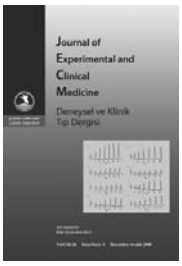
Journal of Experimental and Clinical Medicine
Deneysel ve Klinik Tıp Dergisi



Teknik Not*/Technical Note*

Solucan: Biyolojik veri madencisi ve kişisel ilişkisel veri tabanı yönetim sistemi

Worm: Biological data miner and personal relational database management system

M. Kamil Turan[a]*, Hasan Bağcı[a], Sedat Doğan[b]

[a] *Ondokuz Mayıs Üniversitesi, Tıp Fakültesi, Tıbbi Biyoloji Anabilim Dalı, Samsun*
[b] *Ondokuz Mayıs Üniversitesi, Mühendislik Fakültesi, Harita Mühendisliği Bölümü, Samsun*

MAKALE BİLGİLERİ

Makale Geçmişi:
Geliş 12 / 02 / 2010
Kabul 06 / 04 / 2010

*** Yazışma Adresi:**
M. Kamil Turan
Ondokuz Mayıs Üniversitesi,
Tıp Fakültesi,
Tıbbi Biyoloji Anabilim Dalı,
Kurupelit, Samsun
e-posta : mkamilturan@gmail.com

Anahtar Kelimeler:
Biyoinformatik
Veri Madenciliği
Genetik Veri Tabanları
Düzenli İfadeler

Key Words :
Bioinformatics
Data Mining
Genetic Databases
Regular Expressions

ÖZET

Genetik bilginin her gün değişen ve gelişen doğası bugün önemli bir soruna neden olmuştur. Aradığınız bilgi internette neredедir? O bilgiye nasıl ulaşabilirsiniz? Ulaştıktan sonra nasıl depolayabilirsiniz? Depolanan bilgileri nasıl güncel tutabilirsiniz? Şüphesiz bu soruların cevabını vermek genetik bilgi uzayının devasa yapısı düşünüldüğünde oldukça zor olacaktır. Bizim bu konudaki sorulara kısmi de olsa çözüm önerimiz Solucan ismini verdiğimiz yeni bir yazılımdır. Solucan, araştırmacının kendi istek ve öncelikleri doğrultusunda hazırladığı ilişkisel veri modelini, sizin istediğiniz ham veri kaynaklarından, sizin istediğiniz kurallar dizesini işletmek suretiyle bilgi grupları oluşturup otomatik olarak doldurmanızı sağlar. Ham veri kaynaklarının ve ilişkisel veritabanının sabit olmaması, bilgi elde etmek kurallarının da kullanıcı tarafından belirleniyor olması Solucan'ı etkin bir araç haline sokmuştur. Toplam 32004 gen için, ham veri kaynağımızdan, erişim numaralarını toparlayıp istediğimiz şekilde ilişkilendirmesi Solucan'ın yaklaşık 18 dakika 46 saniyesini almıştır.
J. Exp. Clin. Med., 2009; 26:180-185

ABSTRACT

Today, ever growing and changing nature of the genetic information has caused a serious problem. Where is the information, you are looking for, in the internet? How could you get that information? After getting it how would you store it? How would you keep that stored information up-to-date? In view of the monstrous structure of the genetic information space, there is no doubt that answering those questions will be very difficult. Our solution, even if it is a partial one, to the questions related to these topics is a new software that we called as the Worm (Solucan in Turkish). The Worm, let researchers fill the database model they prepare according to their desires and priorities, by running in line with series of their principles, forming information groups from row databases of their interests. Dynamic nature of row data sources and relational databases and the fact that the rules of information gathering are determined by the user marks the Worm as an effective tool. For a total of 32.004 genes, it took the Worm approximately 18 minutes and 46 seconds to gather from row data sources the accession numbers for these genes and associate them relationally as we wish.
J. Exp. Clin. Med., 2009; 26:180-185

1. Giriş

Genetik bilginin her gün değişen ve gelişen doğası bugün önemli bir soruna neden olmuştur. Aradığınız bilgi internette neredедir? O bilgiye nasıl ulaşabilirsiniz? Ulaştıktan sonra nasıl depolayabilirsiniz? Depolanan bilgileri nasıl güncel tutabilirsiniz? Şüphesiz bu soruların cevabını vermek genetik bilgi uzayının devasa yapısı düşünüldüğünde oldukça zor olacaktır. Varolan bilgi birikiminin veri tabanları şeklinde organize olması, bilgiyi elde etme, hazırlama ve kullanıcıya sunma çeşitlilikleri de hesaba katıldığında problemlerin arttığını ve daha da karmaşıklaştığını görmek mümkündür. Bu bilgi birikimine tam olarak hâkim olmak mümkün gibi görünmemektedir. Fakat bilginin etkili bir şekilde kullanılmasına imkân tanıyan yardımcı sistemlerin varlığına ihtiyaç artmaktadır. Araştırmacıların aynı konuyu çalışırken dahi önceliklerinin ve ihtiyaçlarının farklı olması benzer vasıftaki verilere ulaşımda sabit bir yol haritası oluşturulmasını zorlaştırmaktadır. Araştırmaların multidisipliner yapısı gerek duyduğu bilgiyi de multidisipliner kılmakta ve ulaşmayı da bir o kadar zorlaştır-

maktadır. İhtiyaç duyulan bilginin elde var olan verilerden çıkartılıp karşılaştırmaların, hesaplamaların uygulanabilmesi, yeni kazanımlar için depolanabilir kılınması ilişkisel veri modellerine olan ihtiyacımızı arttırmıştır. Verilerimiz üzerinde her an değişiklik yapabilmek, yenilerini eklemek, yeni ilişkiler tanımlamak, hipotezlerimizi test etmek, benzeri kazanımları kendi veri bloklarımıza entegre etmek için hem yerel hem de dağıtık çalışabilen kişisel sistemlerin, yerel ya da dağıtık proje gruplarının oluşturulması artık gerekli hale gelmiştir. Solucan, biyolojik veri madencisi ve kişisel ilişkisel veri tabanı yönetim sistemi uygulaması olarak bahsedilen sorulara kısmi yanıtlar vermekte kanımızca uygun bir çözüm yoludur. Solucan kısaca üç ana varlık ve bu varlıkların arasında tanımlanan ilişkilerin yürütülmesi ile görevini yerine getirir. Bu varlıklar Web (World Wide Web, W3) varlığı, düzenli ifade varlığı, SQL (Structured Query Language, yapısal sorgu dili) bağlantı varlığı ve bunların grup bağlantılarıdır. Öncelikle kullanıcının bağlantı kurduğu web sunucusundan istenilen veri indirilir. Daha sonra düzenli ifade varlığı kullanılarak veri bilgi grupları haline dönüştürülür. Bağlantı grupları kullanılarak bu bilgi grupları bir veri tabanına SQL bağlantı varlığı üzerinden kaydedilir. Veri tabanı yönetim sistemi sayesinde kullanıcının istekleri, öncelikleri ve ihtiyaçları doğrultusunda oluşturduğu ilişkisel veri modeli sorgulanabilir kılınmıştır.

2. Genetik bilgi uzayı

Genetik bilgi uzayı araştırmalar sonucunda elde edilmiş verilerin veri tabanları şeklinde organize edilip araştırmacılarının kullanımına sunulması neticesinde oluşmuş ve büyümüştür. 1993 yılında yaklaşık 24 veri tabanı tanımlı iken; 1995 yılında bu rakam 179 veri tabanına ulaşmıştır. 2010 yılı için ise bu rakam 58 yeni tanımlanmış ve 73 güncellenmiş veri tabanı ile 1230 olarak açıklanmıştır (Cochrane ve Galperin, 2010). Bu veri tabanları şüphesiz pek çok bilgiye yer vermektedir. En sık gezilen ve en sık yararlanılan veri tabanları şunlardır: GenBank, EMBL ve dbSNP (Baxevanis, 2006). Ulusal Biyoteknoloji Merkezi (NCBI), Ulusal Tıp Kütüphanesi (NLM) ve Ulusal Sağlık Enstitüsü (NIH) tarafından geliştirilen GenBank araştırmacıların en sık tercih ettikleri veri tabanıdır. GenBank kısaca, kapsamlı herkese açık kullanıma sahip, nükleotid veri bankasıdır. GenBank ayrıca bibliyografik biyolojik notları da kullanıcısına sunar. Bunların yanında bu veri tabanından GSS (genome survey sequence), EST (expressed sequence tags) ve WGS (whole genome shotgun sequencing) gibi pek çok bilgiye de referansları ile ulaşmak mümkündür (Benson ve ark., 2008). Sadece bu üç veri tabanı düşüldüğünde bile rakamlar inanılması zor seviyelere çıkmaktadır. GenBank 3 Şubat 2009'da 99.116.431.942 adet baz çiftine ve 98.868.465 adet diziye ev sahipliği yapmaktaydı (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html). 24 Ocak 2010 sabahında ise EMBL 273.873.018.011 adet nükleotid ve 175.210.541 adet veri tabanı girişine sahipti (http://www.ebi.ac.uk/embl/Services/DBStats/). GenBank altında yine oldukça sık ziyaret edilen diğer bir veri tabanı olan dbSNP ise 130 farklı organizma hakkında polimorfizm verilerini bize sunmaktadır. dbSNP'de var olan 130 organizmadan birisi olan insan hakkında 24 Ocak 2010 itibarı ile toplam 12.878.918 giriş bulunmaktadır (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi). Kasım 2009'da PubMed 73.094.038 adet interaktif aramaya ve 93.022.771 adet görüntülemeye ev sahipliği yapmıştır (http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmeddata.html). Medline ise 1966 'dan bu yanan 10 milyon atıfa, sadece NLM üzerinden yürütülen 120 milyondan fazla aramaya sahiptir. Medline veri tabanına her yıl ortalama 400.000 adet yeni atıf eklenmektedir (http://www.nlm.nih.gov/bsd/history/tsld024.htm). Görüldüğü gibi bu rakamlar her saniye değişmektedir. Nükleotid dizileme makinelerinin artması, teknolojilerinin her geçen gün değişmesi ve iyileşmesi, biyoinformatik yazılımlar ve bunların güncel teknoloji ile buluşması, yürütüle gelen çalışmalar ve sonuçları gibi pek çok etken nedeniyle genetik bilgi uzayı büyümesine her geçen gün daha büyük bir ivme ile devam edecektir (Cochrane ve ark., 2009). Genetik veri oldukça hızlı bir şekilde artmakta ve her gün veri uzayına yenileri eklenip, onaylamalar yapılıp, düzeltmeler ve güncellemeler ile desteklenmektedir. Bugün artık biyolojik çeşitliliğe, genetik veri uzayı da; yaşayan canlı bir sistem olarak girmeyi hak etmiştir. Bu bağlamda genetik veri uzayı hakkında yeni bir genom projesi başlatılması gerekliliği karşımıza çıkmaktadır.

3. Veri tabanları ve önemli bazı sorunlar

Birkaç on yıl önce bilim adamları biyolojik bilgi birikimini tek merkezden yönetmek ve araştırmacılara uzun dönemli veri olarak sunmak amacıyla çalışmalara başlamıştı. Erken dönemlerinde bu çalışmada kelime işlemciler ve hesap tabloları kullanılarak bu işlem yapılmaya çalışılıyordu. Bu çabalar sınırlı da olsa verilerin depolanmasına ve araştırmacılar arasında değiş tokuşuna müsaade ediyordu. Fakat bilginin inanılmaz artışı karşısında bu yöntemler, verimli ve etkili depolama, paylaşma ve yararlanmak amacıyla sonra kullanma gibi ihtiyaçlara cevap veremez oldu. Bu nedenle web tabanlı daha karmaşık organizasyona sahip yönetim sistemlerinin geliştirilmesi gerekti. Web tabanlı olarak verileri sunmak oldukça etkilidir. Kitlelerin rahat bir şekilde verilere ulaşmasına ve rahat bir şekilde veri girmesine izin verir. Araştırmacılar ilgi duydukları konular hakkında detaylı aramalar ile istediklerine ulaşabildiler. Fakat çoğunlukla yapılan sorguların sonuçları aşırı derecede veri içermekteydi. Bugün

için araştırmacılar sorgulardan dönen büyük veriyi sıklıkla tek tek inceleyip ilgi duydukları kısmı kendileri seçmek durumundadırlar (Philippi ve Kohler, 2006). Bazı veri tabanlarının sorgu sonuçları düz metin dosyası (flat file fromat) şeklinde kullanıcına ulaşmaktadır. Bu tip verileri organize etmek daha da zor bir işlem gerektirir (Ellis ve Attwood, 2001). Bir diğer önemli durum da artan biyolojik bilginin entegrasyonunda yaşanmaktadır (Philippi ve Kohler, 2006). Farklı veri tabanlarında aynı konunun farklı yönlerine ait bilgiler bulunuyor olması bunlar arasındaki ilişkinin zayıf ya da hiç olmaması olası yeni kazanımları engellemektedir. GenBank, EMBL, TrMBL\UniProt gibi veri tabanları bu konuda oldukça başarılı veri entegrasyon örneği sunmaktadır. Çözüm bekleyen diğer bir sorun da herkese açık kullanıma sahip veri tabanlarının alt yapılarını güncel ve etkili tutabilmek için ihtiyaç duydukları fonun yaratılmasıdır (Ellis ve Attwood, 2001). Benzer verilerin farklı veri tabanlarında farklı şekillerde ifade ediliyor olması, depolanan biyolojik verinin karmaşık yapısı ve sözel verinin fazlalığı, girişlerde doğal dilin kullanılıyor olması ve bundan kaynaklanan diğer sorunlar, veri tabanları üzerinde kullanıcıların sınırlı yetkiler dahilinde kullanım hakkına sahip olması ve ilişkisel yapılarını bu nedenle etkili ve akıcı bir şekilde kullanamıyor olmaları sayılabilecek diğer önemli sorunlardır.

4. Yapılan çalışmalar

Genetik veri uzayının yerel çevrede organizasyon ve optimizasyonu hakkında pek çok çalışma yapılmıştır. Bunlardan bazıları şu şekilde sayılabilir: GeneRecords, GeneNotes, SnpHunter, Atlas, GeneKeyDB. GeneRecords, GenBank veritabanının düz metin dosyalarının ayrıştırılıp kişisel bilgisayarınızdaki ilişkisel veri tabanında saklanmasını sağlayan bir yazılımdır. Ayrıca elde edilen diziler üzerinde analiz yapma imkanını da sağlar (D'Addabbo ve ark., 2004). GeneNotes, farklı veri tabanlarından özellikle genlerin farklı formatlar (text, imaj, PDF dosyası vb.) şeklinde saklanmış özelliklerini yönetmeyi sağlayan ilk örnektir (Hong ve Wong, 2005). SnpHunter ise, seçtiğiniz bir gendeki tüm SNP noktalarını bilgisayarınıza indirip, filtre edip üzerinde çalışabileceğiniz bir programdır. SnpHunter ayrıca indirdiği SNP noktalarını görsel olarak da kullanıcısına sunabilmektedir (Wang ve ark., 2005). Atlas, kullanıcısına yeniden ilişkilendirilmiş biyolojik veri deposunun yerel depolama biriminde entegre edebilmesinin önünü açan önemli bir biyoinformatik yazılımıdır. Atlas biyolojik veri deposu olarak oldukça geniş ham veriyi ,örneğin diziler, moleküler etkileşimler, homoloji, fonksiyonel ve biyolojik ontoloji gibi, işleyebilmektedir. Atlas, bu yönü dışında, biyoinformatiğin önemli bir amacı olan farklı kaynaklardan gelen verilerin entegrasyonu konusunda da çözüm sunmaktadır (Shah ve ark., 2005). GeneKeyDB ise Atlas gibi farkı veri tabanlarından alınan ham veriyi veri madenciliği kuralları dahilinde işleyerek kullanıcısına sunan bir diğer yazılımdır (Kirov, 2005).

Solucan'ın bahsi geçen bu yazılımlardan bazı farkları bulunmaktadır. Bu farkların ilki Türkçe bir yazılım olmasıdır. İkinci önemli farkı ise, sabit bir veri tabanı değil de istekleriniz doğrultusunda sizin hazırladığınız bir veri tabanı kullanmasıdır. İhtiyaçlarınıza göre şekillenebilir. Ayrıca Solucan sadece genetik alanında değil veri madenciliğin her aşamasında rahatça kullanılabilir (örneğin arama sonuçları için herhangi bir arama motoruna yönlendirilebilir ya da hikâye, roman gibi geniş veri kaynaklarında istenilenleri arayabilir ya da imla denetçisi olarak kullanılabilir. Kullanım alanı ve amacı tamamen sizin elinizdedir). Veri kaynağınız, formatının değişmesi durumunda sadece düzenli ifadesi değiştirilerek istenen sonuç yeni formattan çıkartılabilir.

5. Solucan

Solucan kısaca kullanıcının tanımladığı direktif (direktif, kullanıcının isteklerine göre belli görevi tamamlamak için Solucan'a sunulan kurallar dizisi) varlıklarını (direktifleri oluşturan alt kural parçaları, verinin alınma yeri, şekli, v.b. gibi) kullanarak hedefteki veriyi, yerel çevredeki ilişkisel veri modeline kaydırarak bilgi haline dönüştüren ve bu model üzerinde sorgulama yapma imkanı tanıyan bir uygulamadır. Veri tabanı ve direktifler tamamen kullanıcının kendi istek ve öncelikleri doğrultusunda ayarladığı bileşenlerdir. Solucan'ı oldukça esnek kılan ve onu sabit ham veri topluluğundan bünyesindeki sabit veri tabanına bilgi çıkartıp yazan katı kurallı bir yazılım olmaktan uzak tutan kural da budur. Web varlığı sayesinde kullanıcısının istediği herhangi bir kaynaktan (ki bu kaynak Web sunucusundan bir sayfa, FTP sunucusundan bir dosya ya da yerel saklama biriminden herhangi okunabilir bir kaynak olabilir) ham veriyi temin eder. Bu ham veri üzerinde Solucan düzenli ifade varlıklarını kullanarak istenilen bilgi gruplarını oluşturabilir. Bu bilgi gruplarını aynen saklayabilir ya da üzerinde sayma, frekans hesaplama gibi sayısal işlemleri uygulayabilir. Daha sonra SQL varlıkları üzerinden kullanıcının tanımladığı veri tabanına bunları kaydeder. Solucan'ın çalıştırılmasından önce tamamlanması gereken bazı altyapı çalışmaları vardır.

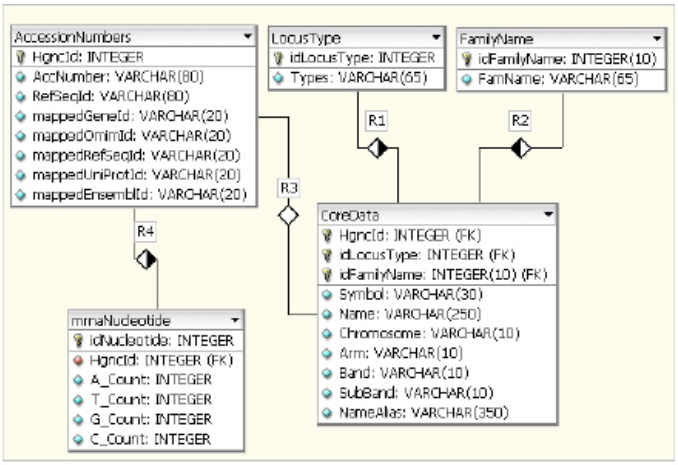
6. Solucan altyapı çalışmaları

Solucan, kullanıcının tanımladığı direktifleri çalıştırarak veriyi bilgi halinde organize eden bir sistemdir. Bu nedenle direktiflerden önce Solucan'ın verileri bilgi şeklinde organize edebileceği veri tabanına ihtiyacı vardır. Veri tabanı oldukça basit, örneğin, tek bir tablo olabileceği gibi, pek çok tablo ve ilişki içeren karmaşık bir sistem de olabilir. Solucan, ayrıca, birden çok veritabanında birden çok sistemi de desteklemektedir. Alt yapı malzemesi olarak kullanılacak veri tabanı yönetim sistemi uygulamasının mutlaka yapısal sorgulama dilini (SQL) tam desteklemesi gerekmektedir. Biz bu çalışmada, görsel veri tabanlı tasarım yazılımı olarak serbest, açık kaynak kodlu lisansa sahip olması nedeniyle DbDesigner 4 'ü kullandık (http://fabforce.net/dbdesigner4/). İlişkisel veri tabanı yönetim sistemi olarak serbest, açık kaynak kodlu olması ve SQL

'i tam desteklemesi nedeniyle MySQL 5.1'i kullandık (http://www.mysql). Ne yazık ki hem DbDesigner 4 hem de MySQL 5.1'in İngilizce yazılmış olmaları ve Türkçe değişken isimlerine izin vermiyor olmaları nedeniyle değişken isimleri programsal karmaşıklığa neden olmaması için İngilizce olarak tercih edilmiştir. Düzenli ifade motoru olarak yine herkese açık, açık kaynak kodlu olan deelx kütüphanesi kullanılmıştır. Biz bu çalışmamızda çok basit yapılı bir veri modelini düşündük. Ve bu model ile Genlerin isimlerinin, Sembollerinin, eskiden kullanılan takma isimlerinin, kromozomal lokalizasyonlarının ve ait oldukları gen ailesi ile ait oldukları lokus tipini ve bu genlere ulaşmak için kullanılan erişim numaralarını bulmayı hedefledik. Ayrıca elde ettiğimiz genlerin mRNA dizileri üzerinde nükletidleri sayıp depolayabilir olmasını hedefledik. Bu projeye CoreData adını verdik. CoreData, projesinde kullandığımız tablolar aşağıda Tablo-1'de sunulmuştur. Bu tablolar arasındaki ilişki ise DbDesigner 4 programı kullanılarak görsel hale getirilip Şek. 1 'de sunulmuştur. Bu veri tabanının adı Solucandır.

Tablo 1. Kullanılan tablolar ve bunların kullanım amaçları

TABLO ADI	KULLANIM AMACI
Locus Type	Var olan lokus tiplerini depolar
Family Name	Var olan gen aile adlarını depolar
Core Data	Genlerin sembolleri, adları, kromozomal yerleşimleri gibi bilgileri depolar
Accession Numbers	Genlerin erişim numaralarını depolar
mRNA Nucleotides	mRNA nükleotid dizilerinde A,T,G,C miktarlarını depolar

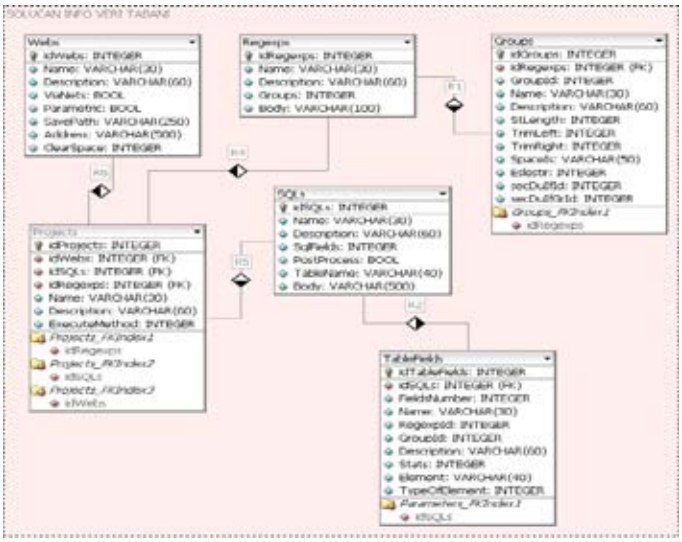


Şek. 1: Solucan için hazırlanmış CoreData veri tabanı (Solucan).

7. Solucan varlıkları

Solucan üç ana varlık ve bu varlıklar arasında tanımlanmış bağlantı gruplarının derlenmesi ile çalışmaktadır. Bu üç ana varlık Solucan'ın aynı zamanda ham veriden bilgi elde etmek için kullandığı yol haritasıdır. Kendi içinde kayıtlı sabit yol haritaları yoktur. Harita sizin tarafınızdan tanımlanan direktifler ve varlıklardan oluşur. Bu sayede kullanıcılar ihtiyaç duydukları verilere kendi alıştıkları şekilde ulaşabilmektedir. Bu üç ana varlık sırası ile şunlardır. Web varlığı, düzenli ifade varlığı, SQL varlığı. Ayrıca düzenli ifade varlığının grup bağlantı varlıkları ile SQL varlığının alan bağlantı varlıkları, üç ana varlığı ilişkisel olarak birbirine bağlamaktadır. Solucan varlıkları için kullanılan veri tabanının adı SolucanInfo olarak verilmiştir. SolucanInfo veri tabanında bulunan tablolar ve bunların işlevleri Tablo-2'de sunulmuştur.

Varlıkları tanımlayan tablolar SolucanInfo veritabanı içerisinde DbDesigner 4 kullanılarak ilişkilendirilip görsel hale getirilip Şek. 2 'de sunulmuştur.



Şek. 2. Solucan ınfo veri tabanı

8. Web varlığı

Web varlığı ham verinin alınacağı yeri belirler. Ham veri web varlığı tarafından bir web sunucusundan, FTP sunucusundan ya da yerel depolama biriminden alınabilir. İndirilen ham veri belirtilen adrese kaydedilir. Ayrıca, Solucan pek çok web sitesinden ya da ham veri kaynağından verileri sırası ile alacak şekilde de kullanılabilir.

9. Düzenli ifade varlığı

Düzenli ifade varlığı ile kaydedilen ham veri bilgi grupları şeklinde parçalanır. Bir düzenli ifade varlığının kendi içindeki bir grup için tekrar bir düzenli ifade varlığının tanımlanması mümkündür. Kendi üzerine dönüşlü veri modeli sayesinde ham veri üzerinde sayma, frekans hesaplama gibi istatistiki işlemler de yapabilir. Aslında ana

Tablo 2. Solucan veri modelimizdeki tablolar ve kullanım amaçları

TABLO ADI	KULLANIM AMACI
Webs	Web varlıkları için tablo
RegExps	Düzenli ifade varlıkları için tablo
SQLs	SQL ifadeleri için tablo
Groups	Düzenli ifade grupları için tanımlayıcı tablo
Table Fields	SQL alanları için tanımlayıcı tablo
Projects	Solucan direktifleri için tanımlayıcı tablo

düzenli ifade grubuna dahil olmayan bu şekildeki gruplar Solucan için sanal gruplar olarak adlandırılmıştır.

10. SQL varlığı

Ham veri; bilgi gruplarına, Solucan sanal grup-

larına parçalandıktan sonra, hedef ilişkisel veritabanına SQL varlığı kullanılarak kaydedilir. Her SQL varlığı bilgi grupları ile veri tabanı alanları arasında ilişkileri içerir. Bu ilişkiler tasarlanırken SQL alanlarının bilgi grupları kümesinin bir alt kümesi olması kuralı ihlal edilmemelidir.

11. Solucan projeleri

Solucan projesi, yukarıda vurgulanan üç varlığın bileşimi ve tanımlanan projenin Solucan tarafından ne şekilde derleneceğini içeren kullanıcı direktifidir. Solucan' ın iki derleme şekli bulunur, bunlardan ilki düzenli derleme ve ikincisi parametrik derlemedir. Düzenli derleme, bir projenin tek kaynaktan ham veri şeklinde alınıp içindeki istenilen bilgi gruplarının çıkartılıp kaydedilmesi şeklinde yapılır. Parametrik derleme ise, birden çok kaynaktan birden çok kere indirilen farklı ham verilerin içindeki istenilen bilgi gruplarının çıkartılıp kaydedilmesi şeklinde tanımlanır. İstenilen genlerin nükleotid dizilerine, amino asit dizilerine sahip olmak parametrik derleme ile mümkün kılınmıştır. Bu şekilde seçilen parametreler için farklı kaynaklardan gelen farklı ham verilerin bütünleşmesine müsaade edilir ki bu Solucan'ın, yazından farklı bir uygulamasıdır.

12. Derleme sonuçları

Biz hedefimize uygun olarak HGNC web sitesinden (http://www.genenames.org/) aldığımız ham verileri kullandık. Örneğin gen aileleri tablosunu oluşturmak için web sayfasını HGNC web sitesinden seçtik ve buradan istediğimiz bilgiyi elde eden (\d{1, })\t(.*) şeklindeki 2 gruptan oluşan düzenli ifadeyi oluşturduk. Bu düzenli ifadenin ikinci grubu aradığımız gen aile adını taşıdığından tek alan grubuna sahip SQL varlığı ile Solucan veri tabanındaki Family Name tablosuna kaydedilir. Bu tip tek alan adları içeren tablolar Solucan tarafından otomatik olarak tekilleştirilir ve tekrar eden kayıtlar tablodan uzaklaştırılır. Bu işlem Solucan tarafından 6 dakikada tamamlanmıştır. 6 dakika sonunda Family Name tablosunda tekrarsız olarak 332 kayıt göze çarpmıştır. Benzer bir yaklaşım aynı düzenli ifade kullanılarak lokus tipi üzerinde uygulanmıştır. 6 dakika derleme zamanı hesaplanmıştır. HGNC web sitesinden bu sefer gen adı, gen sembolü, lakapları, kromozomal yerleşimi (ki bizim projemizde bu yerleşim kromozom no, kromozom kolu, ana bant, alt bant olarak tekrar parçalanarak kaydedilmiştir) içeren ham veri 8 dakikalık bir zamanda derlenmiştir. Bunun için kullanılan düzenli ifade (\d{1,})\t(.*)\t(.*)\t(.*)\t(.*)\t(.*)\t(.*) ve (\d{1, })\t(\d{1, }|X|Y)(p*q*[ter]*[cen]*)(\d*)\.*(\d*) şeklinde yazılmıştır. Erişim numaraları da (burada kullanılan düzenli ifade (\d{1, })\t(.*)\t(.*)\t(.*)\t(.*)\t(.*)) 6 dakikalık zamanda derlenmiştir. Bu şekilde oluşturulan CoreData projesi için toplam derleme zamanı 26 dakikanın biraz altındadır. Sonuç olarak bu kısa zamanda CoreData projesi altında Tablo-2'deki kayıtlar elde edilmiştir.

13. Tartışma

Görüldüğü gibi Solucan, benzer yazılımlardan farklı olarak, ham veri kaynaklarından yararlanabilmesi ve bunları önceden araştırmacının kendisinin öncelik ve isteklerine göre belirlemiş olduğu ilişkisellik içinde kaydetmesi yönüyle daha kullanışlı gibi görünmektedir. Fakat her araştırmacının Solucan benzeri yazılımlar kullanması da bizi aynı sorunlar dizisinin farklı bir yüzüne taşımaktan başka bir işe yaramaz. Bu durumda önemli olan şey biyoinformatiğin amacı olan farklı veri kaynaklarından farklı formlarda gelen farklı verileri ilişkisellik içinde bütünleştirip araştırmacıya bir bilgi paketi şeklinde sunmak olmalıdır. Bu bağlamda her kullanıcı, farklı veritabanlarından aldığı farklı bilgileri elinde tutacağından Solucanlararası bir bütünleşme ağının da kurulması bizi amaca biraz daha yaklaştıracaktır. Bir diğer tartışılması gereken sorun da araştırmacıların bir yerde yayınlamadığı yastık altında kalmış, sınıflandırılmamış verilerin durumudur. Yeni kazanımlar için artık verileri oldukça spesifik küçük parçalarına bölmek bu şekilde veri primitifleri oluşturmak gerekliliği tartışılmalıdır. Ancak bu şekilde yeni ilişkiler değerlenebilir. Belki de halka açık (public network) ağlar yerine dağıtık veri paylaşım ağları kurulmalıdır.

Tablo 3. Projesi altında derlemeler sonucu harcanan zaman ve elde edilen veri miktarı.

Tablo Adı	Web Varlığı	Düzenli İfade Varlığı	SQL Varlığı	Toplam	Kayıt Sayısı
Locus Type	00:00:06:628	00:02:12:406	00:15:13:891	00:17:42:925	26
Family Name	00:00:03:172	00:00:45:854	00:16:05:531	00:16:54:557	332
Core Data 1	00:00:16:922	00:06:37:000	00:15:14:781	00:22:08:703	32004
Core Data 2	00:00:04:125	00:00:55:703	00:12:30:515	00:13:30:343	32004
Accession Numbers	00:00:09:172	00:03:20:656	00:15:16:250	00:18:46:078	32004
mRNA Nucleotides	00:00:01:745	00:05:36:005	00:12:04:211	00:17:41:961	178

Tablodaki veriler saat: dakika: saniye: milisaniye formatında sunulmuştur.

14. Sonuç

Genetik veri uzayının devasa büyüklüğü ile boğuşup galip gelmek mümkün gibi gözükmemektedir. Verileri parçalayıp dağıtık sistemlere geçmek sorunumuzu azaltacaktır. Bu bağlamda Solucan soruna kısmi de olsa çözüm üretebilmektedir. Gelecek çalışmalar Solucanlar arasında olan entegrasyonun kuvvetlendirilip dağıtık veri paylaşım ağı kurmak konusunda yapılacaktır.

15. Teknik altyapı

Solucan programı C++ dilinde yazılmıştır. Gerekli olan veri tabanı yazılımı MySQL olup görsel tasarım aracı olarak DbDesigner kullanılmıştır. Programa ulaşmak için mkamilturan@gmail.com ya da hasanb@omu.edu.tr adreslerinden yazarlar ile temasa geçmek gerekmektedir. Geliştirildiği bilgisayar Core 2 Duo E4300 1.8Ghz 2Gb Ram olup Solucan için 80Gb yerel hard disk alanı ayrılmıştır. Düzenli ifadeleri derlemek için açık kaynak kodlu serbest bir C++ kütüphanesi olan deelx.h kullanılmıştır (http://www.regexlab.com/en/deelx/).

KAYNAKLAR

Baxevanis, A.D., 2006. Curr Protoc Bioinformatics. Chapter 1:Unit 1.1. The importance of biological databases in biological discovery.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L., 2008. Genbank. Nucleic acids research, 36(Database issue).

Cochrane and Galperin, 2009. The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources Nucleic Acids Research Advance Access published on December 3, Nucl. Acids Res. 2010 38, D1-D4; doi:10.1093/nar/gkp1077

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N., Gibson, R., Hoad, G., Hubbard, T., Hunter, C., Jang, M., Juhos, S., Leinonen, R., Leonard, S., Lin, Q., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Plaister, S., Radhakrishnan, R., Robinson, S., Sobhany, S., Hoopen, P. T. T., Vaughan, R., Zalunin, V., and Birney, E., 2009. Petabyte-scale innovations at the european nucleotide archive. Nucleic acids research, 37(Database issue), 19–25.

D'Addabbo, P., Lenzi, L., Facchin, F., Casadei, R., Canaider, S., Vitale, L., Frabetti, F., Carinci, P., Zannotti, M., and Strippoli, P., 2004. Generecords: a relational database for genbank flat file parsing and data manipulation in personal computers. Bioinformatics, 20, 2883–2885.

DBDesigner 4 [http://fabforce.net/dbdesigner4/]

Drug Discov Today. 2001. Molecular biology databases: today and tomorrow. Ellis LB, Attwood TK. 6, 509-513

Deelx Regular Expresşion Engine V1.2 [http://www.regexlab.com/en/deelx/]

DbSNP [http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi]

EMBL [http://www.ebi.ac.uk/embl/Services/DBStats/]

GenBank [http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html]

Hong, P. and Wong, W. H. 2005. Genenotes–a novel information management software for biologists. BMC bioinformatics, 6.

Kirov, S., A. Peng, X., Baker, E., Schmoyer, D., Zhang, B., and Snoddy, J., 2005. Genekeydb: a lightweight, gene-centric, relational database to support data mining environments. BMC bioinformatics, 6.

MEDLINE verileri [http://www.nlm.nih.gov/bsd/history/tsld024.htm]

MySQL 5.1 [http://www.mysql]

Philippi, S. and Kohler, J., 2006. Addressing the problems with life-science databases for traditional uses and systems biology. Nat Rev Genet, 7, 482–488.

PubMed [http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmeddata.html]

Shah, S. P., Huang, Y., Xu, T., Yuen, M. M., Ling, J., and Ouellette, B. F. 2005. Atlas - a data warehouse for integrative bioinformatics. BMC Bioinformatics, 6.

Wang, Liu S, Niu T, Xu X. 2005. SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management. BMC Bioinformatics. 6,60.