

NEDEN BİYOİNFORMATİK?

Rengün Çetin ATALAY*

Recent progresses in molecular biology have revealed the large fractions of the genome sequences of several species during the last decades. Public sequence databases, such as GenBank and SwissProt, have been growing at exponential rates. Storage, organization and cataloging of this information become indispensable by information science methods under the field called bioinformatics. Bioinformatics deals with the recording, storage, annotation, analysis and retrieval of sequence (DNA, RNA or protein) and structural information. Eventually this pioneering new field aims to enable the discovery of new biological insight as well as to create a global perspective of virtual cells that can be used as models for disease prediction, diagnosis and treatment.

Geçtiğimiz son 15-20 yıl içerisinde geliştirilen teknikler sayesinde moleküler biyoloji bilimi çok çeşitli seviyelerde kapsamlı gelişmelere tanık olmaktadır. Hızlı DNA dizi analiz yöntemleri çok çeşitli türlerin genomlarının DNA dizilerinin belirlenmesini sağlamıştır (Venter, 2001; Lander, 2001). Ancak inanılmaz bir hızla ve miktarda biriken bu verileri saklamak ve analiz etmek giderek zorlaşmıştır (Tablo1). Bu bağlamda bilgisayar biliminin giderek artan ivmelenmeyle biriken bu verilerin dikkatli bir şekilde saklanması, düzenlenmesi, birleştirilmesi, kataloglanması ve kolayca erişilmesinde katkısı büyük olmuş ve biyoinformatik bilimi böylece doğmuştur (Altschul, 1994).

Biyoinformatik biyoloji, bilişim teknolojisi ve bilgisayar bilimini bir araya getiren bir bilim dalıdır. Yakın gelecekte DNA dizileri bünyesinde genom düzeyinde saklanan ve gerektiğinde kullanılan yaşam şifresi biyoinformatik çalışmalarının da katkısıyla daha hızlı çözülebilir ve böylece gelişimsel biyolojiden kanserin tanı ve tedavisine kadar bir çok biyolojik olayın altında yatan mekanizmalar açığa çıkarılabilir (Scherer, 2001).

veri tabanı	saklanan veri tipi	dizi sayısı	
		1986	2001
EMBL-GENBANK	Nükleotid dizileri	9978	14976310
SWISSPROT	Protein dizileri	4060	114033
PDB	Protein yapısı	296	16973

Tablo1. Biyolojik verilerin yıllar içindeki değişimi

* Yrd. Doç. Dr., Moleküler Biyoloji ve Genetik Bölümü Bilkent Üniversitesi

Başlangıçta biyolojik veri tabanlarının kurulması ve verilerin kataloglanıp saklanması ile ilgili çalışmalar sadece biyoinformatiğin ana amacını kapsarken, günümüzde birçok biyolojik veriye birden ulaşımı sağlayan, yeni verilerin kullanıcılar tarafından veritabanına girişini ya da var olan verilerin değiştirilmesi ile ilgili oldukça karmaşık altyapıya sahip veri tabanlarının oluşturulması bu bilim dalını gelişmesine ve daha kapsamlı çalışmalara yöneltmiştir. Ayrıca var olan biyolojik verilerin normal fizyolojik koşullarda ve hastalık sırasında hücresel olayların nasıl düzenlendiği konusunda kullanmak amacıyla çeşitli biyoinformatik analiz yöntemleri de geliştirilmektedir. Bu nedenle biyoinformatik bilimindeki gelişmeler daha çok bu verilerin analizi ve yorumlanması yönündedir. Bu bağlamda nükleotid ve amino asit dizi analizi, proteinlerin işlevsel alt birimleri ve protein yapı analizleri bilişimsel biyoloji (computational biology) ile çözümlenmektedir.

Biyoinformatik ve bilişimsel biyoloji çalışmaları başlıca çeşitli tipdeki biyolojik bilgilere verimli erişilmesi, bu bilgilerin verimli kullanım ve idaresi için gerekli bilgisayar yazılımlarının geliştirilmesi ve gerçekleştirilmesini ve bu çalışmalar ile gerçekleştirilen büyük veri kümeleri arasındaki ilişkilerin değişik algoritmalar ve istatistiksel yöntemler ile açığa çıkarılmasını kapsamaktadır (Baxevanis, 2001). Topluca bu çalışmaları 5 ayrı grupta toplamak mümkün olmaktadır: 1. Biyolojik bilgiler ile ilgili veri tabanı oluşturulup yönetilmesi, 2. Genom ve dizi analizleri, 3. İşlevsel genomik ve ilgili veri analizi, 4. Birincil protein dizisinden protein üç boyutlu yapı tahmini, 5. Hücresel olayların modellenip insliko gerçekleştirilmesi.

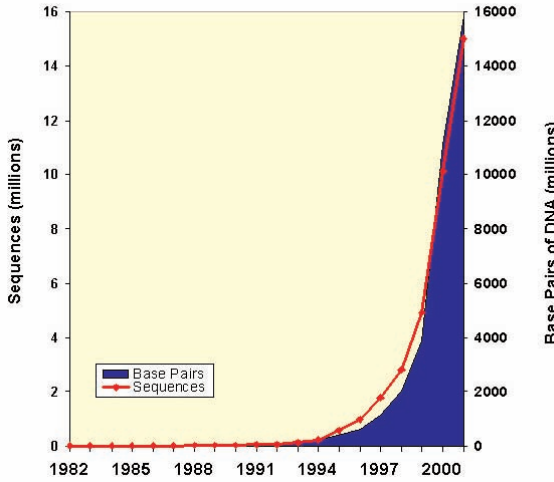
Biyolojik veri tabanları

Biyolojik veri tabanları belirli tipte (DNA, Protein) organize edilmiş verileri içermektedir. Çoğunlukla bu verilere ulaşmak, yenilerini eklemek veya değiştirmek için o veri tabanına uygun bir yazılım ile birlikte sunulur. Nükleotid dizi, protein dizi ve protein yapısı ile ilgili olarak biyolojik veri tabanları başlıca üç ana grupta toplanır. Ayrıca çeşitli analiz yöntemleri kullanılarak bu başlıca 3 tip veriden kaynaklı özelleştirilmiş veri tabanları da bulunmaktadır.

Nükleotid veri tabanları başlangıçta birbirlerinden bağımsız olarak nükleotid dizi verilerini toplamakta idi. Ancak yeni moleküler biyolojik yöntemler sayesinde hızla artan dizi verilerini tek bir kaynaktan toplamak hem biyologlar hem de biyoinformatisyenler için bir gereklilik haline geldi. Böylece tüm nükleotid dizi verilerini bir araya toplayan uluslararası işbirliği gurubu (International Nucleotide Sequence

Database Collaboration) başlıca DNA dizilerini içeren üç merkez olan Japon DNA veri tabanı (DNA DataBank of Japan (DDBJ)), Avrupa Moleküler Biyoloji Laboratuvarları (the European Molecular Biology Laboratory (EMBL)), ve Amerikan Ulusal Biyoteknoloji Bilgi Merkezi GenBank veri tabanı (GenBank, NCBI) tarafınca kuruldu (Benson, 2002). Böyle bir ortak grup çerçevesinde nükleotid dizilerinin sunumunda ve betimlenmesinde ortak kurallar uygulanmasının yanı sıra var olan bütün DNA dizilerinin kaynaklandığı ülke kayıtları tutulmakta ayrıca türler içi ve arasında evrimsel bağlantılar kurulması (taksonomi) amacıyla ortak projeler yürütülmektedir. Bu üç veri tabanından herhangi birine yollanan her nükleotid dizi verisi günlük olarak diğerlerinde de aynı veri yapısında yayımlanmaktadır.

Growth of GenBank



Şekil 1: GenBank verilerinin büyüme grafiği (<http://www.ncbi.nlm.nih.gov> adresinden alınmıştır.)

GenBank'ta Haziran 2002 tarihinde yaklaşık 20.649.000.000 bazı içeren 17.471.000 nükleotid dizisi bulunmaktadır (Şekil 1). Uluslararası ortak işbirlikleri her ne kadar var olan nükleotid verilerine bir düzenleme getirmiş olmalarına rağmen bu verilerin de analizi ve kataloglanması yanlış veya tekrarlayan dizilerin var olan veriler kaybolmaksızın ayıklanması gerekmektedir. Ayrıca kullanıcıların istedikleri verilere hızlı kolay anlaşılabilir ve amaçlarına uygun ulaşabilmeleri için gerektiğinde yeni dizileri yollayabilecekleri yazılımlar da bu veritabanlarının sağlanmaktadır.

Protein dizi veri tabanları aynen canlı bir hücrede olduğu gibi, proteom verilerini genom verilerine göre daha az sayıda ve daha düzenli bir şekilde sunmaktadır. Başlıca SWISS-PROT ve tarafınca sunulan protein dizi veri tabanları, nükleotid dizi veri tabanlarına göre daha düzenlidirler. SWISS-PROT veri tabanında Ağustos 2002 tarihinde araştırmacılar tarafından insan eliyle düzenlenmiş 112.892 ve EMBL nükleotid dizi verilerinden kaynaklanan trEMBL (translated EMBL) veri tabanı başlığı altında 668.930 protein dizi verisi bulunmaktadır. Nükleotid veri tabanlarında olduğu gibi, protein dizi bilgisi ve bu diziyeye ait diziyeye özellikleri betimleyen veriler olmak üzere protein dizi verileri de iki tip bilgi içerir. Ana protein dizi verisi (dizi bilgisi, organizma türü, kaynağı) yanısıra, proteinin işlevi, Post-translasyonel değişiklikler, proteinin işlevsel bölgeleri, protein yapı bilgileri, diğer proteinlerle olan benzerlikleri, hastalıklarla ilişkileri ve diziyeye ait çeşitlilikler ve hatalar ilgili protein dizisini betimleyen özelliklerdir.

Biyolojik veri tabanları arasında ilk kurulanı protein yapı bilgisi içeren protein veri tabanıdır (Protein DataBank, PDB). 1971'de protein yapı arşivi olarak Brookhaven National Laboratories tarafından kurulan PDB ilk protein yapı verilerini (sadece 7 protein) araştırmacıların kullanımına sunmuştur. 1957 de ilk protein yapısı belirlendikten bu yana geçen 45 yıl içerisinde PDB deki veri sayısı ancak 1980'lerde gelişen teknik imkanlar sayesinde artmaya başlamıştır. Ağustos 2001 tarihinde 18.488 veri içeren PDB'sına verilerin girişi, protein yapısının değerlendirilip doğrulanıp onaylanması ve kullanıma sunulması Rutgers, The State University of New Jersey, University of California, San Diego Supercomputer Center ve National Institute of Standards and Technology tarafından ortaklaşa kurulan işbirliğince (Research Collaboratory for Structural Bioinformatics (RCSB)) yürütülmektedir.

Genom ve dizi analizleri

Günümüzdeki geniş amaçlı genom araştırmaları ve gelişmiş veri tabanları sayesinde biyolojik dizi bilgilerine internet ve World Wide Web üstünden ulaşmayı mümkün kılmaktadır. Sonuç olarak bu veri tabanlarının çeşitli amaçlarla kullanımı hemen hemen her moleküler biyoloji laboratuvarının günlük aktiviteleri arasına girmiştir. Dizi eşleme yöntemlerine (sequence alignment) (Şekil 2) dayanan dizilerinde belli bir oranın üstünde benzerlik gösteren (homolog) genlerin araştırılması moleküler biyologlar tarafından en sıkça veri tabanlarının kullanılmasını gerektirmektedir. Farklı organizma türlerindeki homolog genler ortolog (orthologous) genler olarak adlandırılır. Ortolog genlerin farklı türler arasında genom analizi aracılığı ile bulunması yeni bulunan

genlerin işlevlerinin açığa çıkarılmasında büyük rol oynar. Bu amaçla sekans eşleme amacıyla bilişimsel biyoloji yöntemleri (başlıca BLAST ya FASTA) geliştirilmiştir (Altschul, 1990). Bu yazılımlar verilen bir nükleotid ve protein dizisini kullanarak ilgili veri tabanlarını tarar ve olası homolog genleri bulurlar. Bu bağlamda ekmek mayası, sirke sineği gibi hem hızla çoğalan model organizmalar kullanılarak bulunan genlerden elde edilen veriler daha sonra memeli organizmalarının ortolog genlerinin bulunmasında işlevlerinin açığa çıkarılmasında önemli role sahiptirler (Lipman, 1984).

Tam eşleme =47/153 (30%), boşluklar= 18/153 (11%)

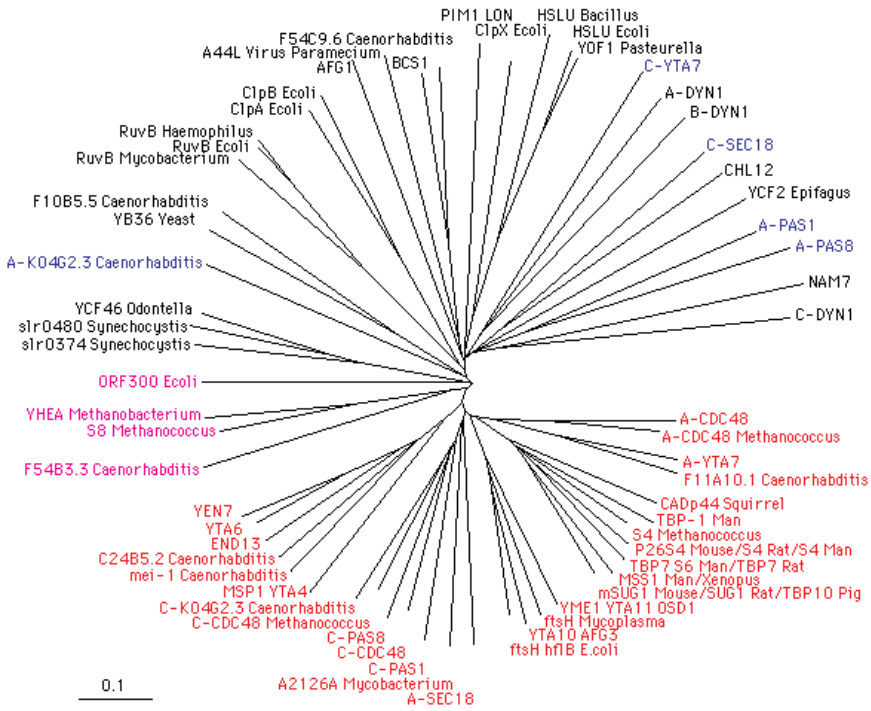
Dizi1: 8 ILYPTDFSETAEIALKHVKAFKTLKAEVILLHVIDEREIKKRDFSLLLGVAGLNKSVE 67
 +L PTD SE + L+++ FK + EE+ +L VI+ ++ S + G ++ ++
 Dizi2: 1 MLLPTDLSSENSFKVLEYLGDFFKVGVEEIGVLFVINLTKL-----STVSGGIDIDHYID 54
 Dizi1: 68 EFENELKKNLTEEAKNKMENIKKELEDVGFVKVDI-IVVGIPHEEIVKIAEDEGVDDIII 125
 E ++E+A+ + + +++E G K + I G P EI+K +E+ I
 Dizi2: 55 E-----MSEKAEVLPFEVAQKIEAAGIKAEVIKFPFAGDPVVEIHKASEN-YSFIA 104

Şekil 2: BLAST formatında protein dizi analizi örneği. Herbir harf bir amino asit için kullanılmıştır. İki dizi rarsındaki benzerlik kırmızı ile belirtilmiştir. + benzer amino asitler için kullanılmıştır.

Aynı tür içinde de evrim süresince oluşan tekrarlayan genom organizasyonu sonucunda bazı DNA parçacıkları genom içinde birden fazla gen de yer alabilirler bu tür benzerlikleri içeren genler de paralog (paralogous) genleri oluştururlar (Dayhoff, 1976). Bu benzer DNA parçacıklarının kodladıkları protein alt üniteleri hücre içinde aynı ya da çok benzer (örn. proteinleri fosforlayan kinazlar) üç boyutlu yapıya böylece de benzer görevlere sahiptirler. Proteinlerin bu özelliğini kullanarak çeşitli bilişimsel biyoloji yöntemleri ya da araştırmacılar tarafından elle protein veri tabanları incelenmiş benzer işleve sahip proteinlerin sınıflandığı yada sadece kısa (ortalama 5-30 amino asit) peptid dizilerini içeren özelleşmiş veri tabanları geliştirilmiştir. PFAM benzer işlevli protein ailelerine ait işlevsel protein altünitesi verilerini, BLOCKS ise benzer işlevli oligopeptidlere ait verileri içeren özelleşmiş veri tabanları olarak internet üzerinden hizmet vermektedir.

Dizi eşleme yöntemleri ayrıca genler arası ve dolayısıyla türler arası benzerlikleri açığa çıkardığı için türler arası evrimsel yakınlıkların (taksonomi) bulunmasını sağlar. Bu veriler kullanılarak türler arası filogenetik ağaçlar (Şekil 3) yaratılmaktadır.

Genom analizi arařtırmalarında geniş ölçekli DNA dizilerinin açığa çıkarılmasına paralel olarak kromozomlar üzerinde genlerin yerlerinin bulunmasında da sekans eşleme yöntemleri kullanılır. Böyle bir arařtırma sonucunda elde edilen verilerde ayrıca oluşturulan sadece gen bölgesi bilgisini içeren alt özelleşmiş veri tabanlarında tutulur. Bu bilgilere de gerektiğinde internet aracılığı ile ulaşmak mümkündür.



Şekil 3: Çeşitli organizmalarda ATP'ye bağlı işlev gören proteinlerin evrimsel ilişkilerini gösteren filogenetik ağacı (AAA veritabanından alınmıştır).

İşlevsel genomik ve ilgili veri analizi

Moleküler biyolojideki son gelişmeler arařtırmaları artık daha geniş kapsamlı olmaya yöneltmektedir. Günümüzde gen çipleri ile bir anda hücrenin bir anındaki binlerce genin durumu hakkında bilgi edinilebilmektedir. Böylece bilim adamları biyoloji ve tıpta daha kapsamlı arařtırmalar yapabilmekte ve birden bire binlerce gene ait ya da transkript ya da protein seviyesine ait veya proteinler arası ilişkilere açığa çıkaran binlerce bilgiye sahip olabilmektedirler (Claverie, 2001).

Tabii ki bu aşamada bu binlerce bilginin anlamlı hale getirilmesinde bilişimsel biyolojinin ve biyoinformatiğin rolü büyük olacaktır. Halen kullanımda olan yazılımlar istatistiksel yöntemler kullanarak bu tür verilerin sınıflamakta ve nükleotid veya protein veri tabanları ile ilişkilendirmektedirler. Örneğin aktiviteleri benzer şekilde kontrol edilen genler bu tür yazılımlar ile sınıflanabilmektedir. Geniş ölçekli genom ve proteom araştırmalarından elde edilen sonuçları kullanarak bu genlerin işlevleri ve birbirleri ile olan ilişkilerini açığa çıkaran yazılımlar geliştirmek bilişimsel biyoloji çalışmalarını oldukça uzun bir süre meşgul edecektir.

Geniş ölçekli gen analizlerinin bir başka amacı da yeni genlerin açığa çıkarılmasıdır. Genom projeleri sonuçlandııkça elde edilen DNA dizilerinde yeni genler çeşitli bilişimsel yöntemlerle tahmin edilebilir (Davuluri, 2001). Bu işlevsel genom araştırması sonucunda aktivitesi benzer şekilde kontrol edilen genlerden elde edilen deneysel bilgiler insliko gen yakalamak için kullanılabilir çünkü bu genler aktivitelerini kontrol eden bölgelerde benzer DNA dizilerine sahiptirler. Bu tür DNA sekaslarının otomatik olarak bilişimsel yollarla bulunması moleküler biyoloji araştırmalarına hem yeni genlerin ve işlevlerinin bulunması ve buna bağlı olarak hücresel olayların gelişiminin takibi açısından katkısı büyük olacaktır.

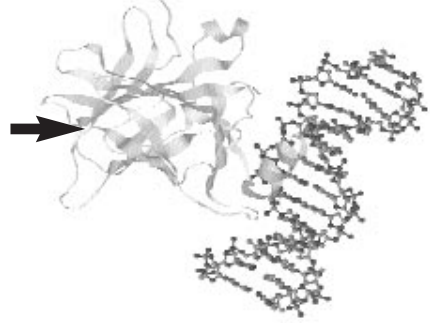
Birincil protein dizisinden protein üç-boyutlu yapı tahmini

Proteinlerin büyük çoğunluğunun işlevi bu makromoleküllerin üç-boyutlu yapılarına bağlıdır. Hücre içinde gelişen olaylar büyük çoğunlukla proteinler ve daha az oranda RNA moleküllerince sağlanır. Bu bağlamda protein ve RNA moleküllerinin yapısal özelliklerine ait her türlü bilgi bir hücrenin aktivitelerinin normal koşullarda, patolojik durumlarda belirlenmesinde ve ayrıca hastalıklarla savaşmak için yeni ilaçların geliştirilmesinde kullanılmaktadır. Biyolojik makromoleküllerin üç-boyutlu yapılarının deneysel yöntemlerle belirlenmesi hem çok pahalı hem de çok yavaştır. Bu nedenle bu makromoleküllerin üç-boyutlu yapılarının yeni geliştirilecek in sliko bilişimsel biyoloji yöntemleri ile belirlenmesine büyük gereksinim vardır. Herbiri ayrı özelliğe sahip ortalama 100-1000 adet, 20 değişik amino asitin ardarda sıralanıp daha sonra da birbirleri ile ana zincir ve yandalları ile ilişki içinde olduğu proteinlerin üç-boyutlu yapılarının belirlenmesi problemi günümüzde kullanılan bilişimsel yöntemlerin ve süper bilgisayarlar ile çözülememektedir (Şekil 4). Ancak bu yönde yapılan araştırmalar hızla ilerlemektedir ve çeşitli matematiksel modeller kullanılarak olumlu sonuçlar elde edilmektedir. Bundan önceki çalışmalar ile matematik-

sel modeller (yapay sinir ağları, saklı Markov modelleri) kullanılarak proteinlerin ikincil yapıları olan alfa sarmalı, beta tabaka ve serbest katlanmaları % 70'lerin üstünde bir doğruluk ile tahmin edilmiştir.

>p53 Tümör baskılayıcı Proteini

```
MDFVVDLPESQGSFQELWETVSYPPLETL-
SLPTVNEPTGSWVATGDMFLLDQDLSGTFDD-
KIFDIPIEPVPTNEVNPPPTTVPVTTDYPGSYELELR-
FQKSGTAKSVTSTYSETLNKLYCQLAKT-
SPIEVRVSKPEPPKGAILRATAVYKKTEHVADVVR-
RCPHHQNEDSVEHRSHLIRVEGSQLAQYFED-
PYTKRQSVTPYEPQPGSEM TILLSYMCNSSCMG-
GMNRRPILITLLETEGLVLRRCFEVRICACPGR-
DRKTEEESRQKTQPKRRKVTPTSSSKRKKSHSS-
GEEEDNREVFHFVEVYGRERYEFLKKIND-
GLELLEKESKSNKDSGMVPSGKKLKS
```



Şekil 4: Protein birincil yapısından üç-boyutlu yapıya.

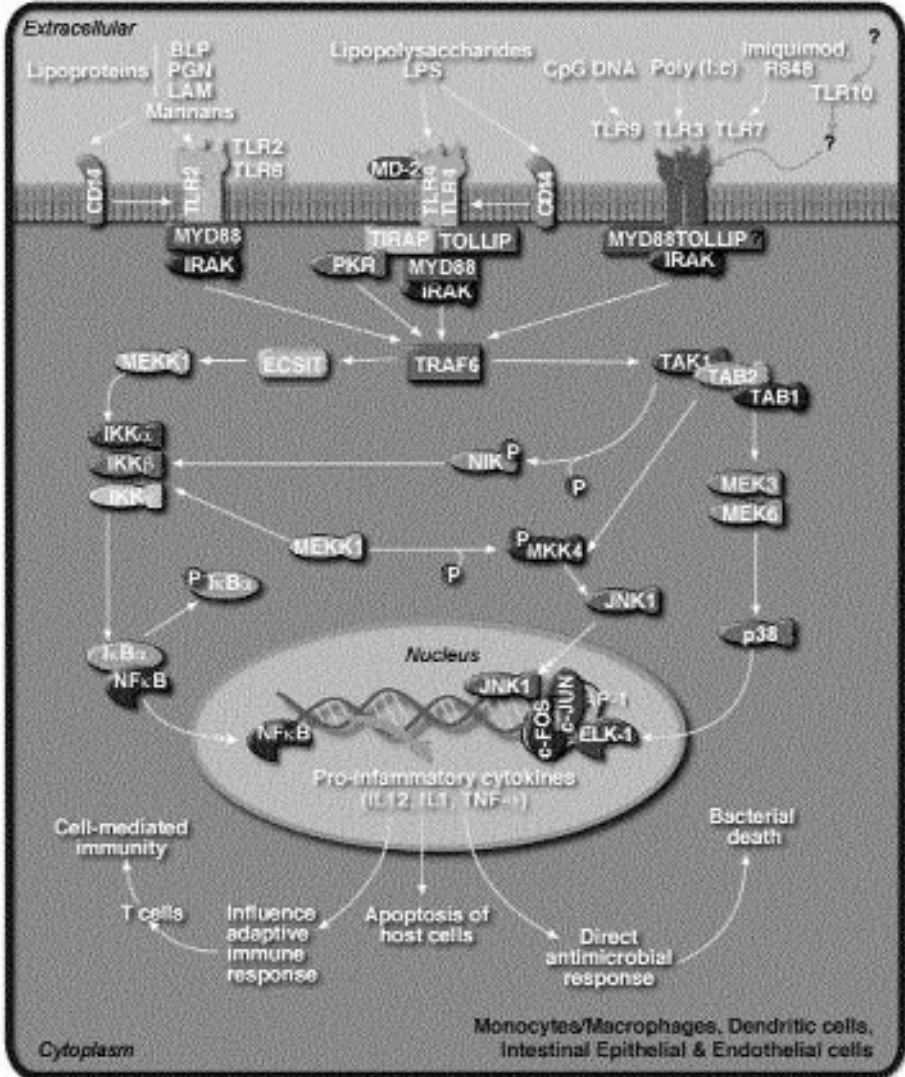
Proteinlerin üç-boyutlu yapılarının her ne kadar birincil yapıdan başlayarak tahmin etmek şu anki imkanlarla oldukça zor görünse bile proteinler arası homolojiden yararalanılarak benzer proteinlerin üç-boyutlu yapılarını tahmin etmek mümkündür. Amino asit dizilerinde ortalama % 25'in üstünde birebir eşleme gösteren proteinlerden birinin yapısı daha önce deneysel yöntemler ile belirlenmiş ise diğerinin yapısını in silico homolojiye dayalı yapı tahmini (homology based modelling) yöntemi ile belirlemek mümkündür. Başlıca "WHATIF" ve "Modeller" yazılımları bu yöntemle üç-boyutlu yapı tahmini için kullanılabilir. Proteinlerin üç-boyutlu yapılarının tahmininde kullanılan bir başka yöntem ise yine bu makromoleküllerin genelde benzer yapısal katlanma kurallarına uyarak üç-boyutlu yapılarını almaları fikrini kullanan iz sürme (threading) yöntemidir. "Threader", iz sürme yöntemi ile yapı tahmini için kullanılan bir yazılımdır.

Bilişimsel biyolojinin 3-boyutlu yapı tahmini çalışma alanı altında ayrıca proteinler arası veya protein bir başka molekül arası ilişkilerin tahmin edilmesini amaçlayan protein demirleme (Protein docking) alt çalışma alanı da bulunmaktadır. Deneysel olarak ya da modelleme ile belirlenmiş 3-boyutlu yapı bilgileri kullanılarak proteinlerin nasıl birbirleri ile etkileştikleri hakkında bilgiler edinebilmek hastalıkların temelini anlaşılmasında büyük öneme sahiptir. Viral bir proteinin hücre proteinleri ile olan ilişkisi moleküler düzeyde belirlenebilirse bu virüse karşı savaşılan ilaçların nasıl geliştirilebileceği yönünde bilgiler elde

edilebilir. Yeni geliştirilen bir ilaç molekülünün viral proteinler ile olan ilişkisi ve proteinlerin aktif bölgeleri protein demirleme yöntemi ile belirlenebilir. Proteine bağlanan bu küçük molekülün proteinin aktif bölgesine olan afinitesi, hangi koşullarda bağlantının kuvvetleneceği veya azalacağı tahmin edilebilir. Bu tür çalışmalar ilaç endüstrisinin yeni ilaç bulunması ve geliştirilmesi ile ilgili gelecekteki aktivitelerinin büyük çoğunluğunu oluşturacaktır.

Hücreyel olayların modellenip in sliko gerçekleştirilmesi

Geniş ölçekli işlevsel genom ve proteom araştırmaları sonucunda, yakın gelecekte hücreyel olaylar hakkındaki bilgimiz artan bir ivmeyle çoğalacaktır (Gribskov, 1999) (Şekil 5). Bu verileri saklayacak, birleştirecek, erişimini ve çözümlenmesini sağlayacak araçların geliştirilmesi öncelikli gereksinimlerden biridir. Bu bağlamda hücre içi fizyolojik olayların, proteinler arası ilişkiler sonucu oluşan ileti yollarının bilişimsel biyoloji ile modellenmesi gerekmektedir. Bir sinyal yolağının biyolojik anlamının saptanmasıyla elde edilen bilgiler bir sonraki dönemde ilgili hastalığın tanı ve tedavisinde geliştirilecek gen tedavisi, yeni ilaçların bulunması ve genetik tanı kitleri gibi yöntemlerin bulunması için kullanılabilir.



Şekil 5: Örnek hücresel ileti ağı (BioCarta'dan alınmıştır).

Geniş ölçekli işlevsel genom ve proteom arařtırmaları verilerinin analizi için kullanımda olan yazılımlar genelde elde edilen sonuçları tek gen seviyesinde incelemektedir. Bu yazılımlar bir deney sonucu çıkan en azından yüzlerce ayrı veriyi aynı anda genlerin işlevleri açısından analiz etme yönünde yetersiz kalmaktadırlar . Ayrıca, bu tür yazılımlar tek bir hastalık tipine yönelik (örneğin kanser) bilgi içermektedirler. Mikroarray sonuçları bir hücrenin biyolojik örneđi alındığı anda içinde bulunduđu durumun bir aynasını oluşturmaktadır. Bu nedenle elde edilen sonuçlar tek bir gen düzeyinde kalmayıp, genler ve ürünlerinin birbirleri ile ilişkilerinin gösterilebilmesi ve biyolojik anlamının açığa çıkarılması için kullanılabilmesi gereklidir.

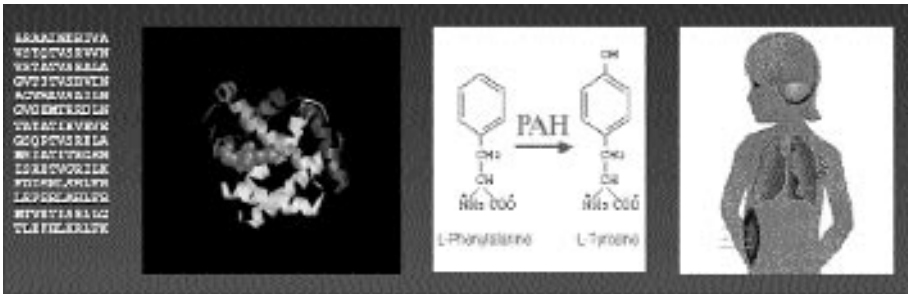
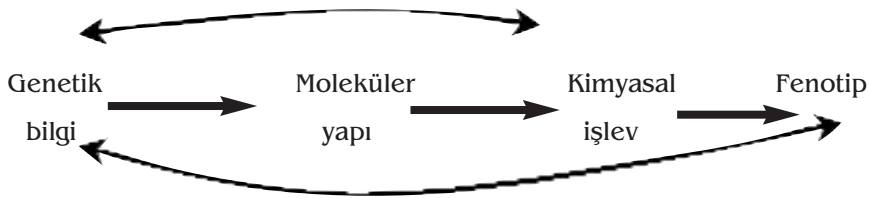
Geniş ölçekli ekspresyon analizlerinin başlangıç aşamasında sinyal yollarının seçimi, mikroarray sonuçlarının bilişimsel analizi, biyolojik bilgi bankalarına ulaşım, verilerin biyoinformatik analizi gibi konularda uzmanlaşacak olan biyoinformatik yazılımlarının geliştirilmesi ve faaliyete geçirilmesi gerekmektedir. Bu çalışmalarla binlerce genin ifadesini gelişigüzel incelemek yerine, daha önceden in silico yöntemler ile hücre içindeki moleküler ilişkileri modellemek amacıyla aday hücrenel yolak ve genlerin seçiminde kullanılacak biyoinformatik yazılımları geliştirilmelidir. DNA’da yer alan tek boyuttaki otuz bin genin hücre içerisinde üç boyutta işlev gören protein ürünleri oldukça karmaşık bir ağ oluşturmaktadır. Proteinler üç boyutlu yapılarına bađlı olarak 4-5 ayrı formda bulunup şekilde de belirtildiđi gibi her bir form da en az 5-10 ayrı protein veya diđer moleküller ile etkileşebilir. Bu nedenle hücre içi moleköl ilişkilerini modelleyip web üstünden kullanılabilir yazılımlar geliştirilmektedir. Moleküller arası ilişkiler ađı uygun bir veri tabanında saklanmalı ve çeşitli sorgulama yöntemleri ile yeni (bilinmeyen) ilişkilere ulaşabilmek olanađı sağlanmaktadır.

“Transpath” hücrenel ileti yollarını ve “KEGG” ise metabolik ileti yollarının verileri ile oluşturulmuş veritabanlarıdır. Ancak, bu yazılımlar, verileri daha çok yazı ile gösterdikleri için bilginin hücre yolađı olarak takibi zor olmaktadır. Bu nedenle göze hitap eden grafik ortamda bu bilgiler sunulmalıdır. BioCarta göze hitap eden resim halinde sinyal ileti yollarını içeren www sunucularından biridir (Şekil 5). Ancak, bu tür uygulamalarda bilgiyi resim olarak sakladıkları Transpath ve KEGG gibi veri tabanı olmaktan uzaktırlar. Bilgiye erişim ve sorgulama kapasiteleri sınırlıdır ve ileti bilgisi yolları arası ilintilenmemiştir. Bu bağlamda son zamanlarda geliştirilen bir yazılım olan PATİKA (Pathway Analysis Tool for Integration and Knowledge Acquisition) hem sorgulanabilen bir veri tabanına sahip olması hem de grafik ortamda birbirleri ilintili olarak

hücre içi ileti yollarını göstermesi açısından yukarıda bahsedilen iki tip yöntemi bir araya getirmektedir (Demir, 2002). Bu yazılımda, moleküller arası ilişkiler ağı nesneye yönelik bir veri tabanında saklanmakta ve çeşitli sorgulama yöntemleri ile yeni (bilinmeyen) ilişkilere ulaşılma olanağı sağlanmaktadır. Bu yazılımın bir diğer önemli özelliği de, ortaya çıkan bilgilerin genellikle daha küçük bir ilişkiler ağı olarak otomatik görüntüleyebilmesidir. PATİKA yazılımı veritabanı, sunumcu ve web den kullanıma uygundur. Böylece geniş ölçekli işlevsel genom ve proteom araştırmaları sonucunda elde edilen yüzlerce gene ait veriler bu tür yazılımlarla daha kapsamlı ve birbirleri ile ilintili olarak analiz edilebilir.

Biyoinformatiğin bu alt alanında tabii ki en son hedef hücre içi ileti ağını kullanarak hücre içindeki süreçlerin benzetimidir (simulasyon). Hücresel olayların betimlenmesi ile genomdan gelen bilginin hücrede, çevreden ve kendi içinden gelen uyarılara karşı nasıl kullanıldığı ve böylece hücrenin normal fizyolojik durum dışında ve hastalık durumunda nasıl davrandığı tahmin edilebilir. Bu tür bilgiler hastalıklarla tanı ve tedavisinde yeni yöntemler geliştirmek için kullanılabilir.

DNA → RNA → Protein → Fenotip
A. Biyolojinin ana dogması



B. Biyoinformatiğin ana dogması

Şekil 6: A ve B : Biyoloji ve biyoinformatikte bilginin kullanım akışı (D. Broutlag'dan alınıp değiştirilmiştir).

Biyoinformatik ve Gelecek

Yukarıda verilen bilgilerin ışığı altında biyoinformatiğin ana hedefi, kitle halindeki dizi verilerini özel analiz yöntem ve araçlar tasarlayarak anlamlı küçük bilgi odakları yaratmaktır. Biyolojik dizi dili kullanılarak olabildiğince çok sayıda biyokimyasal işlev, yapısal bilgi ve evrimsel ipuçları bulunarak hastalıkların temeli anlaşılabilir (Şekil 6). Bu bağlamda başlıca proteinlerin diziden yapı belirlenmesi gibi, biyoinformatiğin önünde aşılması zor ancak imkansız olmayan bir dizi hedef vardır. Önümüzdeki 15-20 yıl içerisinde gerçekleşecek olan, biyoinformatik bilimi ile ilgili araştırmalar biyoloji biliminin önünde yer alan bir ışık görevi üstlenecektir.

KAYNAKÇA

Altschul, S.F., Boguski, M.S., Gish, W. & Wootton, J.C. (1994) "Issues in Searching Molecular Sequence Databases." *Nature Genet.* 6:119-129

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic Local Alignment Search Tool." *J. Mol. Biol.* 215:403-410.

Baxevanis, A. D. and Ouellette, B. F. F. (2001). "Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins" 2nd Edition. New York, NY: John Wiley & Sons, Inc., 356.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2002) "GenBank" *Nucleic Acids Res* Jan 1;30(1):17-20

Claverie, J. M. (2001). "Gene number. What if There Are Only 30,000 Human Genes?" *Science* 291(5507): 1255-7.

Davuluri, R. V., I. Grosse and M. Q. Zhang (2001). "Computational Identification of Promoters and First Exons in the Human Genome." *Nat Genet* 29(4): 412-7.

Dayhoff, M. O. (1976). "The Origin and Evolution of Protein Superfamilies". *Fed Proc*, 35(10), 2132-8.

Demir E, Babur O, Dogrusoz U, Gursoy A, Nisanci G, Cetin-Atalay R, Ozturk M. (2002) "PATIKA: An Integrated Visual Environment for Collaborative Construction and Analysis of Cellular Pathways." *Bioinformatics* Jul;18(7):996-1003

Gribskov, M. (1999). The New Biological Literature. *Bioinformatics* 15(5): 347.

Lander, E. S., L. M. Linton, B. Birren, et al. (2001). Initial Sequencing and Analysis of the Human Genome. *Nature* 409(6822): 860-921.

Lipman, D.J., Wilbur, W.J., Smith T.F. & Waterman, M.S. (1984) "On the Statistical Significance of Nucleic Acid Similarities." *Nucl. Acids Res.* 12:215-226

Scherer, S. W. and J. Cheung (2001). "Discovery of the Human Genome Sequence in the Public and Private Databases." *Curr Biol* 11(20): R808-11.

Venter, J. C., M. D. Adams, E. W. Myers, et al. (2001). "The Sequence of the Human Genome." *Science* 291(5507): 1304-51.