğu kullanarak (28) %95’lik benzerlik için değerleri optimize eder. *Discontiguous megablast* ve *blastn* daha kısa kelime uzunluğu kullanarak (11) %85’lik benzerlik için değerleri optimize eder [12].

BLAST çalıştırdıktan sonra Özet, Grafiksiz Özet, Açıklamalar ve Hizalamalar başlıkları altında dört bölümden oluşan bir rapor sayfası açılır. Özet kısmı yapılan iş ile ilgili kısa açıklamaların yanında farklı raporlama formatlarına linkler içerir. Grafiksiz Özet, sorgu dizisi ile veritabanında bulunan benzer dizilerin benzerliklerini grafiksiz olarak ifade eder.



Şekil : BLAST Grafiksiz Özet

Şekil 28’de görülen grafikte, sorgu dizilimi üstteki numaralandırılmış kırmızı çizgi ile ifade edilir. Bunun altındakiler veri tabanında bulunan dizilimleri ifade eder. Sorgu dizilimine en benzer dizilim kırmızı çizgiye en yakın olanıdır, en yüksek puanlı olan dizilimdir. Daha alt seviyelerde olan çizgiler puanı daha az olan hizalamaları gösterir. Çizgilere tıklayarak veritabanında yer alan o dizilim ile sorgu diziliminin hizalaması görüntülenebilir [14].

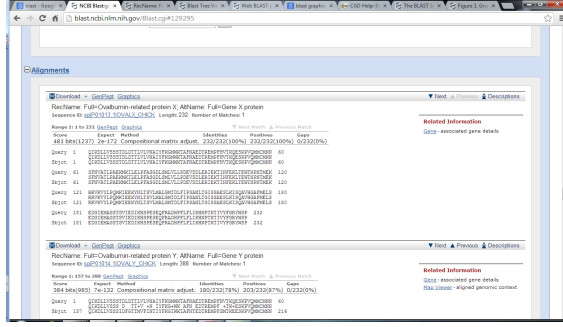
Açıklama bölümünde, her dizilim için puanlama bilgisi, dizilimin kromozom üzerindeki yeri ve veritabanındaki belirleyici kimlik numarası yer alır. Kimlik numarasına tıkladığında dizilimin detaylarına ulaşılır.

The image shows the BLAST Results page. The table lists sequences producing significant alignments with E-value <= 10^-10. The columns are: Description, Max Score, Total Query, E-value, Ident, Accession, and Hit. The table contains 11 rows of data.

Description	Max Score	Total Query	E-value	Ident	Accession	Hit
EMBL:AB011011.1	481	100%	2e-172	100%	EMBL:AB011011.1	95
EMBL:AB011011.1	388	100%	7e-102	99%	EMBL:AB011011.1	95
EMBL:AB011011.1	315	100%	6e-105	65%	EMBL:AB011011.1	95
EMBL:AB011011.1	304	100%	7e-97	93%	EMBL:AB011011.1	95
EMBL:AB011011.1	295	100%	6e-98	64%	EMBL:AB011011.1	95
EMBL:AB011011.1	294	100%	7e-97	93%	EMBL:AB011011.1	95
EMBL:AB011011.1	293	100%	2e-96	63%	EMBL:AB011011.1	95
EMBL:AB011011.1	213	100%	3e-65	45%	EMBL:AB011011.1	95
EMBL:AB011011.1	211	100%	2e-64	44%	EMBL:AB011011.1	95
EMBL:AB011011.1	209	100%	3e-64	44%	EMBL:AB011011.1	95
EMBL:AB011011.1	206	100%	1e-62	43%	EMBL:AB011011.1	95
EMBL:AB011011.1	205	100%	6e-62	43%	EMBL:AB011011.1	95
EMBL:AB011011.1	204	100%	5e-62	42%	EMBL:AB011011.1	95

Şekil : BLAST Açıklama

Hizalamalar kısmında ise sorgu dizilimi ile veritabanında bulunan dizilimlerin hizalamaları görüntülenir.



Şekil : BLAST Hizalamalar

5. Sonuç

Bu çalışmada yaygın olarak kullanılan biyolojik veritabanları ve analiz araçlarını tanıtmıştır. Biyolojik veritabanları sayesinde nükleotidler ve proteinler hakkında yalın veya ek açıklamalar ile desteklenmiş kapsamlı verilere kolaylıkla ulaşılır iken, bu verilerden anlamlı bilgiler çıkarabilmek için Biomart, Blast ve Clustal gibi analiz araçları kullanılmaktadır. Bu veritabanlarının yönetilmesi ve analiz araçlarının geliştirilmesi biyoinformatik alanının temel araştırma konularındandır.

Bilgisayar teknolojileri, internet teknolojileri ve bilişsel yöntemler ile moleküler biyoloji ve genetik çalışmalarının işbirliği, sağlık ve tıp alanında gelişmeleri doğrudan etkilemektedir. Avrupa, Asya ve Amerika'daki kuruluşlar bu konuda disiplinler arası ortak çalışmalar yapmakta olup, ülkemizde de disiplinler arası ortak çalışmaların artırılması önem kazanmalıdır.

6. Kaynaklar

- [1] Bal, S. H., & Budak, F. (2013). Genomik, Proteomik Kavramlarına Genel Bakış ve Uygulama Alanları. *Uludağ Üniversitesi Tıp Fakültesi Dergisi* , 65-69.
- [2] Biomart. (2013, 12 29). *BioMart*. 12 29, 2013 tarihinde BioMart: <http://www.biomart.org/> adresinden alındı
- [3] Clark, F. (2013, 6 1). *How Blast work*. 12 29, 2013 tarihinde Blast Central: http://www.clarkfrancis.com/blast/Blast_what_and_how.html adresinden alındı

[4] Consortium, T. U. (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research* .

[5] Curwen, V. (2004). The Ensembl Automatic Gene Annotation System. *Genome Research* .

[6] Dennis A. Benson, E. W. (2010). GenBank. *Nucleic Acids Research* .

[7] EMBL Outstation European Bioinformatics Institute. (2012). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research* .

[8] Galperin, M., & Fernandez-Suarez, X. (2011). The 2012 Nucleic Acids Research Database Issues and the Online Molecular Biology Database Collection. *Nucleic Acids Research* .

[9] *GenBank*. (2013, 7). 7 2013 tarihinde GenBank: <http://www.ncbi.nlm.nih.gov/genbank/statistics> adresinden alındı

[10] Gümüsel, F. (2002). *BİYOTEKNOLOJİ, GENETİK VE SAĞLIK SEKTÖRÜ*.

[11] Luscombe, N., Greenbaum, D., & Gerstein, M. (2001). *What is bioinformatics? An introduction and overview*. New Haven, USA: Yearbook of Medical Informatics.

[12] Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research* .

[13] Mümin Polat, A. G. (2009). Multidisipliner yeni bir bilim dalı: biyoinformatik ve tıpta uygulamaları. *.D.Ü. Tıp Fak. Derg* , 41-50.

[14] NCBI. (2013). *The NCBI Handbook*. 12 29, 2013 tarihinde NCBI: <http://www.ncbi.nlm.nih.gov/books/NBK21097/> adresinden alındı

[15] Protein Data Bank. (2013, 12 29). *Protein Data Bank*. 12 29, 2013 tarihinde Protein Data Bank: <http://www.rcsb.org/pdb/protein/P26367> adresinden alındı

[16] *Reseach Genome Institute*. (2013, 7). 7 2013 tarihinde Reseach Genome Institute: <http://www.genome.gov/25020001> adresinden alındı

[17] Rhoda J. Kinsella, P. F. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* .

[18] S. Velankar, G. J. (2013). PDBe: Protein Data Bank in Europe. *Nucleic Acid Research* .

[19] Suzek, B. E. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* .

[20] The UniProt Consortium. (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research* .

[21] Tübitak. (2004). *BİYOTEKNOLOJİ ve GEN TEKNOLOJİLERİ STRATEJİSİ*.

[22] Wikipedia. *Intron*. 1 16, 2014 tarihinde Wikipedia: <http://tr.wikipedia.org/wiki/%C4%B0ntron> adresinden alındı

[23] Wikipedia. (2013, 11 18). *List_of_biological_databases*. 12 27, 2013 tarihinde Wikipedia: http://en.wikipedia.org/wiki/List_of_biological_databases adresinden alındı

[24] Wikipedia. (2013, 12 3). *UniProt*. 12 27, 2013 tarihinde Wikipedia: <http://en.wikipedia.org/wiki/UniProt> adresinden alındı